



AI Technology Landscape

(AIStatus 3)

**The Technologies, System Forms & Enabling
Developments Shaping Applied AI**

AI Sourced Facts (AISF) Pte. Ltd.

© AI Sourced Facts (AISF)

All rights reserved.

Edition: Version 1.0

Publication Date: September 2026

Preface

Artificial intelligence is often discussed as though it were a single technology.

It is not.

The current AI landscape includes general-purpose foundation models, specialised systems, reasoning capabilities, agents, multimodal systems, conversational and voice technologies, search and research systems, coding tools, synthetic-media technologies, physical AI, on-device systems, synthetic data, world models and the computing infrastructure that enables them.

These technologies are also developing at very different rates.

Some have been deployed for years and operate at substantial scale. Some have become widely available only recently but are already used by millions of people. Others demonstrate significant capability while remaining less established in dependable operational use. Research breakthroughs, commercial availability, pilot deployment and mature adoption therefore coexist across the landscape.

The boundaries between technologies are also becoming less distinct. A single AI system may combine a foundation model, reasoning, multimodal capability, memory, retrieval and agentic action. A robot may combine perception, world modelling, reasoning and physical control. An AI service may operate partly on a device and partly through large-scale remote infrastructure.

These overlaps can make the technological landscape difficult to understand.

AIStatus has been developed to provide a structured view of what is happening at a defined point in time.

This publication examines artificial intelligence through the **Technology Landscape** lens. Its purpose is not to catalogue every model, product or technical architecture. Nor is it to predict which technologies will ultimately prevail.

It is to establish, as coherently as the available evidence permits, what significant AI technologies, system forms and enabling developments currently constitute the applied AI landscape.

The Status presented in this publication is assessed as at **30 August 2026**.

Purpose of This AIStatus

AIStatus — Technology Landscape is designed to help readers **DISCOVER** what the **current applied AI technology landscape actually consists of**.

The publication examines fifteen elements:

1. Physical AI & Robotics
2. Generative AI & Foundation Models
3. Reasoning AI
4. AI Agents & Agentic Systems
5. Multimodal AI
6. Conversational & Voice AI
7. AI Search & Research Systems
8. AI Coding & Software Development
9. Image, Audio & Video AI
10. AI Memory, Context & Personalisation
11. Edge & On-Device AI
12. Specialised, Small & Domain AI
13. AI Data & Synthetic Data
14. World Models & Simulation
15. AI Infrastructure & Compute

Each element is examined independently but through a common set of questions.

The purpose is not to produce a technical catalogue, model leaderboard or vendor comparison. Technologies are explained in terms intended to be accessible to informed non-specialists while retaining the distinctions necessary to understand their present capabilities, maturity and application.

The publication concentrates on the current state.

It distinguishes, where the evidence permits, between research capability, demonstrated technical capability, product or service availability, pilot deployment, operational deployment, adoption at meaningful scale and demonstrated effectiveness.

These conditions are not treated as interchangeable.

This is not a forecast of the future of artificial intelligence. It is not a technology-selection guide, implementation manual, investment analysis, consultancy report or set of recommendations.

It is a point-in-time examination of the technological landscape as it can reasonably be understood from available evidence.

Global Perspective & Neutrality Statement

Artificial intelligence is a global technological development.

AI research, model development, semiconductor production, data-centre construction, robotics, device manufacturing, software development and AI deployment occur across interconnected international ecosystems.

No single country or region therefore represents the complete Technology Landscape.

The evidence examined in this publication includes developments from North America, Europe, China and wider Asia, the Middle East, Africa, Latin America and Oceania where they materially contribute to understanding the Status.

Geographical participation is not uniform. Frontier model development, semiconductor manufacturing, cloud infrastructure, robotics, language technology, device ecosystems and research capability are distributed differently across countries and regions.

The availability of public evidence is also uneven.

A genuinely global assessment therefore requires both active consideration of international developments and caution against manufacturing geographical symmetry where comparable evidence does not exist.

AISF is neutral, educational and non-prescriptive.

This publication does not rank AI models, companies, countries or technologies. It does not identify technological “winners”, recommend particular systems or providers, or assume that newer or larger technologies are inherently preferable.

References to particular organisations, systems and technologies are included where they provide material evidence of capability, availability, deployment or technological development.

Their inclusion does not constitute endorsement.

Editorial Standard

Technology develops through stages that are easily confused.

A research result may demonstrate that a capability is technically possible without establishing that it is commercially available.

A product release may establish availability without demonstrating substantial adoption.

A pilot can demonstrate operation in a defined environment without establishing dependable performance at scale.

High performance on a benchmark can establish capability under specified evaluation conditions without establishing equivalent performance across unfamiliar real-world situations.

Large numbers of users can establish adoption without demonstrating effectiveness.

Announced infrastructure investment does not establish that the infrastructure is operational.

These distinctions are central to the Technology Landscape assessment.

The publication therefore distinguishes, where materially relevant, between:

research capability → demonstration → product availability → pilot deployment → operational deployment → adoption at scale → demonstrated effectiveness

The sequence should not be interpreted as a mandatory development path. Different technologies develop and diffuse in different ways. Its purpose is to prevent evidence from one condition being treated automatically as evidence of another.

Terminology also requires care.

Terms such as *reasoning*, *agent*, *autonomous*, *multimodal*, *foundation model*, *world model* and *memory* can describe materially different technical arrangements or levels of capability. Where terminology is ambiguous, the publication explains what is meant rather than relying on the label alone.

Evidence is interpreted according to what it can reasonably establish.

Provider evidence can establish the existence, architecture or availability of a technology. Academic research can establish performance under defined conditions. Operational examples can establish real deployment. Surveys can provide evidence of reported use. Independent evaluations can test particular aspects of capability or reliability.

None automatically establishes the whole Status.

Where evidence is incomplete, uneven or uncertain, that uncertainty is preserved.

How to Use This AIStatus

Readers seeking an overall understanding of the current AI Technology Landscape may read the fifteen elements sequentially.

Readers interested in a particular technology or system form may go directly to the relevant element. Each has been designed to stand independently and provide sufficient context without requiring the reader to consult the other elements.

Each element begins with a concise **Status snapshot**, providing rapid orientation to its present maturity, demonstrated capability, application and principal qualifications.

Every element is then examined through the same seven questions:

1. What is it?

Explains the technology or system form in accessible terms and establishes its principal boundaries.

2. What can it currently do?

Examines demonstrated current capability while distinguishing technical achievement from dependable general performance.

3. Where is it currently being applied?

Identifies significant areas of real-world application and distinguishes availability, experimentation and operational deployment where necessary.

4. How developed is it?

Examines maturity without assuming that every capability or application within the category is at the same stage.

5. What is driving its development and use?

Identifies important technological, economic, operational, research and other evidenced drivers.

6. What are the present limitations?

Examines material constraints affecting capability, reliability, deployment, integration, scale or effectiveness.

7. What you should be aware of

Provides the principal distinctions and qualifications necessary to interpret the current Status without overstating capability, maturity or adoption.

Each element concludes with **Evidence Notes** identifying principal sources supporting its Status assessment.

The common structure enables comparison. It does not imply uniform maturity.

Differences between the technologies are themselves part of the Status.

Version & Governance

AIStatus 3 — AI Technology Landscape

The Technologies, System Forms & Enabling Developments Shaping Applied AI

Edition: Version 1.0 / 2026 Edition

Status Reference Date: 30 August 2026

Publication Date: September 2026

Series: AISF AIStatus

Published by: AI Sourced Facts (AISF) Pte. Ltd.

AIStatus publications provide point-in-time examinations of a changing artificial intelligence landscape.

The Status contained in this edition has been assessed using information available up to and including **30 August 2026**. Developments occurring after that date do not form part of the Status assessment for this edition.

Future editions may reassess the Technology Landscape at a later reference date. Technologies may develop, converge or become less significant, while new system forms and enabling developments may emerge.

The elements used to describe the landscape may therefore be retained, added, modified, combined, separated or removed where the evidence relevant to a future Status warrants it.

An earlier edition remains a record of the Status assessed at its stated reference date.

Editorial oversight of the AISF AIStatus series is maintained by AI Sourced Facts (AISF).

Table of Contents

Preface**Purpose of This AIStatus****Global Perspective & Neutrality Statement****Editorial Standard****How to Use This AIStatus****Version & Governance****Introduction — Understanding the Technology Landscape Status****Technology Landscape**

- 1 — Physical AI & Robotics
- 2 — Generative AI & Foundation Models
- 3 — Reasoning AI
- 4 — AI Agents & Agentic Systems
- 5 — Multimodal AI
- 6 — Conversational & Voice AI
- 7 — AI Search & Research Systems
- 8 — AI Coding & Software Development
- 9 — Image, Audio & Video AI
- 10 — AI Memory, Context & Personalisation
- 11 — Edge & On-Device AI
- 12 — Specialised, Small & Domain AI
- 13 — AI Data & Synthetic Data
- 14 — World Models & Simulation
- 15 — AI Infrastructure & Compute

The Status Across the Technology Landscape**AIStatus — Continuing the Exploration****Future Status Updates & User Submissions****Acknowledgements****AISF Universal Disclaimer****About AI Sourced Facts (AISF) Pte. Ltd.****Publication & Governance Information****Back Page**

Introduction — Understanding the Technology Landscape Status

The fifteen elements point to a distinct Technology Landscape story: AI is no longer adequately described by a single progression from “smaller” to “larger” models. The landscape now consists of interacting layers of capability, system architecture, specialised technologies and enabling infrastructure.

At one level are the models and capabilities that process information: foundation models, specialised models, reasoning systems and multimodal systems.

Around them are technologies that affect how AI interacts with people and information: conversational and voice systems, search and research systems, coding systems, media-generation technologies, memory and personalisation.

Another development concerns **action**. Agentic systems increasingly allow AI to interact with software and undertake bounded digital activity, while Physical AI & Robotics connects AI to machines capable of sensing and acting in the physical world.

The location and scale of computation are changing as well. Edge & On-Device AI moves increasing amounts of inference closer to users, sensors and machines, while large-scale AI Infrastructure & Compute continues to expand and specialise. These are not contradictory directions. They demonstrate that AI computation is becoming more distributed across a technological environment.

Data remains another foundational layer. AI Data & Synthetic Data examines the material from which systems learn, while World Models & Simulation illustrates the increasing ability of AI systems to learn, represent and generate environments that can support training, planning and action.

These categories necessarily overlap.

A multimodal foundation model may also reason. A research system may use retrieval, memory and agentic actions. A coding system may itself be an agent. A robot may combine multimodal perception, reasoning, world modelling and specialised on-device inference. Synthetic data may be generated through simulation and then used to train a physical AI system.

The elements therefore should not be understood as fifteen sealed technological compartments.

They are fifteen perspectives on significant parts of an increasingly interconnected technological landscape.

That distinction also explains why this publication does not attempt to identify one overall maturity level for AI.

The evidence across the fifteen elements shows several technological generations operating simultaneously.

Some forms of specialised AI have been operational for many years. Speech recognition, computer vision, recommendation, prediction, classification and other bounded capabilities are deeply established in particular applications. Edge inference likewise has a long operational history, while large-scale computing infrastructure is already foundational to modern AI.

Generative AI and foundation models have reached substantial commercial availability and real-world use, but their capabilities remain uneven. Multimodal systems are increasingly central. Conversational interfaces are becoming more natural. Coding assistance has become mainstream within parts of software development. AI-enhanced search has reached consumer scale.

Other capabilities are materially significant but less uniformly mature.

Reasoning systems demonstrate strong performance in selected bounded and verifiable domains while remaining less dependable across ambiguous real-world problems. Agentic systems can increasingly undertake multi-step digital activities, but dependable long-horizon autonomy remains more difficult. Persistent AI memory is operational but continues to face problems of relevance, accuracy, updating and user control.

Physical AI shows perhaps the widest maturity range. Specialised industrial and autonomous machines are already extensively deployed, while adaptable general-purpose robots remain much less mature.

World models similarly combine established simulation concepts with newer generative approaches whose broader role in AI planning, robotics and interactive environments is still developing.

The result is not a simple technological ladder.

Research capability, product availability, operational deployment and mature effectiveness coexist within individual categories and sometimes within the same system.

Another important feature of the current landscape is **convergence**.

Capabilities that were once developed more separately are increasingly being combined.

Language systems incorporate vision and audio. Conversational systems connect to tools. Search systems undertake multi-stage research. Coding assistants become agents. Memory gives systems greater continuity. World models connect simulation to planning. Robots combine perception, reasoning and action.

This convergence can create the appearance of increasingly general capability.

It should be interpreted carefully.

Combining several capable components does not automatically eliminate the limitations of those components. A system that can see, speak, remember, reason and act may still misinterpret information, retrieve the wrong memory, make an incorrect inference or execute an inappropriate action.

As capability becomes integrated, system-level reliability becomes increasingly important.

Infrastructure provides a further qualification.

Advanced AI does not exist independently of processors, memory, networking, electricity, cooling, data centres, devices and communications systems. The rapid expansion of AI capability is occurring alongside major expansion and redesign of the physical infrastructure required to train and operate AI.

At the same time, more AI computation is moving onto phones, computers, vehicles, industrial equipment and other devices.

The Technology Landscape is therefore simultaneously becoming more computationally concentrated at its largest scale and more distributed at its point of use.

Geography matters throughout.

Different countries and regions occupy different positions in model development, semiconductor manufacturing, robotics, consumer devices, data-centre infrastructure, language technology, research and deployment. Access to compute, capital, electricity, data, skills and digital infrastructure remains highly uneven.

The current AI Technology Landscape is consequently global but not geographically uniform.

Most importantly, the fifteen elements demonstrate why technological visibility should not be confused with technological maturity.

The technologies attracting the greatest public attention are not necessarily those most deeply deployed. Long-established specialised systems can have substantial economic and operational significance while attracting relatively little attention. Conversely, a highly visible demonstration may represent an important frontier without yet representing normal operational capability.

The Status as at **30 August 2026** is therefore one of substantial but uneven technological development.

AI already comprises mature specialised systems, widely deployed generative technologies, increasingly multimodal and conversational interfaces, significant reasoning and agentic capabilities, rapidly advancing media and coding systems, expanding memory and personalisation, established edge intelligence, developing world models and physical AI, increasingly sophisticated data practices, and a rapidly expanding infrastructure base.

What is emerging is not one AI technology replacing all others.

It is an increasingly interconnected technological ecosystem in which different forms of AI provide different capabilities, operate at different levels of maturity and increasingly combine within larger systems.

That complexity is not something to simplify away.

It is the Technology Landscape Status.

1. Physical AI & Robotics

Status as at 30 August 2026

Status snapshot

Physical AI & Robotics spans a wide maturity range. At its established end, millions of industrial robots operate worldwide, while specialised autonomous systems are already deployed in warehousing and logistics, mining, agriculture, healthcare, cleaning and other structured environments. These systems demonstrate that physical AI is already economically significant, although much of its current value comes from machines designed for particular tasks rather than general-purpose robots.

The principal direction of development is towards **greater adaptability and autonomy**. Advances in perception, multimodal and vision-language-action models, sensors and computing are enabling robots to interpret changing environments, accept higher-level instructions and perform less rigidly programmed tasks. Operational pilots and limited commercial deployments—including humanoid systems in manufacturing and logistics—show meaningful progress beyond laboratory demonstrations.

Maturity nevertheless differs substantially across the landscape. Established industrial and specialised robotics should not be equated with emerging general-purpose embodied intelligence. Humanoid robots can perform increasingly credible physical tasks, but current deployments remain narrow compared with mature specialised automation, and demonstrations do not establish reliable or economical operation at scale. Dexterity, robustness in unstructured environments, safety, training data, energy, hardware cost and integration remain important constraints.

As at 30 August 2026, Physical AI is already extensively deployed in specialised forms while progressing towards more adaptable machines; general-purpose physical intelligence remains a developing rather than established capability.

1. What is it?

Physical AI describes artificial intelligence that is connected to machines capable of sensing, interpreting and acting in the physical world. Robotics is its most visible expression, but the category is broader than humanoid robots. It includes industrial robot arms, autonomous mobile robots in warehouses, self-operating mining and agricultural machinery, inspection robots, medical robots, delivery systems and emerging general-purpose machines.

The important distinction is between **digital AI**, which principally produces or processes information, and **physical AI**, whose outputs can cause something to move, manipulate an object or otherwise alter the physical environment. A language model can propose how a task might be performed; a physical AI system must also perceive its surroundings, translate an intention into physical actions, execute those actions through hardware and continually respond to what happens.

This does not mean that every robot is necessarily an AI system. Industrial robots have performed precisely programmed movements for decades. What is changing is the extent to which machine learning, computer vision, force and touch sensing, autonomous navigation and more recently multimodal and vision-language-action models are enabling machines to respond to changing conditions rather than simply repeat predetermined movements.

The resulting landscape therefore contains technologies at very different stages of maturity. Conventional industrial robotics is deeply established. Autonomous mobile robots and specialised autonomous machines have reached substantial real-world deployment in several settings. AI systems capable of more flexible manipulation and adaptation are developing rapidly. General-purpose humanoid robots remain much less mature.

The distinction matters because current public discussion can create the impression that Physical AI & Robotics began with the recent wave of humanoids. In reality, specialised robots and autonomous machines already perform a considerable amount of economically useful physical work. The International Federation of Robotics (IFR) recorded approximately **4.66 million industrial robots in operational use worldwide in 2024**, with 542,000 new installations that year. Asia accounted for 74% of new installations, Europe 16% and the Americas 9%.

At the same time, newer AI techniques are attempting something more ambitious: enabling machines to understand instructions, unfamiliar objects and changing environments with less task-specific programming. Google DeepMind's Gemini Robotics, for example, applies multimodal AI to robotic perception, reasoning and action, while its later on-device variant was designed to run robot-control capabilities locally. These systems demonstrate the direction of research and development, but their capabilities should not be confused with broad commercial autonomy.

Physical AI in August 2026 is therefore best understood not as a single technology but as a **continuum from highly mature specialised automation to experimental general-purpose embodied intelligence**.

2. What can it currently do?

Current robotic systems can perform many physical tasks extremely well when the task, environment and operating conditions are appropriately constrained.

In manufacturing, robots routinely weld, paint, assemble, cut, place components, load machines, inspect products and handle materials. Computer vision and force sensing increasingly allow these systems to accommodate variations in objects and positioning that older automation handled less easily. The scale of this activity is substantial rather than experimental: more than half a million industrial robots were installed worldwide in 2024 alone.

Mobile robots can navigate warehouses, factories and hospitals, move goods between locations and coordinate as fleets. Amazon reported in June 2025 that it had deployed its **one-millionth robot**, across a network spanning more than 300 facilities. Its DeepFleet AI model was introduced to coordinate the routes of large robot fleets, illustrating how AI is increasingly

being applied not only to individual machines but to the optimisation of many machines operating together.

Specialised autonomous machines can operate with considerably greater independence in structured outdoor environments. Caterpillar reported **827 autonomous haul trucks operating in 2025**, while autonomous haulage at a quarry deployment had moved more than two million tons. Such systems demonstrate that meaningful physical autonomy is already possible where the task is repetitive, the operating domain can be controlled and the economics justify specialised equipment.

Agricultural machinery illustrates another form of physical intelligence. John Deere's autonomous tractors combine cameras, machine vision and onboard processing to navigate fields and detect obstacles, while its See & Spray systems use machine learning and cameras to distinguish plants from weeds and activate individual spray nozzles selectively. The intelligence here is narrow but practically valuable: the machine perceives a physical condition and immediately determines a physical action.

Robots can also perform specialised medical and laboratory tasks. Robotic-assisted surgery is already widely used, although most surgical systems remain surgeon-controlled rather than autonomous. Intuitive reported an installed base of approximately **11,106 da Vinci systems at the end of 2025**, with about 3.15 million da Vinci procedures performed during the year. This is important to the Technology Landscape because it demonstrates that sophisticated physical robotic systems can achieve extensive adoption without possessing general-purpose autonomy.

The newer frontier concerns machines that can perform a wider variety of tasks from visual or natural-language instructions. Vision-language-action models attempt to connect what a robot sees and is told directly to what it does. Demonstrations now include manipulating unfamiliar objects, folding or sorting items, handling kitchen objects and combining walking with manipulation. Google DeepMind, Figure and other developers have demonstrated increasingly capable systems.

Some humanoid systems have also moved beyond laboratory demonstrations. BMW reported that Figure 02 robots worked in its Spartanburg plant during 2025, handling and positioning sheet-metal parts. Across approximately 1,250 operating hours they moved more than 90,000 components and supported production of more than 30,000 vehicles. BMW subsequently began another humanoid pilot in Leipzig involving the AEON robot.

These are significant demonstrations of operational capability. They do **not**, however, establish that humanoids can presently replace general human physical labour. The BMW use case was a defined manufacturing task in a managed environment. Even leading humanoid developers continue to focus heavily on collecting data, improving manipulation, training control models and testing relatively narrow workflows.

The present capability boundary is therefore clear. Physical AI can already be excellent at **specific physical tasks**. It is becoming more adaptable. It does not yet possess the general physical competence, reliability and robustness of people across ordinary unpredictable environments.

3. Where is it currently being applied?

The largest current applications are found where tasks are repetitive, physically demanding, hazardous, time-sensitive or capable of being sufficiently structured.

Manufacturing remains the largest established domain. Automotive, electronics, metalworking and machinery production employ robots extensively for production and material handling. China alone installed about 295,000 industrial robots in 2024—54% of the global total—and had more than two million industrial robots operating. Japan, South Korea and Germany also remain major robotic manufacturing economies. India has been growing from a smaller base.

Warehousing and logistics constitute another major area. Autonomous mobile robots transport inventory through fulfilment centres and factories, while robotic arms increasingly sort, pick and handle packages. IFR reported that transportation and logistics accounted for more than half of professional service robots sold in its 2024 supplier sample, with approximately 102,900 units. Amazon's million-robot fleet illustrates the scale that specialised warehouse automation can reach within a single global organisation.

Mining, quarrying and heavy industry provide favourable environments for autonomy because machines often perform repeated movements along known routes and removing operators from hazardous locations has direct safety value. Autonomous haulage is already commercially established at some large mines and is spreading into quarry applications.

Agriculture is applying AI to navigation, selective spraying, harvesting, milking and other tasks. Current systems remain highly application-specific, but they show how perception and robotics can be combined to reduce repetitive labour and make more precise use of agricultural inputs. John Deere has also extended its autonomy technology towards orchard spraying, construction equipment and commercial mowing, although different products occupy different stages between release and development.

Professional cleaning, hospitality and service environments contain substantial numbers of comparatively simple service robots. IFR's 2024 supplier sample recorded more than 25,000 professional cleaning robots and more than 42,000 hospitality robots. These machines generally perform narrower functions than the humanoid assistants often portrayed in public discussion.

Healthcare and laboratories use surgical, rehabilitation, diagnostic, laboratory and logistical robotics. IFR's service robotics survey recorded strong growth in medical robots in 2024, although its service-robot statistics are based on participating suppliers and are not estimates of the entire global market. Robotic-assisted surgery provides one of the clearest examples of long-established human-machine collaboration: the robot increases precision and capability while clinical responsibility remains with human professionals.

Inspection, security and infrastructure are smaller but important applications. Ground robots, drones and specialised machines can enter confined, dangerous or difficult-to-access locations, collect imagery and sensor data, inspect assets and reduce human exposure to hazards. IFR recorded particularly rapid growth from a small base in inspection and maintenance robot sales in its 2024 sample.

Humanoid robots occupy a distinctive but presently much smaller category. Their attraction is that factories, warehouses and buildings have been designed around human dimensions: stairs, shelves, tools, workstations and passageways can theoretically be used without redesigning the environment. Commercial deployment has begun in limited cases. GXO and Agility Robotics entered a multi-year agreement to deploy Digit robots in logistics after an earlier proof of concept, while BMW has undertaken substantial factory trials.

China is developing particularly rapidly as a humanoid manufacturing ecosystem. Government policy explicitly identified humanoid robotics as a strategic industry, with objectives covering core components, mass production and demonstration applications. Chinese companies have subsequently expanded production and trials. Yet evidence available immediately before the reference date also indicated a significant gap between impressive mobility demonstrations and reliable economically useful factory work.

The current application landscape therefore remains dominated by **specialised machines rather than universal robotic workers.**

4. How developed is it?

Physical AI & Robotics cannot sensibly be assigned a single maturity level.

At the most mature end are **conventional industrial robots enhanced by better sensing, software and AI.** Their commercial status is unquestionable. More than 4.6 million industrial robots were operating worldwide in 2024, and annual installations have exceeded 500,000 for four consecutive years. Standards, integrators, established suppliers and mature safety practices support this ecosystem.

A second established category comprises **specialised autonomous and mobile systems.** Warehouse robots, autonomous mining vehicles, selected agricultural systems, cleaning robots and robotic-assisted medical systems are commercially deployed, although adoption varies greatly between applications and regions.

A third, developing category consists of **more adaptive AI-enabled robots** capable of interpreting scenes, handling greater object variation or accepting higher-level instructions. Foundation models and vision-language-action models are important here. Research has demonstrated significant improvements in semantic understanding, task generalisation and manipulation, while commercially oriented systems increasingly incorporate these techniques. However, robot-specific training data remain scarce compared with the enormous amount of text and imagery available for digital AI, and real-world robustness remains a major challenge.

General-purpose humanoids sit largely in this developing-to-experimental category. They are no longer purely laboratory prototypes: some have completed substantial pilots or entered limited commercial deployment. Figure's BMW programme is particularly notable because it accumulated meaningful operating hours in real production. Agility's Digit has been commercially deployed in logistics. Chinese manufacturers are producing increasing numbers of humanoid platforms.

But the evidence does not support describing humanoids as a mature general-purpose workforce technology. IFR itself concluded that the timing and extent of mass adoption remained uncertain and that humanoids were more likely to complement rather than simply replace existing robot forms.

A useful indicator of the gap is the continuing effort devoted to collecting physical-world training data. Figure announced in August 2026 a programme to build a large and diverse physical dataset, while Apptronik established robot-training environments in which fleets collect real-world interaction data. These initiatives are evidence of technological progress—but also evidence that the learning infrastructure required for general physical competence is still being built.

The Status as at 30 August 2026 is therefore one of **deep maturity at the specialised end, rapid development in adaptive robotics, and substantial but still limited real-world progress towards general-purpose embodied machines.**

5. What is driving its development and use?

Several forces are converging.

First is the rapid improvement of **AI perception and reasoning**. Computer vision has long allowed robots to recognise objects and positions, but multimodal models increasingly combine vision, language and action. This raises the possibility that machines can be given goals rather than having every movement explicitly programmed. Google DeepMind's robotics models and similar work elsewhere illustrate the attempt to transfer some of the generalisation achieved by foundation models into physical systems.

Second is improvement in **hardware**. Cameras, lidar, force sensors, tactile sensing, actuators, batteries and edge computing continue to improve. Faster local processing is especially important because physical systems frequently must respond in fractions of a second and cannot depend entirely on cloud connectivity.

Third is **labour availability and demographics**. Robotics suppliers and users repeatedly cite shortages of workers for repetitive, physically demanding or specialised tasks. IFR identifies staff shortages as an important driver of professional service robotics. John Deere similarly cites labour availability in agriculture, construction and commercial landscaping. Ageing populations reinforce interest in both automation and healthcare robotics.

Fourth is **safety**. Removing people from dangerous mines, repetitive lifting, hazardous inspection locations or physically exhausting tasks can have value independently of labour cost.

Fifth is **productivity and operating economics**. Robots can operate consistently, execute high-volume repetitive work and sometimes extend operations across more hours. At scale, AI can also improve the orchestration of robotic fleets. Amazon states that DeepFleet improves robotic travel efficiency by coordinating movement across fulfilment facilities.

Sixth is the availability of more flexible **commercial models**. Robot-as-a-Service arrangements reduce upfront investment and shift some technical responsibility towards providers. IFR's 2024 service-robot sample showed strong growth in RaaS fleets.

Government policy is another material driver. China has explicitly made humanoid and embodied-intelligence development part of its industrial strategy. Other countries are supporting robotics through manufacturing, research, industrial-policy and defence programmes. This means technological development is shaped not only by immediate commercial returns but also by strategic competition over manufacturing capacity, AI, components and supply chains.

6. What are the present limitations?

The central limitation is that **the physical world is less forgiving than the digital world**.

A language model can produce an incorrect answer without directly moving machinery. A robot's mistake can damage equipment, spoil material or injure someone. Reliability requirements are consequently much higher.

Robots also encounter enormous physical variation. Objects deform, slip, break and obstruct one another. Lighting changes. Floors become wet. People move unpredictably. Doors stick. Tools are misplaced. A task that appears trivial to a person may require perception, balance, dexterity, force control and judgement simultaneously.

Manipulation remains especially difficult. Human hands combine extraordinary dexterity with tactile sensing and continuous feedback. Robotic grippers can be exceptionally effective when designed for a known object or process, but general handling of arbitrary objects remains substantially harder.

Training data present another constraint. Digital foundation models can learn from enormous quantities of text, images and video. High-quality physical-action data are more expensive because robots or people must generate interactions involving particular bodies, sensors and environments. Research literature continues to identify data scarcity, uncertainty, real-time execution, safety evaluation and transfer between simulation and reality as important obstacles.

Hardware economics also matter. A robot must offer sufficient productive value to justify not only its purchase price but integration, safety engineering, maintenance, energy, downtime, supervision and process redesign. A technically impressive humanoid may have little economic advantage if a conveyor, fixed robot arm or simpler mobile machine can perform the same job more reliably and cheaply.

Energy remains a practical constraint for mobile and humanoid robots. More capable bodies require motors, processors and sensors while carrying their own power source. Battery duration, charging and thermal management can therefore affect productive operating time.

Integration is another major barrier. BMW's own humanoid pilot identified requirements involving production IT, occupational safety, process management, shop-floor logistics,

communications infrastructure and physical barriers. The robot cannot be evaluated independently of the system in which it operates.

Safety and regulation also vary by application. The revised ISO 10218 standards published in 2025 establish safety requirements for industrial robots and their integration, but their scope does not encompass every category of medical, service, consumer or mobile robot. The widening range of physical AI therefore creates multiple regulatory and standards environments rather than one universal robotics regime.

Finally, highly choreographed demonstrations remain an unreliable measure of useful autonomy. Walking, dancing or handling an object once under favourable conditions can demonstrate engineering progress without establishing reliability, cycle time, maintenance requirements or economic performance over thousands of operating hours. Evidence immediately before the reference date from China's rapidly developing humanoid sector highlighted precisely this distinction: impressive hardware and demonstrations coexisted with continuing difficulty performing flexible factory work reliably.

7. What you should be aware of

The most important point is that **Physical AI is already economically significant, but its significance does not principally depend on humanoid robots.**

Industrial robots, warehouse systems, autonomous mining machines, agricultural equipment, cleaning robots and medical robotic systems demonstrate that machines need not resemble people to create substantial value. In many circumstances a specialised form is more efficient because it has been designed around the task rather than around the human body.

Humanoids nevertheless deserve attention. Human environments are already built around human dimensions, tools and workflows, so a sufficiently capable human-shaped machine could potentially operate without extensive redesign of workplaces. Evidence from BMW and logistics deployments shows that the concept has moved beyond laboratory demonstration. What remains uncertain is whether the economics, reliability and flexibility will support deployment at the scale sometimes suggested by public discussion.

Safety takes on unusual importance when AI has physical agency. Cybersecurity is also a physical-safety issue: interference with a connected robot, vehicle or machine can potentially produce consequences in the real world. Responsibility can be distributed across robot manufacturers, AI-model developers, software providers, integrators, operators and organisations using the equipment, making accountability more complex than for many conventional machines.

Human oversight remains material across much of the landscape. “Autonomous” does not always mean completely independent. Machines may operate autonomously within a defined area while humans supervise exceptions, define missions, maintain systems or intervene remotely. Surgical robotics demonstrates an even clearer distinction: highly sophisticated robotics can be widespread while humans remain directly in control.

Workforce effects are similarly unlikely to be uniform. Robotics can remove particular manual tasks, reduce physical strain and change staffing requirements while simultaneously increasing demand for maintenance, integration, engineering and operational roles. The consequences depend on the activity, organisation and labour market rather than on robotics alone.

There are also strong geographic differences. East Asia dominates industrial robot deployment, with China now the largest individual market by a substantial margin. Japan and South Korea retain major robotics capabilities. Europe has strong industrial automation and an emphasis on collaborative and safety-conscious deployment. The United States combines a large industrial market with substantial AI, warehouse robotics, autonomous-systems and humanoid investment. China's scale, supply chain and policy support make it particularly important in humanoid development.

Most importantly, **capability, deployment and maturity should not be inferred from one another**. A research system performing a new task proves that the task is possible under the test conditions. A pilot demonstrates that it can operate in a real environment. Commercial deployment shows that someone is prepared to use and pay for it. Repeated operation at scale provides substantially stronger evidence of maturity.

As at **30 August 2026**, Physical AI & Robotics therefore occupies several technological eras simultaneously. Traditional and specialised robotics already perform extensive real-world work. AI is making these machines more perceptive, autonomous and adaptable. Foundation-model approaches are beginning to broaden what robots can understand and do. Humanoid systems are producing increasingly credible operational evidence but remain far from demonstrating human-like general physical competence.

The central current development is consequently **not the arrival of a universal robot worker**. It is the gradual expansion of the range of physical activities that machines can perceive, decide upon and perform—and the movement of robotics from fixed automation towards greater adaptability and autonomy.

Evidence Notes

1. **International Federation of Robotics — World Robotics 2025, Industrial Robots**, 25 September 2025. Principal global adoption evidence: 542,000 installations during 2024; 4.664 million operational industrial robots; regional and national deployment data. [IFR World Robotics 2025](#)
2. **International Federation of Robotics — World Robotics 2025, Service Robots**, 7 October 2025. Professional service, logistics, cleaning, agricultural and medical robot evidence. IFR cautions that service-robot figures are based on supplier samples and should not be projected to the entire industry. [IFR Service Robots summary](#)
3. **IFR — Humanoid Robots: Vision and Reality**, 14 August 2025. Maturity, regional differences and IFR's assessment that mass adoption remains uncertain and humanoids are expected to complement other robots. [IFR Humanoid Robots paper summary](#)
4. **Amazon Robotics**, 30 June 2025. One-millionth robot deployed; operations across more than 300 facilities; DeepFleet AI fleet-coordination model. Company evidence, used principally as evidence of deployment scale rather than independently verified performance claims. [Amazon robotics milestone](#)

5. **Caterpillar 2025 Highlights**, published 2026. Reports 827 autonomous haul trucks operating during 2025 and deployment beyond mining into quarry operations. Company evidence. [Caterpillar autonomy highlights](#)
 6. **John Deere — Autonomous Machines**, 6 January 2025, together with current product information. Evidence concerning autonomous agriculture, construction and landscaping systems, computer vision and machine learning. [John Deere autonomous machines](#)
 7. **John Deere — Autonomous Tractor / See & Spray**. Current product evidence concerning camera-based perception, neural-network obstacle recognition and AI weed identification. [John Deere autonomous tractor](#)
 8. **Intuitive Surgical**, January 2026 and 2025 reporting. Evidence of mature robotic-assisted healthcare deployment: approximately 11,106 da Vinci systems installed and about 3.15 million da Vinci procedures during 2025. These systems should not be interpreted as autonomous surgical AI. [Intuitive 2025 results](#)
 9. **Google DeepMind — Gemini Robotics**, 12 March 2025, and **Gemini Robotics On-Device**, 24 June 2025. Evidence of development of multimodal and vision-language-action robotics models. These are capability/research-development sources rather than evidence of general commercial deployment. [Gemini Robotics](#)
 10. **BMW Group**, 9 March 2026 and February 2026 materials. Independent customer-side evidence of Figure 02's Spartanburg deployment and BMW's subsequent AEON pilot in Leipzig. Particularly useful for distinguishing real production operation from demonstrations. [BMW humanoid deployment evidence](#)
 11. **GXO Logistics / Agility Robotics**, 27 June 2024. Commercial agreement following proof-of-concept testing of Digit in warehouse operations. Evidence of early commercial humanoid deployment. [GXO–Agility agreement](#)
 12. **Figure**, January–August 2026 materials. Evidence concerning Helix 02, Figure 03 manufacturing and data collection. Treated as developer claims except where corroborated by BMW. [Figure news and development record](#)
 13. **Apptronik**, 30 June 2026. Evidence concerning dedicated robot-training environments and collection of physical-world data. Developer evidence. [Apptronik Robot Park](#)
 14. **China Ministry of Industry and Information Technology**, 2 November 2023. Primary policy evidence establishing strategic objectives for humanoid technology, production and demonstration applications. [MIIT humanoid robotics policy](#)
 15. **Reuters investigation**, 27 August 2026. Recent independent evidence concerning the gap between China's rapidly improving humanoid hardware and demonstrations and dependable factory capability.
 16. **ISO 10218-1:2025 and ISO 10218-2:2025**, February 2025. Current international industrial-robot safety standards and evidence of the importance of system integration and risk reduction. [ISO 10218-1:2025](#)
 17. **Firoozi et al., International Journal of Robotics Research**, 2025. Peer-reviewed survey of foundation models in robotics; identifies robot-data scarcity, uncertainty, safety evaluation and real-time execution among continuing challenges. [Foundation Models in Robotics survey](#)
-

2. Generative AI & Foundation Models

Status as at 30 August 2026

Status snapshot

Generative AI & Foundation Models are now an established and widely deployed part of the applied AI landscape, while remaining a rapidly developing technology. General-purpose foundation models underpin conversational assistants, coding tools, document and knowledge systems and a growing range of organisational applications. Their defining significance is their breadth: one broadly pre-trained model can be adapted or used across many tasks that previously required separate specialised systems.

The landscape is increasingly international and diverse. Frontier development remains concentrated among organisations with access to substantial compute, data, capital and technical expertise, particularly in the United States and China, while important model ecosystems have also developed in Europe, the Middle East and wider Asia. At the same time, smaller and open-weight models are widening the range of organisations able to deploy or adapt foundation-model capabilities.

Capability has advanced rapidly, but maturity is uneven. Generative systems can produce highly capable language, code and other content and are already used at substantial scale, yet fluent output remains distinct from dependable accuracy. Hallucination, uneven performance, limited transparency, data and infrastructure constraints, security and privacy concerns, and variable performance across languages and contexts remain material limitations.

Status as at 30 August 2026: the foundation-model paradigm is established; generative AI is in scaled real-world use; frontier capabilities continue to advance rapidly; and dependable effectiveness remains strongly dependent on the task, model and wider system in which the technology is deployed.

1. What is it?

Generative AI refers to artificial-intelligence systems that create new content—such as text, software code, images, audio, video or structured data—by learning patterns from existing data and producing new outputs in response to instructions or other inputs.

Foundation models are related, but they are not synonymous with generative AI. A foundation model is a comparatively general model trained on broad data at substantial scale and designed so that the capabilities learned during that training can be adapted to many different tasks. Instead of building a completely separate model for every application, organisations can use one broadly capable underlying model and adapt, prompt or integrate it for multiple purposes.

Large language models (LLMs) form an important subset of foundation models. They are principally trained on language and related symbolic information and can generate and interpret text, answer questions, summarise material, translate languages, write software code and perform many other language-oriented tasks. Some foundation models operate primarily on

images, biological information or other forms of data. Increasingly, major models combine several forms of information, although the technological implications of that development belong principally within **Multimodal AI**.

The distinctions therefore matter. **Foundation model** principally describes the underlying technological paradigm: broad pre-training followed by adaptation or use across many tasks. **Large language model** describes a major class of models centred on language. **Generative AI** describes the capability to produce new content. Not every generative model is necessarily a foundation model, and foundation models can also be used for classification, representation, prediction and other purposes that do not principally involve generating content.

The modern foundation-model paradigm depends heavily on **pre-training**. A model is exposed to extremely large collections of data and learns statistical relationships within them. In language models, a central training objective has historically involved predicting parts of sequences from surrounding information. The model does not store a conventional database of sentences and retrieve the appropriate one when questioned. Training instead adjusts very large numbers of internal numerical parameters so that the system learns patterns involving language, concepts and relationships.

Pre-training produces a broadly capable but unfinished model. **Post-training** is increasingly important. Developers use techniques including instruction tuning, preference optimisation, reinforcement-learning approaches, specialised data, safety training and other methods to make models more useful, controllable or capable on particular tasks. By 2026, significant capability improvements are being achieved not simply by making initial training runs larger, but through better data, architectures, post-training and greater computation while a model is producing an answer.

This helps explain why the current landscape cannot be understood simply by counting parameters. Model size remains relevant, but architecture, training-data quality, training compute, post-training, inference techniques and system design can materially alter capability. Smaller or more efficient models increasingly achieve performance that previously required much larger systems.

The underlying models are also only one component of most usable generative-AI systems. Applications frequently add document retrieval, search, databases, safety controls, user interfaces, software tools and organisational information around the model. What a user experiences as “the AI” is consequently often a system built around a foundation model rather than the model operating alone.

Generative AI & Foundation Models in August 2026 therefore describes a broad technological platform: large pre-trained models providing adaptable capabilities upon which a rapidly expanding range of AI systems is being constructed.

2. What can it currently do?

Current foundation models can perform an unusually wide range of information-processing and content-generation tasks without separate conventional software being programmed for each one.

Language models can draft, revise, summarise and classify text; explain documents; answer questions; translate between languages; extract information; transform unstructured material into structured formats; generate computer code; analyse code; and hold extended natural-language conversations. They can vary tone and format, combine information supplied in a prompt and produce outputs that resemble many forms of professional and everyday writing.

Their ability to work from instructions is particularly important. Earlier machine-learning systems often required a task to be defined and a model specifically trained for it. Foundation models can frequently perform a new task from an instruction and a few examples supplied at the time of use. The same underlying model can therefore support activities ranging from document summarisation to programming assistance without being retrained from the beginning.

Models can also be adapted to specialised domains using additional training or connected to authoritative information sources. This allows foundation models to support applications involving scientific literature, legal material, organisational documents, technical manuals, healthcare information and many other knowledge domains. The resulting effectiveness depends substantially on the surrounding system and the quality of the information provided to it.

Software development has become one of the clearest demonstrated capability areas. Generative models can produce functions, explain existing code, suggest corrections, translate between programming languages and assist with testing and documentation. Performance on software-engineering evaluations improved sharply through 2025 and 2026. This establishes meaningful technical capability, although benchmark success does not establish that generated software is correct, secure or appropriate for production without checking.

Models can also perform significant analytical and problem-solving tasks. Leading general-purpose systems have achieved very high performance on standardised tests in mathematics, science and professional subjects. The 2026 International AI Safety Report noted capabilities extending to graduate-level science questions and competition-level mathematics. Those developments overlap substantially with the separate **Reasoning AI** landscape and should not be interpreted as evidence that ordinary generative models reason dependably across all circumstances.

The capability profile remains notably uneven. A model may solve an extremely difficult mathematics problem and subsequently make an elementary factual, spatial or commonsense error. Stanford's 2026 AI Index described this as a “jagged” capability frontier. Performance can change materially when a question is reworded, distracting information is introduced or a task moves beyond the conditions represented in evaluations.

Generation itself remains probabilistic. A language model normally produces a plausible continuation conditioned on its instructions and context; it does not inherently perform a conventional database lookup or independently establish that every generated assertion is true. This is why fluent output and factual reliability remain different properties.

Current systems can consequently display impressive breadth while remaining unreliable in particular instances. Their demonstrated capability is substantial, but dependable performance still depends upon the task, model, instructions, supporting information, verification

mechanisms and wider system in which the model is used.

3. Where is it currently being applied?

Generative AI has moved into widespread application because language, documents, software and other information are embedded throughout economic and social activity.

One of the largest application categories is **general knowledge and productivity assistance**. Individuals use conversational systems for drafting, explanation, summarisation, translation, brainstorming, research assistance, learning and information organisation. Generative functions are also increasingly incorporated directly into office software, search services, communications platforms and operating environments.

Software development is another major application. Generative coding assistants are used to suggest code, explain unfamiliar programs, generate tests and documentation and assist debugging. More capable systems can undertake larger bounded software tasks, although the transition from coding assistance towards autonomous software action principally belongs within the separate **AI Agents & Agentic Systems** landscape.

Within organisations, foundation models increasingly support **customer service and communications**, document processing, knowledge retrieval, marketing-content preparation, sales support, administrative work, finance-related analysis and internal information access. Many of these applications combine the model with proprietary organisational data rather than relying solely on information learned during pre-training.

Professional and knowledge-intensive work is a particularly significant area because foundation models can manipulate the natural-language and documentary material on which many such activities depend. Legal, accounting, consulting, financial and technical organisations are integrating generative systems into research, drafting, document review, summarisation and analysis. Use remains more constrained where outputs directly affect regulated or high-stakes decisions.

Education and research have experienced extremely rapid uptake. Students and teachers use generative systems for explanation, tutoring, preparation and writing support, while researchers use them for literature exploration, programming, language assistance and parts of scientific workflows. Stanford's 2026 AI Index reported that four in five university students used generative AI, illustrating availability and behavioural adoption rather than establishing educational effectiveness. UNESCO has similarly documented both extensive opportunities and unresolved issues involving accuracy, assessment, privacy and institutional preparedness.

Science and healthcare increasingly use foundation-model approaches for literature analysis, research support, biomedical information processing and specialised scientific modelling. Foundation models trained on biological sequences, molecules and other scientific data extend the paradigm well beyond conversational language models. Some systems achieve strong research or benchmark results, but evidence of scientific capability should be distinguished from validated clinical or operational effectiveness.

Foundation models are also becoming part of the technological infrastructure behind numerous specialist AI applications. Organisations increasingly access models through cloud services and application-programming interfaces, deploy open-weight models on their own infrastructure or adapt smaller models for particular environments.

This application landscape is internationally distributed. Major frontier-model developers remain concentrated, particularly in the United States and China, but substantial model ecosystems also exist in Europe, the Middle East and wider Asia. China's Alibaba, DeepSeek and other developers have established globally competitive foundation-model capabilities. European developers contribute important commercial and open-weight models. The UAE's Technology Innovation Institute has developed the Falcon family, including models directed towards Arabic-language capability. Similar efforts in other regions reflect demand for models better aligned with local languages, culture, regulation, infrastructure and data sovereignty.

Access nevertheless remains unequal. Model availability does not mean comparable adoption: connectivity, computing infrastructure, language coverage, income, digital skills, organisational capacity and regulation all influence who can use these technologies effectively.

4. How developed is it?

Generative AI & Foundation Models cannot sensibly be given one maturity classification.

At the most established end, **general conversational and text-generation capability is a scaled technology**. Generative systems have hundreds of millions of users, are embedded within widely used digital services and are increasingly incorporated into organisational software. Stanford's 2026 AI Index estimated that generative AI had reached approximately 53% population adoption within three years under the adoption measure used in its analysis.

Organisational deployment is also substantial. The same report found AI use in 88% of surveyed organisations during 2025 and generative AI use in at least one business function in approximately 70%. These are survey measures rather than a census of global enterprises, but they establish that generative AI has progressed far beyond experimental availability.

A second relatively established layer consists of **foundation-model access as technological infrastructure**. Models are routinely supplied through cloud platforms, APIs, enterprise software and downloadable weights. An extensive ecosystem of model hosting, adaptation, evaluation and application development now exists.

A third layer comprises **domain-adapted and organisational applications**. These are growing rapidly but maturity varies considerably. Document assistance or low-risk drafting may be operationally routine, while applications affecting medical, legal, financial or other consequential decisions require substantially higher reliability and governance. The same underlying model can therefore be mature enough for one use and insufficiently dependable for another.

The technological frontier remains highly dynamic. Model performance has continued to improve in coding, mathematics, science and instruction following, but development increasingly involves several interacting routes rather than parameter scaling alone. Better data

selection, specialised architectures, mixture-of-experts approaches, synthetic data, post-training and additional computation during inference can all alter capability and efficiency.

The economics are also changing. The cost of obtaining a given level of model performance fell dramatically during the first half of the decade, while smaller models became substantially more capable. This supports broader deployment, including situations where organisations cannot justify using the largest frontier systems.

Open-weight models constitute another important maturity dimension. Developers in the United States, China, Europe, the Middle East and elsewhere release model weights that can be downloaded, adapted and in many cases run privately. “Open” nevertheless has varying meanings: availability of weights does not necessarily include release of training data, training code or complete development information, and licences can impose differing restrictions.

The frontier-model-development layer is much more concentrated. Stanford reported that industry produced more than 90% of notable AI models in 2025. Developing the most capable systems requires very large quantities of specialised computing infrastructure, data, engineering expertise and capital. The International AI Safety Report estimated that development of leading general-purpose systems from scratch can cost hundreds of millions of US dollars.

Geographically, the frontier is no longer adequately described as a solely American phenomenon. Stanford found that the performance gap between leading US and Chinese models had effectively closed on several major evaluations by early 2026, while US organisations continued to produce more notable frontier models. Other regions are developing important smaller, specialist, open and sovereign-model ecosystems.

The current Status is therefore one of **scaled generative use built upon a technologically mature foundation-model paradigm, combined with continuing rapid change at the capability frontier and highly uneven maturity between applications.**

5. What is driving its development and use?

The first major driver is the unusual **generality of the foundation-model approach**. The cost of developing one large underlying model can be spread across many applications. Organisations do not necessarily need to train a new model from the beginning for every task.

A second driver is continuing **capability improvement**. Larger and better-curated training datasets, increased training compute, architectural improvements and increasingly sophisticated post-training have expanded what models can do. By 2026, post-training and additional computation during use have become particularly important routes to higher performance.

Third is **falling cost at a given capability level**. Hardware has improved, inference has become more efficient, competition has intensified and increasingly capable smaller models have become available. Stanford's 2025 AI Index calculated that the cost of querying a model at approximately GPT-3.5-level benchmark performance fell more than 280-fold between late

2022 and late 2024. Continued model and infrastructure development has reinforced the broader efficiency trend.

Fourth is the simplicity of the **natural-language interface**. Users can instruct many generative systems using ordinary language rather than conventional programming. This reduces the expertise required to experiment with AI and allows a single system to serve users performing very different activities.

Fifth is an expanding **software and infrastructure ecosystem**. Cloud platforms, APIs, open-weight repositories, enterprise applications and development frameworks allow organisations to use sophisticated models without creating frontier models themselves. At the same time, smaller models and local deployment offer alternatives where cost, latency, privacy or control matter.

Sixth is intense **commercial and geopolitical competition**. Private investment, cloud infrastructure expansion and competition among developers are sustaining rapid technical development. Governments are also supporting domestic AI capacity for economic, scientific, linguistic and strategic reasons. “Sovereign AI” initiatives and regional foundation models reflect concern that languages, data, infrastructure and technological capability should not depend entirely upon a small number of foreign providers.

Finally, demand is reinforced by the fact that generative models operate on information itself. Text, software, documents and communications occur across almost every industry and business function. Unlike a technology confined to one physical process, foundation models can therefore diffuse across many activities through existing digital infrastructure.

6. What are the present limitations?

The most visible technical limitation remains **reliability**. Generative models can produce statements that are fluent, detailed and incorrect. Hallucinations have been reduced in some systems and circumstances, but they have not been eliminated. The International AI Safety Report concluded that no current combination of methods guarantees the reliability required across critical applications.

Capability is also **jagged and difficult to predict**. Benchmark excellence in one area cannot safely be extrapolated to another. Models can fail on apparently simple tasks, react unexpectedly to irrelevant information and produce different answers to similar questions. This complicates evaluation and makes broad claims of human-equivalent competence misleading.

The technology also inherits limitations from its **training data**. Data may contain inaccuracies, social biases, duplication, outdated information and uneven representation of languages, regions and cultures. Stanford's 2026 AI Index found continuing performance differences between English and lower-resource languages and even between standard languages and regional dialects.

Data availability itself is becoming an important development constraint. High-quality human-generated text and other training material are finite, while rights to use some material are contested. Synthetic data is increasingly incorporated into training and post-training, but

research indicates that its usefulness depends strongly on how it is generated, selected and mixed with other data.

Transparency is limited. Frontier developers increasingly disclose less about training datasets, parameter counts, compute and training procedures. Stanford's Foundation Model Transparency Index recorded a substantial decline in average developer transparency between 2024 and 2025. This makes independent evaluation of data provenance, environmental impacts, limitations and some safety claims more difficult.

Compute and infrastructure requirements remain substantial at the frontier. Global AI computing capacity has expanded rapidly, but advanced accelerators, data centres, electricity, cooling infrastructure and capital are concentrated. This concentration affects who can train frontier models and creates dependencies in semiconductor and cloud supply chains.

Environmental effects consequently matter. Although hardware and software are becoming more efficient, model scale and aggregate demand have grown quickly enough for training and inference to consume increasing amounts of electricity and water. Stanford estimated very large differences in training emissions between individual frontier models, demonstrating that architecture, hardware and energy sources materially affect environmental impact rather than model size alone determining it.

There are also significant **privacy, security and intellectual-property constraints**. Inputs supplied by users can contain confidential or personal information. Models may be attacked through adversarial prompts or manipulated through information supplied to surrounding systems. Generated software can contain vulnerabilities. Questions concerning copyrighted training material, output similarity, ownership and attribution remain legally and commercially important across jurisdictions.

Evaluation itself remains incomplete. Standard benchmarks can become saturated, contaminated by training data or poorly correlated with real-world performance. New evaluations are continually required as models improve. Responsible-AI evaluation has not advanced as consistently as capability evaluation, and independent access to leading proprietary systems may be restricted.

Finally, the **model is not the complete application**. Real effectiveness depends upon data pipelines, retrieval systems, workflow design, software integration, human review, security controls, organisational processes and the consequences of error. Improving the underlying foundation model does not automatically resolve weaknesses elsewhere in that system.

7. What you should be aware of

The most important point is that generative AI is simultaneously a **widely deployed technology and a still rapidly developing one**.

The foundation-model paradigm is already established. Broadly pre-trained models are now a core part of the applied AI landscape, and generative capabilities have entered everyday consumer software and organisational activity at considerable scale. It would therefore be inaccurate to describe generative AI principally as experimental.

It would be equally inaccurate to infer from widespread availability that every application is mature. Drafting a routine document, generating production software, answering a medical question and supporting a legally consequential decision impose very different requirements for accuracy, evidence and accountability.

A second important distinction is between **model capability and system capability**. Retrieval from authoritative documents, access to software tools, additional databases and validation processes can make an application much more useful than the underlying model operating alone. Conversely, a powerful model can perform poorly when embedded in inadequate processes or supplied with poor information.

Third, **fluency should not be mistaken for understanding or truth**. Generative models are specifically designed to produce plausible outputs. Their ability to express an answer confidently is therefore weak evidence that the answer is correct. Verification remains especially important where errors have material consequences.

Fourth, **benchmark performance is evidence, but not sufficient evidence of dependable real-world performance**. The rapid improvement of model scores is technologically significant, yet the continuing jaggedness of capability means that performance cannot be represented adequately by one leaderboard number.

Fifth, the technology landscape is becoming more geographically diverse. The United States remains central to frontier development, infrastructure and investment, but China has developed models competitive with the strongest global systems. Europe has significant model developers and a distinctive regulatory environment. The Middle East and wider Asia are developing regional and language-specific foundation-model capacity. This diversification matters because models reflect the data, languages, institutions and technical infrastructure from which they are developed.

Sixth, the terms **open, open-source and open-weight** should not automatically be treated as equivalent. A model's weights may be downloadable while its training data, development process or code remain undisclosed. The practical freedoms provided by licences also differ.

Finally, generative AI should not be interpreted as a single march towards one universal model. The landscape increasingly includes very large frontier models, smaller efficient models, proprietary services, open-weight systems, domain-adapted models and models designed around particular languages or deployment environments.

As at 30 August 2026, the central technological development is therefore larger than the success of conversational chatbots. **Foundation models have established a new general-purpose layer of AI infrastructure: systems can be pre-trained broadly, adapted repeatedly and used across activities that previously required many separate specialised technologies.**

That generality explains both their unusually rapid diffusion and their central qualification. A broadly capable model can be useful in many places without being completely dependable in any arbitrary circumstance. The current Status is consequently one of remarkable breadth, substantial real-world adoption and continuing technical progress—alongside important limitations in reliability, transparency, evidence, infrastructure and control.

Evidence Notes

1. **International AI Safety Report 2026 — 3 February 2026.** Major international synthesis involving more than 100 experts and contributors nominated by more than 30 countries and international organisations. Particularly useful for current general-purpose-AI capability, post-training, development cost, jagged performance, hallucination, evaluation gaps and reliability. Its principal focus is advanced general-purpose systems and emerging safety risks rather than overall commercial adoption.
2. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Research and Development.** Principal evidence concerning the international model-development landscape, industry concentration, US-China development, model transparency, compute capacity, infrastructure and environmental effects. Reports that industry produced more than 90% of notable models in 2025 and documents increasing non-disclosure of model-development information.
3. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Technical Performance.** Evidence concerning narrowing differences among leading developers, open-versus-closed model performance and the convergence of leading US and Chinese systems. Benchmark evidence should not be interpreted as general real-world effectiveness.
4. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Economy.** Evidence concerning organisational and population adoption, consumer use and the widening geographical spread of generative AI. Survey and estimated adoption measures vary in methodology and should not be interpreted as a census of all organisations or individuals.
5. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Responsible AI.** Evidence concerning hallucinations, uneven language capability, responsible-AI evaluation, governance practices, adversarial robustness and increasing documented AI incidents.
6. **Stanford Center for Research on Foundation Models — Foundation Model Transparency Index 2025.** Independent assessment of major foundation-model developers across 100 transparency indicators. Found average transparency falling markedly from the 2024 assessment, with continuing weaknesses concerning training data, compute and downstream impacts. The index measures disclosed information rather than underlying model quality or safety.
7. **NIST — Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 26 July 2024; updated 8 April 2026.** Authoritative cross-sector framework establishing useful distinctions concerning generative AI and generative foundation models and documenting risks associated with model development and deployment. NIST notes explicitly that not all generative AI is derived from foundation models.
8. **European Commission — General-Purpose AI obligations and guidelines under the EU AI Act, obligations applicable from 2 August 2025.** Primary regulatory evidence concerning general-purpose-model documentation, copyright policies, training-content summaries and additional obligations for models considered to pose systemic risk. Relevant principally to providers operating within the EU legal framework rather than a universal definition of foundation models.
9. **Stanford AI Index Report 2025 — Research and Development.** Evidence concerning the rapid decline in inference cost, improvement in hardware price-performance and energy efficiency, and continued growth in model-training compute. Its widely cited estimate shows a more than 280-fold fall between November 2022 and October 2024 in the inference cost required to reach approximately GPT-3.5-level MMLU performance.
10. **UNESCO — Guidance for Generative AI in Education and Research, 7 September 2023; updated 16 January 2026.** International evidence concerning the rapid diffusion of generative AI into education and research and continuing concerns involving privacy, institutional preparedness, validation and equitable access.
11. **Technology Innovation Institute, UAE — Falcon model materials, 2025–2026.** Evidence of foundation-model development outside the principal US, Chinese and European ecosystems and of efforts directed towards Arabic-language and more computationally efficient models. Developer evidence is used principally

to establish the existence and orientation of these systems rather than independently verify comparative performance claims.

12. **Nature, 21 January 2025, together with subsequent research reporting on hallucination.** Independent scientific reporting concerning continued factual fabrication by LLMs and attempts to reduce its frequency. Useful evidence that improved model capability has not eliminated hallucination as an operational concern.
 13. **Qin et al. — Scaling Laws of Synthetic Data for Language Models, March 2025, together with subsequent large-scale empirical work.** Research evidence concerning the growing use and potential of synthetic training data while showing that effectiveness depends on data type, mixture and training conditions. Synthetic data should therefore not be treated as a simple unlimited substitute for high-quality human-generated training material.
-

3. Reasoning AI

Status as at 30 August 2026

Status snapshot

Reasoning AI is the increasingly important capability of AI systems to perform structured problem-solving, inference, deliberation, mathematical or logical processing, planning and verification rather than simply generate an immediate response. As at 30 August 2026, it is commercially available within major general-purpose AI systems and has demonstrated particularly strong capability in bounded, verifiable domains such as mathematics, programming and formal problem-solving.

Its maturity is uneven. Reasoning assistance is already widely available as a product capability, while performance on selected mathematical and coding tasks has reached exceptionally high levels. Reasoning is also being applied to research, technical analysis and professional knowledge work. However, dependable open-ended reasoning, unfamiliar problem formulation and longer-horizon planning remain materially less mature.

The principal development direction is towards combining improved post-training with greater inference-time computation, multiple candidate solutions, external tools and verification. This is shifting reasoning from producing plausible step-by-step answers towards systems that can search, test and revise possible solutions.

The central qualification is that benchmark achievement is not equivalent to general reasoning reliability. Performance can change substantially with problem formulation, available tools, computation budget and evaluation method, and longer reasoning does not invariably improve an answer. As at 30 August 2026, Reasoning AI represents a substantial advance in structured AI problem-solving, but not yet universally dependable machine reasoning.

1. What is it?

Reasoning AI describes artificial-intelligence systems and techniques designed to perform tasks that require more than straightforward recognition, retrieval or fluent generation. Such tasks may involve breaking a problem into parts, drawing inferences from evidence, applying rules, planning a sequence, comparing alternatives, checking intermediate results, revising an approach or constructing a mathematical, logical or scientific argument.

The term does not identify one single technology. Much contemporary Reasoning AI is built on general-purpose foundation models, particularly large language models (LLMs), but reasoning techniques can also involve specialised mathematical systems, symbolic solvers, theorem provers, search algorithms, verifiers and combinations of neural and rule-based methods. Reasoning AI therefore overlaps with Generative AI & Foundation Models but is better understood as a particular capability and system-development direction rather than as another name for an LLM.

A useful distinction is between producing an answer that appears reasoned and reliably solving a reasoning problem. Language models have long been able to generate convincing step-by-step explanations. The more recent development is that particular training and inference techniques have materially improved performance on tasks for which several intermediate decisions or computations are needed before a correct answer can be produced.

One important approach is to allocate additional computation after a user presents a problem. Instead of immediately generating the final answer, a model may perform additional internal computation, generate candidate approaches, reconsider earlier steps, use tools or check candidate solutions. This is variously described as extended thinking, deliberation or **test-time/inference-time scaling**. OpenAI, Anthropic, Google DeepMind, DeepSeek, Alibaba's Qwen programme and other developers have all incorporated forms of extended reasoning into publicly available models or systems.

Training has changed as well. Reinforcement learning—training a model by rewarding successful behaviour—has become particularly important where answers can be checked relatively objectively. Mathematics, coding and formal logic are favourable examples because a numerical answer, program execution or formal proof can often provide stronger feedback than subjective human preference. DeepSeek's R1 work demonstrated publicly that substantial reasoning behaviours could be developed using large-scale reinforcement learning, while OpenAI similarly reported continuing performance gains as reinforcement-learning and inference-time computation were increased.

Verification is consequently becoming part of the reasoning landscape. Rather than expecting one reasoning path to be correct, systems can generate several candidates and use rules, executable tests, formal methods or another model to identify better answers. Research also explores process evaluation, in which intermediate reasoning steps are assessed rather than only the final response.

None of this establishes that present systems reason in the same way as people. A visible “chain of thought” is generated model output and should not automatically be interpreted as a transparent record of an underlying human-like mental process. Research has shown that stated reasoning can sometimes be an incomplete or unfaithful explanation of what caused a model to reach its answer.

Reasoning AI as at August 2026 is therefore best understood operationally: AI systems are becoming substantially better at solving certain problems requiring multi-step inference, computation, search, revision and verification. How closely the mechanisms correspond to human reasoning remains a separate scientific question.

2. What can it currently do?

The clearest improvements are visible in mathematics, programming, scientific problem-solving and other tasks where correctness can be evaluated reasonably clearly.

Mathematical reasoning has advanced particularly rapidly. Systems can solve complex algebra, geometry, combinatorics and number-theory problems, derive proofs in natural language and, when connected to formal tools, construct or verify machine-checkable arguments. In 2025 an

advanced version of Google DeepMind's Gemini with Deep Think achieved the gold-medal threshold on that year's International Mathematical Olympiad, solving five of six problems and receiving 35 of 42 points. The solutions were officially graded under the same criteria applied to human contestants. This represented a substantial advance over earlier mathematical AI systems that required expert translation into formal languages and much longer computation.

By 2026, reasoning systems were also being tested on research-level scientific and mathematical problems. Google DeepMind reported systems that generate candidate mathematical solutions, apply a verifier, revise the result and, importantly, sometimes conclude that they cannot solve a problem. Such examples demonstrate the growing usefulness of iterative generation-and-verification approaches, although developer-reported research results do not establish that independent scientific discovery can generally be delegated to AI.

Programming is another strong domain. Reasoning-oriented models can analyse existing code, diagnose faults, formulate implementation approaches, write and execute code, interpret test results and revise unsuccessful solutions. Tool access matters considerably: an AI that can run code, inspect a repository and test its own output has capabilities different from a model answering a programming question from text alone. This distinction between model reasoning and the wider tool-enabled system is increasingly important.

Logical and deductive tasks have also improved. Models can follow chains of conditions, identify implications, analyse structured arguments and solve many puzzles that defeat earlier language models. Hybrid systems can go further by using an LLM to translate a natural-language problem into formal representations that a symbolic solver executes. Such combinations can improve reliability because the language model handles interpretation while deterministic software performs formal deduction. Research published in July 2026 continued to investigate the relative strengths of chain-of-thought and solver-augmented approaches.

Reasoning systems can also evaluate alternatives and construct multi-stage answers in less formal settings. Examples include analysing documents, synthesising evidence, comparing hypotheses, working through technical questions and deciding when to search for information or use a computational tool. This has increased the usefulness of reasoning models for research assistance and complex professional tasks.

However, performance is uneven. A model that solves an exceptionally difficult Olympiad problem may fail when an easier mathematical problem is embedded in an unfamiliar real-world description. The 2026 CORE-MATH study evaluated 61 proprietary and open models on benchmark mathematics transformed into contextual scenarios. Performance fell materially when models first had to formulate the mathematical problem from the surrounding situation, showing that identifying the correct problem can remain as difficult as solving it.

Planning remains another significant weakness. ICLR 2026's ACPBench Hard examined open-ended problems involving actions, changing states and planning. With limited exceptions, tested frontier and reasoning models scored below 65%, and researchers concluded that the systems remained substantially short of reliable planning capability on these tasks.

Nor does extra “thinking” invariably produce better answers. More inference-time computation can markedly improve difficult reasoning under some conditions, but research has found diminishing returns, overthinking and task categories in which longer reasoning can reinforce an incorrect approach or increase hallucination. Contemporary test-time scaling consequently

includes not only making one reasoning chain longer, but generating several candidates, searching different solution paths, voting between outputs or verifying them.

The current capability is therefore substantial but bounded: AI can solve classes of reasoning problems that were difficult or inaccessible to earlier language models, sometimes at very high levels of benchmark performance, while still failing unpredictably on unfamiliar formulations, planning tasks and apparently simpler problems.

3. Where is it currently being applied?

Reasoning capability is increasingly incorporated into general AI services rather than existing only in research systems. Users of widely available AI assistants can invoke systems that devote additional computation to difficult questions, while developers can access reasoning-oriented models through APIs and incorporate them into applications. Product availability is consequently well established even though dependable use varies greatly by task.

Software development is among the most visible applications. Reasoning models assist with architecture, coding, debugging, repository analysis, testing and complex technical questions. When connected to development environments they can iterate against compiler output or tests rather than simply suggest code. This area overlaps with AI Agents & Agentic Systems when the system begins executing extended sequences of actions; the underlying ability to diagnose a problem, choose an approach and revise it remains a Reasoning AI capability.

Scientific and technical research is another developing application. Mathematics provides particularly strong evidence because answers and proofs can often be checked. Experimental systems are also being applied to physics, chemistry, engineering and computer science, with researchers using AI to explore hypotheses, solve technical problems, examine candidate proofs and assist scientific analysis. Google DeepMind's 2026 research provides examples of expert-supervised use on professional mathematical and scientific questions, but remains stronger evidence of advancing technical capability than of widespread scientific adoption.

Professional knowledge work is beginning to use reasoning capability for document analysis, structured research, quantitative questions and the comparison of alternative interpretations. Finance illustrates both the potential and the need for caution. A 2026 study comparing reasoning-enhanced and conventional LLMs on financial-news sentiment and earnings-direction prediction found that the reasoning-oriented systems did not outperform ordinary models on those tasks; model scale was more predictive of performance. This is useful evidence that stronger reasoning benchmarks do not automatically translate into superior domain outcomes.

Healthcare research offers similar examples of domain-specific experimentation. Research published in August 2026 examined retrieval-augmented reasoning for text-based radiology question answering. Such work demonstrates active development of reasoning-assisted specialist workflows, but benchmark question answering should not be confused with autonomous clinical judgement.

Reasoning capability is also increasingly used inside AI systems themselves. One model may evaluate another model's answer, choose between candidate outputs, inspect a proposed

solution for errors or determine whether additional tools are required. Reasoning can therefore function as both a user-facing capability and an enabling component of more complex AI systems.

Global development is broad. United States-based developers have been prominent in commercial reasoning systems, but the technological landscape is not confined to the United States. DeepSeek's R1 work in China was especially influential because it demonstrated powerful reasoning behaviour using reinforcement-learning approaches and released model weights and technical information for wider experimentation. Alibaba's Qwen family subsequently offered both dedicated thinking models and systems able to move between thinking and non-thinking operation.

Research institutions across China, Hong Kong, Singapore, Europe and elsewhere are also contributing to reasoning evaluation, verification, mathematical reasoning and hybrid symbolic approaches. The current landscape is consequently internationally distributed, even though the most compute-intensive frontier systems remain concentrated among organisations with access to substantial models, infrastructure and capital.

4. How developed is it?

Reasoning AI has moved decisively beyond laboratory demonstration, but different aspects of the technology occupy very different maturity levels.

At the most established level is **reasoning assistance within general-purpose AI systems**. Extended-thinking or reasoning modes are commercially available from multiple developers and can be used directly by individuals and organisations. Models incorporating reasoning have become ordinary components of major AI platforms rather than specialist research prototypes.

A second comparatively developed category is **bounded reasoning in verifiable domains**. Mathematics, coding and formal tasks benefit from clear feedback, executable tests and objective answers. Performance in these environments can be extremely high, and verification mechanisms provide opportunities to detect errors before returning an answer.

A third category—**open-ended reasoning over unfamiliar, messy or weakly verifiable problems**—is materially less mature. Real-world problems often do not arrive already formulated, may contain missing or contradictory information, and may not have one objectively correct answer. Current systems can be highly useful in such settings, but reliability is considerably harder to demonstrate.

Planning provides a particularly important boundary. Models may write plausible plans or successfully reason through bounded sequences, yet systematic evaluation continues to find substantial weaknesses when actions, changing states and constraints must be represented correctly over extended problems.

Longer-horizon execution should also be distinguished from reasoning itself. METR's evaluations of frontier AI agents measure performance on software tasks according to the amount of expert human work needed to complete them. Its 2026 measurements demonstrate continuing progress on increasingly difficult tasks, but METR explicitly cautions that its “time

horizon” does not measure how long a model can autonomously operate. Such evidence is relevant to the growing connection between reasoning and agentic systems while illustrating why capability metrics require careful interpretation.

Evaluation science itself remains immature. The UK AI Security Institute notes that frontier-AI evaluation faces inconsistent terminology, developing measurement methods, small samples and new complications from inference-time scaling. A benchmark score is therefore not a fixed physical property of a model: prompting, tool access, computation budget, sampling method and scoring procedure may materially alter the result.

As at 30 August 2026, Reasoning AI is therefore **commercially established as an available capability, highly advanced in selected bounded domains, increasingly useful in complex professional and scientific assistance, but still developing as a dependable general problem-solving technology.**

5. What is driving its development and use?

The first major driver is the recognition that simply increasing pre-training scale is not the only way to improve model performance. Reasoning-oriented development shifts more attention towards post-training and computation at the point when a problem is being solved. Developers have reported continued gains from reinforcement learning and from allowing models greater inference-time computation.

A second driver is **verifiable training feedback**. Mathematics and programming offer large numbers of tasks for which the outcome can be automatically checked. Code can be compiled or tested; calculations can be verified; formal proofs can be checked mechanically. This makes it possible to reinforce successful problem-solving behaviour at greater scale than would be practical if every reasoning step required human judgement.

Third is **verification at inference time**. Generating one answer is increasingly being supplemented by producing alternative candidates, assessing intermediate reasoning or allowing another system to check the proposed solution. Research on process evaluation and reasoning verifiers reflects the growing importance of this approach. Yet verifier technology is itself imperfect: the 2026 VerifyBench study found specialised verifiers could achieve high accuracy in individual domains while still having recall weaknesses and limited cross-domain generalisation.

Fourth is **tool integration**. Calculators, code execution, web search, databases, symbolic mathematics and external software allow a reasoning system to obtain information or perform operations more reliably than relying entirely on the model's internal representation. This creates increasingly capable combinations of neural reasoning and conventional computation.

Fifth is strong user demand for AI that can address complex tasks rather than merely generate fluent text. Coding, research, quantitative analysis and technical problem-solving provide economically valuable applications in which an improvement in reasoning can directly improve usefulness. This demand encourages developers to offer different reasoning budgets, specialised modes and smaller reasoning models intended to balance capability, latency and cost.

Sixth is international competition and diffusion of methods. The emergence of DeepSeek-R1 materially widened access to reasoning-model techniques and demonstrated that strong results were not confined to a small group of proprietary Western systems. Open-weight and commercially accessible models from China and elsewhere have enabled researchers and developers to study, modify, distil and deploy reasoning approaches more widely.

Finally, reasoning has become important because it acts as an enabling capability for other parts of the AI landscape. Agents need to decide what to do next; scientific AI must evaluate hypotheses; multimodal systems must connect information across different forms; robots may need to plan actions; and verification systems must judge whether another AI's output is correct. Improvements in reasoning can therefore influence multiple technology categories even when Reasoning AI itself is not the final application.

6. What are the present limitations?

Reliability is the central limitation.

Current systems can alternate between remarkable competence and elementary error. Success on difficult benchmarks does not establish dependable performance on a user's next unfamiliar problem. This inconsistency is especially consequential because fluent explanations can make incorrect reasoning appear convincing.

Generalisation remains a substantial problem. Models may learn problem structures, templates or strategies that work extremely well on established benchmarks while performing much worse when equivalent reasoning is required in a different presentation. CORE-MATH's large performance falls when mathematical problems were converted into realistic contextual situations illustrate this difficulty.

Planning remains weak relative to the strongest benchmark achievements in mathematics and coding. ACPBench Hard found that even leading reasoning models struggled with fundamental questions about which actions were possible in particular states. This suggests that the ability to construct an impressive chain of reasoning does not yet imply reliable modelling of a changing environment.

Extended thinking also has limits. More computation often helps, but not uniformly. Models can continue along an incorrect path, reconsider a correct answer into an incorrect one or expend additional tokens without meaningful improvement. Research on knowledge-intensive tasks has found circumstances in which longer reasoning increased hallucination rather than correcting it.

Verification is not a complete solution. A verifier can itself be wrong, fail to detect a subtle error or work well in one technical domain but poorly in another. VerifyBench found substantial sensitivity to input structure and limitations in cross-domain generalisation among contemporary verifier systems.

Benchmarking presents further difficulties. Widely used test sets can become familiar through training data, public discussion or repeated optimisation. Benchmark saturation can then make it difficult to distinguish memorisation, learned strategies and genuinely transferable reasoning.

Research evaluating models on previously unseen or dynamically generated mathematical problems reflects the growing effort to reduce these effects.

Visible reasoning traces require particular caution. A coherent step-by-step explanation is useful for checking an answer, but it is not necessarily a faithful description of all internal computation that produced that answer. Conversely, an incorrect intermediate statement does not always mean that exactly that internal process caused the final answer. Reasoning traces therefore provide evidence to inspect, not a perfect window into model cognition.

Cost and latency also constrain deployment. Extended inference, multiple candidate solutions, search and verification consume more computation than a direct response. The practical question is consequently not simply whether extra reasoning improves accuracy but whether the improvement justifies additional processing time, energy and expense for the task concerned.

Domain knowledge is another constraint. Reasoning cannot reliably compensate for missing or incorrect factual information. A model may construct an internally consistent argument from a false premise or hallucinated fact. This is why retrieval, trusted databases and specialist tools can matter as much as reasoning depth in knowledge-intensive applications.

Finally, real-world effectiveness depends on the wider system. A reasoning model used in medicine, engineering, finance or law operates within data sources, software, professional processes, review mechanisms and accountability arrangements. High technical capability at one component does not by itself establish that the whole process is dependable.

7. What you should be aware of

The most important qualification is that **Reasoning AI is both genuinely more capable and still substantially less dependable than its strongest demonstrations can imply.**

Reasoning systems have crossed meaningful technical thresholds. Gold-medal-level mathematical performance, increasingly strong programming ability, sophisticated scientific problem-solving and the growing effectiveness of reinforcement learning, verification and inference-time computation are not merely changes in marketing terminology. They represent material advances in what AI systems can solve.

But reasoning is not a single general capability that improves uniformly. Mathematical proof, software debugging, deductive logic, factual research, contextual problem formulation and real-world planning impose different demands. A system can be exceptionally strong in one and unreliable in another.

Benchmark results therefore require context. The question is not only **what score was achieved**, but whether the problems were previously unseen; how much computation was allowed; whether multiple attempts were generated; whether tools or external information were available; whether a verifier selected the answer; and whether performance survives changes in wording and problem structure. The growing diversity of test-time scaling methods means apparently comparable “reasoning” scores can embody materially different computational procedures.

Reasoning traces also should not be anthropomorphised. A model that writes “I reconsidered my assumption” is generating a useful representation of a problem-solving process, not demonstrating human consciousness, introspection or understanding. Current evidence supports claims about observable performance much more strongly than claims that these systems reason in the same cognitive sense as people.

Another important distinction is between **reasoning and verification**. In domains where answers can be checked objectively, present systems can combine generation with executable tests, formal proof or other forms of validation. This can create much more dependable outcomes than asking a model to judge its own unconstrained answer. The availability of external verification is consequently part of the reason progress appears stronger in mathematics and coding than in many open-ended areas.

Reasoning should similarly be distinguished from **agency**. Determining a good sequence of steps is reasoning; independently carrying those steps out through software, tools or physical systems moves into agentic operation. Stronger reasoning is helping enable more capable agents, but it does not by itself establish reliable autonomy.

There is also no single geographic centre of Reasoning AI. Major proprietary development remains concentrated among a relatively small number of well-resourced organisations, particularly in the United States and China, but important models and research are being produced across Asia, Europe and other regions. DeepSeek's open release and Alibaba's Qwen development have been particularly significant in widening the technological base beyond the initial proprietary reasoning-model wave.

Most importantly, the present Status should not be inferred from exceptional examples alone. Research capability, product availability, operational deployment and dependable performance are different evidential stages and should not be treated as interchangeable.

As at 30 August 2026, Reasoning AI has become one of the most consequential developments in the applied AI landscape. AI systems can devote more computation to difficult problems, pursue and compare solution paths, use external tools, check candidate answers and reach levels of performance on selected mathematical, programming and scientific tasks that would have appeared exceptional only a short time ago.

The central current development is nevertheless not the arrival of universally reliable machine reasoning. It is the widening range of problems for which AI can perform useful structured problem-solving—and the growing ability to combine learned reasoning with computation, search and verification. The remaining challenge is that impressive reasoning under favourable and verifiable conditions still does not guarantee dependable judgement across unfamiliar, ambiguous and real-world problems.

Evidence Notes

1. **Google DeepMind — Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad, 21 July 2025.** Evidence of officially graded high-

level mathematical reasoning: five of six IMO problems solved for 35/42 points. Particularly useful because the solutions were graded under normal IMO criteria. This is a highly demanding but specialised benchmark and does not establish general reasoning reliability.

2. **Google DeepMind — Accelerating Mathematical and Scientific Discovery with Gemini Deep Think, 11 February 2026.** Evidence concerning iterative generation, verification and revision on professional mathematical and scientific problems and the use of additional inference-time computation. Developer research evidence; selected expert-supervised problems should not be interpreted as evidence of general autonomous scientific discovery.
3. **OpenAI — Introducing o3 and o4-mini, 16 April 2025; associated system card.** Evidence concerning large-scale reinforcement learning, inference-time reasoning, tool use and commercially available reasoning models. Benchmark and performance claims are primarily developer evaluations and should be interpreted according to their stated conditions, including tool availability and reasoning effort.
4. **DeepSeek — DeepSeek-R1 technical report, January 2025.** Important evidence of large-scale reinforcement learning producing strong reasoning behaviour and of dissemination through openly released model weights and distilled models. Benchmark comparisons are developer-reported; the wider importance lies partly in reproducibility and international diffusion of reasoning-model methods.
5. **Alibaba Cloud / Qwen — Qwen reasoning-model documentation and releases, 2025 onward.** Evidence of Chinese commercial and open-model development, including dedicated thinking models and hybrid systems able to switch between thinking and direct-response operation. Company evidence used primarily to establish availability and technological direction.
6. **Anthropic — Claude 3.7 Sonnet, 24 February 2025; Claude 4, 22 May 2025.** Evidence of commercially available extended-thinking systems, adjustable reasoning budgets and integration of reasoning with tool use. Company evidence principally establishing system capability and availability rather than independent effectiveness.
7. **Anthropic — Measuring Faithfulness in Chain-of-Thought Reasoning, 18 July 2023.** Foundational evidence that visible chains of thought need not be faithful descriptions of the process producing a model's answer. Older than the principal reasoning-model wave but still materially relevant to interpreting reasoning traces.
8. **Cao et al. — “From Abstract to Contextual: What LLMs Still Cannot Do in Mathematics,” ICLR 2026.** Evaluation of 61 proprietary and open models showing significant performance declines when established mathematical problems were placed into contextual settings requiring problem formulation. Particularly useful evidence of the distinction between benchmark solution ability and real-world problem formulation.
9. **Kokel et al. — ACPBench Hard: Unrestrained Reasoning about Action, Change, and Planning, ICLR 2026.** Independent academic evidence that current frontier and reasoning models remain unreliable on difficult open-ended planning problems, with most reported results below 65%.
10. **Zhou et al. — Logic-Grounded Metamorphic Testing for Evaluating the Reasoning Reliability of LLMs, Knowledge-Based Systems, 3 August 2026.** Recent peer-reviewed evidence that strong static logical-reasoning benchmark performance can coexist with inconsistency when logically equivalent formulations are introduced.
11. **Li et al. — VerifyBench, AAAI 2026, published 14 March 2026.** Cross-domain assessment of approximately 4,000 expert-level problems in mathematics and sciences. Evidence that model-based and specialised reasoning verifiers themselves have important limitations, including input sensitivity and uneven cross-domain generalisation.
12. **Hong et al. — Benchmarking LLMs' Mathematical Reasoning with Unseen Random Variables Questions, AAAI 2026.** Evidence addressing benchmark contamination and the importance of evaluating reasoning on genuinely unseen problems rather than relying solely on fixed public datasets.
13. **Hariri et al. — Test-Time Scaling in Reasoning LLMs: Inference Regimes, Evaluation, and Reproducibility, 4 August 2026.** Survey and analysis distinguishing longer individual deliberation, multiple

candidate sampling, verification and search over reasoning states. Particularly useful in demonstrating why “more inference compute” is not one uniform evaluation condition. Preprint evidence.

14. **Zhao, Hooi & Ng — Test-Time Scaling in Reasoning Models Is Not Effective for Knowledge-Intensive Tasks Yet, September 2025.** Experimental evidence that longer reasoning does not consistently improve factual knowledge-intensive performance and can under some conditions increase hallucination. Preprint evidence; useful as a qualification rather than a universal claim about test-time scaling.
 15. **Ghosal et al. — Does Thinking More Always Help?, June 2025.** Experimental evidence of diminishing or negative returns from simply extending a reasoning trajectory and of benefits from alternative approaches such as parallel candidate generation. Preprint evidence.
 16. **METR — Task-Completion Time Horizons of Frontier AI Models, updated 8 May 2026.** Independent measurement of frontier-agent performance on software tasks graded by human expert completion time. Useful evidence of progress in complex problem solving, with METR's explicit qualification that its time-horizon metric does not measure how long an AI system can autonomously operate.
 17. **UK AI Security Institute — Research Agenda and evaluation research, current at August 2026.** Public-sector evidence concerning the methodological challenges of measuring frontier AI capability, including limited samples, changing evaluation methods and inference-time scaling. Supports caution in interpreting benchmark performance as a simple or fixed measure of reasoning capability.
 18. **Finance Research Letters — “Is AI reasoning useful in finance?”, May 2026.** Domain-specific evidence showing that general reasoning-enhanced models did not outperform conventional models on the particular financial sentiment and earnings-prediction tasks studied. Useful qualification against assuming benchmark reasoning gains transfer automatically to every professional domain.
-

4. AI Agents & Agentic Systems

Status as at 30 August 2026

Status snapshot

AI Agents & Agentic Systems represent the movement of digital AI from principally generating outputs towards undertaking bounded activities: planning intermediate steps, using tools, interacting with software and data, and executing actions towards an objective. They are not one uniformly mature technology. Tool-using assistants and bounded agentic workflows are increasingly established product capabilities, while coding, research and specialised enterprise agents are moving into operational use. Computer-use and multi-agent systems are developing rapidly but remain less dependable, and broad long-horizon autonomous operation is substantially less mature.

Current agents can demonstrably write and test software, conduct multi-step research, retrieve and analyse information, interact with business systems and operate some computer interfaces. Deployment is growing across software engineering, customer service, enterprise administration and knowledge work, but product availability and successful pilots should not be confused with widespread adoption or demonstrated effectiveness at scale.

The central direction of development is towards AI systems able to carry out progressively longer and more complex digital workflows, supported by improved reasoning, tool access, interoperability and execution environments. The principal qualification is reliability: errors can accumulate across multiple actions, while security, permissions, integration, observability and accountability become more consequential as autonomy increases.

As at 30 August 2026, agentic AI is therefore a significant and increasingly operational part of the applied AI landscape, but its strongest current evidence remains in bounded, controlled and verifiable activity rather than dependable general digital autonomy.

1. What is it?

AI agents are digital AI systems designed not merely to generate an answer, but to pursue an objective by deciding what to do next and taking actions through software, tools or other external systems. An agent might search for information, retrieve files, call an application programming interface, operate software, write and run code, update a business system, communicate with another agent or ask a person for approval before continuing.

The important distinction is therefore between **producing information and undertaking activity**. A conventional generative-AI assistant might draft an email for a user. An agentic system may be able to identify the relevant recipient, retrieve supporting information, prepare the message and—if authorised—send it. The underlying AI model remains important, but the agent is the wider system around that model: instructions, tools, permissions, data connections, execution environment, state, safeguards and mechanisms for determining subsequent actions.

There is no universally agreed boundary around the term. The OECD's 2026 examination of the agentic-AI landscape found recurring characteristics across definitions but also substantial variation. Anthropic similarly distinguishes between **workflows**, in which models and tools are orchestrated through largely predetermined paths, and **agents**, in which the model more dynamically directs its own process and tool use.

This creates a continuum rather than a clean dividing line.

At one end are conventional automated workflows: the sequence of actions has essentially been designed in advance. Adding a language model may make such automation more flexible without making it genuinely autonomous. Further along are agentic workflows in which AI chooses between tools, adapts intermediate steps or hands work between specialised components while remaining inside a substantially bounded process. At the more autonomous end are systems given an objective and considerable discretion over how to pursue it, potentially operating over many steps with limited intervention.

Agentic does not therefore mean fully autonomous. Many commercially useful agents operate within deliberately narrow permissions, predefined workflows and human approval points. Indeed, limiting autonomy can increase practical reliability.

Modern agentic systems have been enabled by several developments occurring together: stronger foundation models and reasoning capabilities; reliable tool or function calling; multimodal understanding of screens and documents; longer context windows and external memory; execution environments for code and files; and increasingly standardised mechanisms for connecting models to external resources.

Two interoperability developments illustrate this emerging infrastructure. Anthropic introduced the open Model Context Protocol (MCP) in November 2024 to standardise connections between AI applications and external data and tools. Google introduced the Agent2Agent (A2A) protocol in April 2025 to enable agents built with different technologies to communicate and cooperate. Such protocols do not themselves make an AI system intelligent, but they make it easier for agentic systems to reach the resources through which useful actions can occur.

AI Agents & Agentic Systems in August 2026 are consequently best understood as an emerging **execution layer for digital AI**: systems that connect model intelligence to tools, applications, data and workflows so that AI can increasingly move from suggesting what should be done towards carrying out bounded portions of the work.

2. What can it currently do?

Current agents can already perform substantial multi-step digital work when the objective, tools and operating environment are sufficiently well defined.

One of the clearest capability areas is **software development**. Coding agents can inspect repositories, locate relevant files, write and modify code, run commands and tests, diagnose failures and iterate towards a solution. This differs materially from earlier coding assistants

that principally suggested individual code fragments. The agent can participate in a longer development loop.

Research also demonstrates rapid improvement in the length of tasks that frontier agents can complete. METR measures a system's "task-completion time horizon" according to the length of tasks, measured by the time required by human experts, that an agent can complete with a specified probability of success. Its May 2026 measurements show substantial improvement in frontier systems, although performance remains probabilistic and its principal task suite is concentrated on software-related work rather than the entire economy.

Agents can also conduct **research and information-gathering workflows**. They can break a question into sub-questions, search multiple sources, retrieve documents, compare evidence, perform calculations and synthesise the results into reports. The agentic feature is not simply producing the final prose: it is determining and executing intermediate information-gathering steps.

A third important capability is **computer use**. Multimodal agents can interpret graphical interfaces and operate websites and desktop applications through actions such as clicking, typing and navigating screens. This matters because much organisational software is not conveniently accessible through modern APIs. Microsoft made computer-using agents in Copilot Studio generally available in May 2026, while OpenAI has developed computer environments in which agents can work with files, networks, APIs and software.

Tool-connected agents can perform more structured actions through APIs and enterprise systems: querying databases, creating tickets, updating records, initiating workflows or retrieving information from organisational systems. Multi-agent arrangements can divide work among specialised agents—for example, one gathering information, another analysing it and another performing an action—although adding agents does not automatically improve outcomes.

The capability boundary remains important. Long chains of activity amplify small errors. An agent that is 95% reliable at each independent step would, in a simplified ten-step example, have only about a 60% chance of completing every step correctly. Actual agent behaviour is more complicated than this calculation, but it illustrates why apparently strong individual capabilities do not automatically produce dependable long-horizon execution.

Benchmarks reinforce the point. Early realistic environments exposed large gaps between demonstrations and robust operation. WebArena reported a 14.41% success rate for its strongest GPT-4-based baseline compared with 78.24% for humans. OSWorld initially reported 12.24% for its best evaluated model compared with more than 72% for humans on real-computer tasks. τ -bench found leading tool-using agents succeeding on fewer than half its tasks and becoming considerably less reliable when required to reproduce successful behaviour repeatedly. Subsequent systems have improved markedly, but the underlying evaluation problem remains: completing a task once is different from completing it consistently.

The present capability is therefore substantial but uneven. Agents can increasingly perform meaningful sequences of digital work. They are strongest where actions can be checked, environments are constrained, tools are well specified and mistakes can be detected or reversed. Dependable unattended execution becomes harder as tasks become longer, environments less

predictable, consequences more significant and required judgement more ambiguous.

3. Where is it currently being applied?

Software engineering is among the most developed applications. Agents are being used to investigate codebases, implement changes, test software, diagnose errors and support development workflows. This domain is particularly suitable because agents can use executable tools and receive relatively rapid feedback from compilers, tests and other machine-readable checks.

Customer service and service operations are another major area. Agents can retrieve customer information, consult knowledge bases, classify requests, propose or execute permitted actions and escalate exceptions to people. Many deployments remain bounded by business rules rather than allowing unrestricted autonomous decision-making.

In **enterprise administration and knowledge work**, organisations are developing agents for research, document processing, employee support, procurement, finance operations, sales support, IT service management and other repetitive digital workflows. Microsoft, Salesforce, ServiceNow, Google, OpenAI, Anthropic, Mistral and other technology providers have all developed agent-building or agent-enabled environments. The significance is less the number of branded products than the broader transition of enterprise AI platforms from conversational assistance towards tool-connected workflow execution.

Computer-use agents extend this possibility into **legacy applications and websites** where conventional integration is difficult. Instead of requiring a dedicated API, an agent can sometimes operate the same interface used by a person. This increases the range of potentially automatable activities, although interface changes, ambiguous screen elements, security controls and execution errors make such automation less predictable than well-engineered API integration.

Research and professional knowledge work are also becoming agentic. Systems can search multiple sources, analyse files, execute code and prepare integrated outputs rather than requiring a user to initiate every individual step.

Public-sector experimentation is emerging internationally. Singapore provides a particularly visible example. Its government and Google launched an AI Agents Sandbox in August 2025 to test computer-use agents in real-world settings including quality assurance, AI safety and social assistance. By May 2026 the programme was reporting both useful capabilities and continuing questions around cybersecurity, privacy, governance and human oversight. Singapore's public healthcare system has also developed AgentSea, a platform through which healthcare professionals had created more than 12,000 agents by August 2026, although the number created should not be interpreted as evidence that all were mature operational deployments.

China has also become a significant centre of agent development. Manus attracted international attention in 2025 with a system designed to undertake multi-step tasks with relatively little continuing user direction, while major Chinese technology companies have increasingly emphasised agentic capabilities and platforms. Alibaba, for example, has explicitly described

a strategic transition towards AI capable of executing actions rather than merely generating content.

European developers are participating in the same transition. France-based Mistral introduced an Agents API in May 2025 and subsequently expanded agentic coding and enterprise capabilities.

These examples demonstrate broad international technological activity, but they do not establish uniform global adoption. Agent development is concentrated where capable foundation models, cloud infrastructure, enterprise software integration, technical skills and investment are readily available. Deployment conditions differ substantially between countries, industries and organisations.

The current application landscape is therefore dominated by **bounded digital workflows**, not generally autonomous digital organisations. Agents are increasingly doing pieces of real work, but people, conventional software and predefined process controls remain integral to most serious deployments.

4. How developed is it?

AI Agents & Agentic Systems cannot sensibly be assigned one maturity level.

At the most established end are **tool-using assistants and bounded agentic workflows**. Function calling, retrieval, API integration and model-directed selection between defined tools are widely available technological capabilities. Organisations can build such systems using commercial and open development environments, and operational deployments exist across multiple business functions.

A second category comprises **coding, research and specialised enterprise agents** capable of executing substantial multi-step tasks. These have moved rapidly from experimentation into commercial use. Their maturity varies considerably by task: software work with strong verification mechanisms can support greater autonomy than activities involving ambiguous business judgement or irreversible external consequences.

Computer-use agents occupy a developing category. The technology is commercially available and can perform useful work, but operating graphical interfaces remains inherently more exposed to interface variation and visual or procedural misunderstanding than calling a stable, structured API.

Multi-agent systems are also developing rapidly. Commercial platforms now support agents delegating work to or coordinating with other agents. The emergence of A2A and related infrastructure reflects serious industry investment in interoperability. Yet the existence of orchestration technology is not evidence that large networks of autonomous agents are already operating reliably across enterprises at scale.

The broader adoption picture similarly requires caution. Deloitte's August 2026 survey found that only 15% of surveyed organisations reported scaled, orchestrated cross-functional multi-agent adoption, while only 5% considered their business processes highly prepared for agents.

At the same time, expectations were much higher: large proportions of respondents anticipated substantial future process redesign. These are survey results rather than comprehensive measures of the world economy, but they illustrate the gap between strong organisational interest and mature implementation.

Interoperability infrastructure is becoming more established. MCP has expanded from an Anthropic initiative into a widely supported mechanism for connecting AI systems with tools and data, while A2A addresses communication between agents. This emerging standardisation is important because agentic capability depends heavily on reliable connections to the wider digital environment.

Governance infrastructure is developing alongside the technology. Singapore launched a dedicated Model AI Governance Framework for Agentic AI in January 2026 and updated it in May with practical case studies and guidance concerning multi-agent systems, third-party agents and automation bias. The appearance of such frameworks is itself evidence that agentic systems have progressed sufficiently far into deployment for questions of authority, accountability and control to become operational rather than purely theoretical.

As at 30 August 2026, agentic AI is therefore neither merely experimental nor generally mature. **Bounded agents are an increasingly established product capability and are entering operational use; sophisticated multi-step agents are growing commercially; high-autonomy and long-horizon operation remain substantially less dependable.**

5. What is driving its development and use?

The first driver is improvement in the **underlying models**. Better reasoning, coding, multimodal perception and instruction-following make it possible for systems to select tools, interpret intermediate results and recover from some failures rather than simply generate a single response.

Second is the development of **tool infrastructure**. An intelligent model becomes considerably more operationally useful when it can search, retrieve files, execute code, query databases or interact with business applications. MCP, A2A, agent software-development kits and expanding libraries of connectors are reducing the amount of bespoke integration required.

Third is **economic demand for automation of knowledge work**. Conventional automation works well when processes are highly structured but often struggles with language, documents, exceptions and changing circumstances. Agentic AI is attractive precisely because it attempts to insert greater flexibility into these workflows.

Fourth is the enormous installed base of **digital enterprise systems**. Organisations already conduct finance, customer management, procurement, software development, human resources, logistics and administration through software. Agents do not need new physical infrastructure to affect these processes; if appropriately authorised, they can operate through existing APIs, connectors or user interfaces.

Fifth is **user demand for outcome-oriented AI**. The generative-AI wave familiarised millions of people with systems that answer questions and create content. Agent development attempts

to reduce the continuing human effort between receiving advice and completing the underlying task.

Sixth is **competition among technology providers and national AI ecosystems**. Agent capabilities have become a major focus for foundation-model developers, enterprise-software companies and cloud providers in North America, China and Europe, while governments including Singapore are actively developing governance and experimentation frameworks. This creates reinforcing investment in models, integration standards, evaluation, security and developer tools.

Finally, agent economics can improve when models become cheaper and more capable. An agent may consume considerably more inference than a single chatbot response because it repeatedly reasons, calls tools and processes intermediate results. Falling inference costs and more efficient models therefore broaden the range of workflows for which agentic execution may be economically practical.

6. What are the present limitations?

The central technological limitation is **reliability over sequences of actions**.

An ordinary chatbot error may produce a poor answer. An agent error can propagate into subsequent actions. The system may choose the wrong tool, misunderstand its result, modify the wrong record, repeat an action, fail to recognise completion or pursue an initially mistaken plan for many steps. Longer tasks create more opportunities for failure.

Evaluation consequently remains difficult. Benchmark success demonstrates capability under specified conditions, but real organisational environments contain changing interfaces, incomplete data, unexpected exceptions, conflicting instructions and consequences that benchmark environments cannot fully reproduce. METR's time-horizon research provides useful evidence of improving autonomous task capability while also illustrating that success must be expressed probabilistically rather than as a simple claim that an agent "can" perform a particular length of task.

Security becomes more consequential because agents have tools and permissions. Indirect prompt injection or "agent hijacking" can occur when malicious instructions are embedded in material that an agent reads, such as a webpage, document or email. NIST's Center for AI Standards and Innovation has demonstrated the continuing difficulty of defending agents against such attacks, including through large-scale red-team evaluation.

OWASP's agentic-security work similarly identifies risks involving goal hijacking, tool misuse, identity and privilege abuse, supply-chain vulnerabilities and unexpected code execution. These are system-level problems: security depends not only on the underlying model but on what the agent can access, which credentials it possesses and what consequences its tools can produce.

Permissions and identity therefore become central design constraints. Giving an agent broad access increases what it can accomplish but also increases the consequences of error or

compromise. Restricting access improves containment but can reduce usefulness. This tension is intrinsic to systems whose purpose is to take action.

There is also an **observability problem**. Organisations need to know what an agent attempted, which information it accessed, which tools it invoked, why a transaction occurred and whether a human approved it. Agent behaviour can involve many model calls and intermediate actions, making debugging and audit more complex than for conventional deterministic software.

Cost and latency remain practical constraints. A multi-step agent may invoke large models repeatedly, perform searches, execute code and interact with multiple systems before completing one task. Greater deliberation can improve outcomes while also increasing time and computational expense.

Integration remains another major obstacle. Enterprise data can be fragmented, permissions inconsistent and legacy systems difficult to connect. Computer use can bypass some integration problems by allowing AI to operate graphical interfaces, but this replaces one challenge with another: interfaces designed for people are not necessarily stable or unambiguous environments for autonomous software.

Governance and accountability become more difficult as autonomy rises. Singapore's agentic-AI framework emphasises bounding agents' powers, establishing human accountability and placing human approval at significant checkpoints. Its government has specifically stated that high-stakes or irreversible actions should not occur without human review.

Finally, organisational processes themselves may not be ready for agents. A poorly defined process does not necessarily become a good process because AI executes it. Deloitte's 2026 survey finding that only 5% of responding organisations considered their processes highly prepared for agents illustrates that the deployment challenge extends beyond model capability.

The limitations of agentic AI are consequently not reducible to whether the model is "intelligent enough". Dependable agency requires the model, tools, permissions, data, workflow, security controls, evaluation mechanisms and human governance to function together.

7. What you should be aware of

The most important distinction is between **assistance and agency**.

A system that recommends an action and leaves execution to a person presents a different operational risk from one authorised to perform that action itself. The same underlying model can therefore have very different significance depending on the tools and authority surrounding it.

Second, **agentic capability is not the same as autonomy**. Many useful agents are intentionally constrained. They may work only within one application, use a limited set of tools, require approval before consequential actions or hand exceptions to people. Such constraints should not automatically be interpreted as technological shortcomings; they can be fundamental to dependable deployment.

Third, **workflow automation and AI agency overlap but are not synonymous**. Conventional automation follows predetermined logic. Agentic systems introduce varying degrees of model-directed choice. In practice, many of the strongest current systems combine the two: deterministic software controls critical boundaries while AI handles the parts requiring interpretation or flexibility.

Fourth, a successful demonstration should not be confused with reliable long-term operation. Agent benchmarks have improved rapidly, and agents can now undertake activities that were difficult only a short time ago. But performance across unfamiliar situations, repeated trials and extended chains of action remains materially more demanding than completing a selected task once.

Fifth, **the agent is a system, not merely a model**. Changing the foundation model can alter capability, but so can changing the available tools, permissions, prompts, retrieval system, execution environment, memory, verification process or human checkpoints. Claims about an "agent's intelligence" can obscure these system-level dependencies.

Sixth, multi-agent systems should not automatically be assumed to be more capable than single agents. Dividing work among specialised agents can improve modularity and enable collaboration, but it also creates additional communication, coordination, security and error-propagation problems.

Seventh, adoption is geographically and institutionally uneven. North American technology companies have played a major role in developing agent platforms and underlying models; China has a substantial and increasingly visible agent ecosystem; European developers are contributing models and agent platforms; and Singapore has become particularly active in practical experimentation and agentic governance. This does not mean that equivalent operational adoption exists throughout those regions or elsewhere.

Most importantly, the present transition is not from human work to unrestricted machine autonomy. It is from AI systems that principally **generate outputs** towards AI systems increasingly able to **participate in processes and execute bounded actions**.

As at 30 August 2026, that transition is technologically significant and already visible in real deployments. Agents can write and test software, conduct multi-step research, interact with business systems, operate computer interfaces and coordinate increasingly complex digital workflows. The surrounding infrastructure for tools, interoperability, evaluation and governance is developing rapidly.

What has not been established is dependable general autonomy across the open-ended digital world. Current agentic systems remain most credible when their objectives are bounded, their tools controlled, their actions observable, their outputs verifiable and human intervention remains available where consequences warrant it.

The defining current development is therefore not the arrival of universally autonomous digital workers. It is the widening range of digital activities over which AI systems can move from **responding to instructions towards deciding and executing the intermediate actions required to fulfil them**.

Evidence Notes

1. **OECD — *The Agentic AI Landscape and its Conceptual Foundations*, 13 February 2026.** International institutional analysis of competing definitions and common characteristics of agents and agentic AI. Particularly useful for avoiding treatment of "agent" as a universally standardised technical category.
2. **Anthropic — *Building Effective Agents*, 19 December 2024.** Developer analysis distinguishing workflows, in which orchestration follows predefined paths, from agents that dynamically direct their own processes and tool use. Developer evidence, but influential in establishing practical architectural distinctions.
3. **Anthropic — *Model Context Protocol*, 25 November 2024.** Primary evidence concerning the introduction of MCP as an open protocol connecting AI applications to external tools and data. Used as evidence of agent infrastructure rather than agent performance.
4. **Google — *Agent2Agent Protocol*, 9 April 2025.** Primary evidence concerning A2A and the developing infrastructure for interoperability between independently built agents. Product/ecosystem source rather than evidence of widespread interoperable multi-agent deployment.
5. **OpenAI — *New Tools for Building Agents*, 11 March 2025, and *From Model to Agent: Equipping the Responses API with a Computer Environment*, 11 March 2026.** Primary developer evidence concerning tool use, computer use, execution environments and agent-building infrastructure. Used principally to establish technological availability rather than independent performance.
6. **Microsoft — *Copilot Studio computer-using agents*, 13 May 2026.** Primary evidence that computer-use capability had progressed into generally available enterprise tooling across commercial Power Platform geographies. Availability does not establish adoption or dependable performance across arbitrary interfaces.
7. **METR — *Task-Completion Time Horizons of Frontier AI Models*, updated 8 May 2026; *Measuring AI Ability to Complete Long Tasks*, 19 March 2025.** Independent evaluation framework measuring autonomous task capability according to human task duration. Particularly useful for demonstrating improving long-horizon capability while preserving the distinction between probabilistic benchmark success and dependable real-world operation. The evaluated tasks are concentrated substantially in software and related technical work.
8. **Zhou et al. — *WebArena*, 2023.** Research benchmark using realistic websites and long-horizon tasks. Early results showed a large gap between agent and human completion rates. The original baseline figures should not be interpreted as measures of 2026 frontier performance; their importance here is in demonstrating the challenge of evaluating realistic multi-step agency.
9. **Xie et al. — *OSWorld*, 2024.** Research benchmark involving 369 tasks across real computer applications and operating systems. Initial results demonstrated substantial weaknesses in graphical-interface grounding and operational knowledge. As with WebArena, initial performance is historical rather than a 2026 frontier score.
10. **Yao et al. — *τ-bench*, 2024.** Research benchmark testing tool-using agents in interaction with users and domain rules. Particularly material for the distinction between one-time task success and repeated reliability.
11. **NIST Center for AI Standards and Innovation — *Agent Hijacking Evaluations*, 17 January 2025, and *large-scale agent-security red-teaming findings*, 23 March 2026.** Public-sector technical evidence concerning indirect prompt injection and continuing vulnerabilities when agents process untrusted external information.
12. **OWASP GenAI Security Project — *Agentic AI Threats and Mitigations*, 17 February 2025; *Top 10 for Agentic Applications*, December 2025; related 2026 security material.** Practitioner-led security evidence identifying risks including goal hijacking, tool misuse, identity and privilege abuse, supply-chain weaknesses and unexpected code execution. These are security taxonomies and guidance rather than measurements of incidence across all deployed agents.

13. **Singapore Government / Google — AI Agents Sandbox, launched August 2025; findings published 20 May 2026.** Public-sector evidence from practical testing of computer-use agents, including potential applications and continuing cybersecurity, privacy, oversight and governance issues.
 14. **Singapore IMDA / MDDI — Model AI Governance Framework for Agentic AI, January–May 2026, with parliamentary clarification in May and August 2026.** Primary governance evidence concerning bounding agent authority, human accountability, approval checkpoints, multi-agent systems, third-party agents and automation bias. Particularly relevant because it addresses deployed agentic systems rather than AI governance only in general terms.
 15. **Singapore Ministry of Health, 24 August 2026.** Public-sector evidence concerning AgentSea, a sector-wide platform for agent development in public healthcare, and more than 12,000 agents created by healthcare professionals. The figure represents agents created and experimentation activity; it should not be interpreted as 12,000 independently validated operational deployments.
 16. **Deloitte — Agentic AI readiness survey, 12 August 2026.** Survey evidence indicating substantial organisational expectations but much lower current process readiness and scaled cross-functional multi-agent adoption. Survey findings reflect the responding population and should not be extrapolated mechanically to all organisations globally.
 17. **Reuters — Manus / Chinese agent ecosystem, 21 March 2025.** Independent reporting concerning the emergence and rapid public attention surrounding Manus and China's developing agent ecosystem. Claims that Manus represented a "first" or fully general agent originated principally with its developer and should not be treated as an independently established technological classification.
 18. **Alibaba Group, 23 March 2026.** Primary company evidence concerning Alibaba's strategic emphasis on agentic AI and action-oriented systems. Statements concerning future market size are corporate expectations rather than evidence of current realised economic impact and are not relied upon for the Status assessment.
 19. **Mistral AI, May–July 2025 materials.** European developer evidence concerning agent APIs and agentic coding systems. Used to establish geographic breadth of active technological development rather than comparative performance.
 20. **World Economic Forum — *AI Agents in Action: Foundations for Evaluation and Governance*, 27 November 2025, and *A Playbook for Trusted Adoption, Authorization and Scaling*, 26 May 2026.** Cross-industry material documenting the movement of agents from prototypes towards deployment and the growing importance of evaluation, authorisation, monitoring and governance. Survey and industry-participant evidence should be interpreted as indicators rather than comprehensive global adoption statistics.
-

5. Multimodal AI

Status as at 30 August 2026

Status snapshot

Multimodal AI has moved from a specialist capability into a central part of the applied AI landscape. Its significance lies not simply in AI systems being able to handle text, images, audio or video separately, but increasingly in their ability to relate information across those modalities within the same task.

Maturity varies considerably. Image-and-language understanding, document interpretation and visual interaction are broadly available in general-purpose AI services, while audio interaction is also commercially established. Video understanding, multi-image reasoning and more fully integrated systems combining text, vision, audio and video are developing rapidly but remain less consistently dependable. Specialised multimodal sensor-fusion systems are already operationally mature in areas such as autonomous systems, while applications including clinical multimodal foundation models remain at materially different stages of research, validation and deployment.

Current systems can demonstrably interpret images and visually structured documents, combine visual and textual evidence, process speech and other audio, analyse video to varying degrees, and increasingly retrieve or generate information across modalities. Their application is expanding across general AI assistance, workplace information, accessibility, search, healthcare research, autonomous systems, robotics and science.

The principal qualification is that multimodal capability does not establish human-like perception or reliable understanding: hallucination, spatial and temporal reasoning, long-video comprehension and integration of conflicting information remain significant limitations. **As at 30 August 2026, multimodality is becoming an increasingly important enabling layer across AI, while dependable integration across complex real-world information remains an active technological challenge.**

1. What is it?

Multimodal AI describes artificial-intelligence systems that can work with more than one form, or *modality*, of information. Those modalities can include text, images, speech and other audio, video, diagrams, documents, computer interfaces, sensor readings and structured data.

The important development is not simply that one AI product can contain separate tools for different media. Multimodality becomes more significant when information from different modalities is represented, related and interpreted together. A system might, for example, examine a photograph while considering a written question about it; interpret a chart embedded in a report; connect speech with what is happening in a video; or use visual information together with language instructions to determine what a robot should do.

This distinction separates **integrated multimodal AI** from a collection of otherwise independent capabilities. A text model connected to a separate image generator may offer several media functions without deeply integrating them. By contrast, newer multimodal foundation models increasingly learn relationships between modalities within the same model or closely integrated architecture.

The technological progression has been substantial. Earlier AI systems commonly specialised in one information type: computer-vision systems classified images, speech-recognition systems converted audio into text, and language models processed words. More recent systems increasingly combine these functions. OpenAI's GPT-4o, for example, was trained across text, vision and audio and was designed to accept combinations of text, audio, images and video. Google's Gemini family was similarly developed as multimodal rather than simply adding vision to a text-only system. Meta's Llama 4 models introduced natively multimodal open-weight models, while Chinese developers including Alibaba's Qwen and Baidu's ERNIE have developed increasingly unified vision-language and omni-modal systems.

By 2026, the boundary has broadened further from predominantly **vision-language models** towards systems combining text, images, audio and video. Qwen3.5-Omni, for example, supports text, image, audio and audio-visual understanding and interaction, while Baidu describes ERNIE 5.0 as integrating text, images, video and audio within a unified architecture. These are developer descriptions and benchmark-backed technical claims rather than evidence that every capability is equally dependable in ordinary use.

Multimodality also extends beyond foundation models. Autonomous vehicles have long combined cameras, radar, lidar and positioning information; healthcare research combines medical images with clinical records and laboratory data; and robotics combines vision, language and physical sensor information. These forms of **sensor and data fusion** belong within the broader multimodal landscape even where they do not resemble conversational AI.

Multimodal AI in August 2026 is therefore best understood as a spectrum: from established systems that fuse particular data sources for specialised tasks, through widely available foundation models that interpret images and documents alongside language, to newer systems attempting increasingly unified understanding, reasoning, generation and interaction across many forms of information.

2. What can it currently do?

The clearest established capability is **combining visual information with language**.

Current systems can inspect photographs, screenshots, charts, diagrams, maps and document pages and answer questions about what they contain. They can identify objects and relationships, extract visible text, explain diagrams, compare images, interpret graphical information and combine those observations with substantial language-based knowledge.

This makes documents an especially important practical multimodal application. A report is rarely just text: it may contain tables, graphs, photographs, diagrams and page layout carrying meaning. Multimodal systems can increasingly process these components together rather than treating the document simply as extracted text. Anthropic, for example, has described visual

processing of photographs, charts, graphs and technical diagrams, reflecting a capability now present across several major AI systems.

The capability extends beyond individual images. Systems can compare multiple images, follow sequences of screenshots and reason about visual changes. They can also analyse video by sampling or processing frames and, increasingly, combining those frames with subtitles and audio.

Video understanding demonstrates both the progress and the remaining difficulty. The peer-reviewed Video-MME benchmark evaluates systems across 900 videos ranging from seconds to an hour and combines visual frames with subtitles and audio. Its research found that additional subtitle and audio information improved video understanding, demonstrating the practical value of integrating modalities rather than relying on images alone. It also found performance deteriorating as videos became longer.

Audio adds another dimension. Multimodal systems can recognise speech, distinguish aspects of tone and delivery, converse using spoken language, interpret sounds and relate audio to visual events. Native audio systems have made interaction substantially more fluid because they need not always reduce speech to a text transcript before reasoning about it. Google, for example, describes native audio capabilities designed to interpret not only words but characteristics such as tone, accent and non-speech vocalisation.

Cross-modal **retrieval** is also becoming more capable. A system can increasingly search for related information even when the query and the target are in different forms—for example, using text to find relevant images or video, or using one image to retrieve related material. Google's Gemini Embedding 2 illustrates this development by mapping text, images, video, audio and documents into a shared representation for retrieval and classification. It moved from public preview to general availability in 2026.

Another developing capability is **multimodal-to-multimodal generation**. Systems can use combinations of text, images, audio or video as references and generate or edit another medium. Google's Gemini Omni, for example, was introduced in 2026 with the ability to combine several input forms in producing and conversationally editing video. OpenAI had earlier integrated image generation into its multimodal model architecture. The detailed quality and reliability of media generation belong principally within Image, Audio & Video AI; its significance here is that understanding and generation are increasingly connected across modalities.

Multimodality can also support action. A model that can see a software interface, understand an instruction and determine what should happen next has capabilities that text-only systems lack. Similarly, vision-language-action systems in robotics connect language and perception with physical action. These applications overlap with AI Agents & Agentic Systems and Physical AI & Robotics rather than making multimodality itself equivalent to agency.

The strongest current systems consequently do considerably more than image recognition. They can relate different information sources, answer questions requiring several modalities, maintain interactive audio-visual exchanges and increasingly generate across modalities.

But capability remains uneven. Recognising the principal subject of an image is considerably easier than accurately counting many objects, interpreting subtle spatial relationships,

following a long video, understanding an unfamiliar specialist diagram or reasoning correctly from several images containing ambiguous evidence. Benchmark success therefore establishes capability under particular conditions, not universal visual or multimodal understanding.

3. Where is it currently being applied?

The most widespread application is increasingly **general-purpose AI assistance**. Consumer and workplace AI services allow users to upload images, photographs, screenshots and documents alongside written or spoken questions. This has made multimodal interaction part of mainstream generative AI rather than a specialist computer-vision function.

Document and workplace analysis is a major practical use. Multimodal systems can examine presentations, forms, diagrams, financial material, graphical reports and other visually structured information. This matters because much organisational knowledge cannot be represented adequately as continuous text. Visual understanding therefore expands the portion of an organisation's information that general-purpose AI systems can use.

Search and information retrieval are becoming multimodal. A person can photograph an object or scene and ask a question about it rather than first describing it in words. Google introduced multimodal querying into AI Mode in Search in 2025, combining Lens-based visual information with Gemini to interpret scenes and answer questions.

Accessibility provides a particularly tangible application. Visual AI can describe photographs and surroundings, read labels and signs and answer questions about visible objects for people who are blind or have low vision. Be My Eyes' Be My AI provides AI-based visual interpretation, while camera-equipped wearable devices increasingly allow visual assistance to be accessed without continuously holding a phone. These systems demonstrate practical multimodal interaction, although reliability remains important when users may depend on descriptions to understand their surroundings.

Healthcare and biomedical research are important but substantially more regulated and less uniformly mature. Multimodal medical research combines radiological or pathological images with electronic health records, clinical notes, laboratory results, genomic information and other patient data. The rationale is strong: clinicians themselves make decisions by combining multiple information sources rather than interpreting an image in isolation. Reviews published in 2025–26 document extensive research across radiology, pathology, clinical prediction and treatment-related tasks.

However, research activity should not be mistaken for routine clinical deployment of general-purpose multimodal foundation models. Reviews continue to identify limited external clinical validation, hallucinations, generalisability problems, privacy requirements and uncertain regulatory pathways. Multimodal medical AI therefore ranges from established specialised data-fusion methods to research-stage foundation-model applications.

Transportation and autonomous systems represent a different and more established form of multimodality. Vehicles combine cameras, radar, lidar, ultrasonic sensors and positioning systems because individual sensors have different strengths and failure modes. A 2026 systematic review covering 66 peer-reviewed studies concluded that multimodal sensor fusion

had become an important architectural approach for improving perception robustness and fault tolerance.

Robotics similarly depends on multimodality. Robots may combine cameras, language instructions, force or tactile sensing and information about their own physical state. Multimodal foundation-model techniques are increasingly being used to connect perception with higher-level instructions and action. This is an important enabling role for multimodality, although the resulting physical capability belongs principally within Physical AI & Robotics.

Scientific research is another developing area. Scientific problems often require simultaneous interpretation of images, measurements, text and domain knowledge. The MicroVQA benchmark presented at CVPR 2025, for example, was developed specifically to test multimodal reasoning over microscopy-based scientific research tasks, illustrating both growing capability and the difficulty of research-level visual reasoning.

These applications show why multimodality matters: much of the world does not arrive as clean text. It arrives as mixtures of documents, images, sound, video, physical measurements and structured records.

4. How developed is it?

Multimodal AI cannot sensibly be assigned a single maturity level.

At the most mature end are **specialised multimodal and sensor-fusion systems**. Combining different sensor streams in vehicles, industrial systems, medical imaging workflows and other specialised applications predates the current foundation-model wave. Such systems can be operationally mature within defined domains even though they do not possess general multimodal intelligence.

Image-and-text foundation-model capability has also moved into broad commercial availability. Visual questioning, screenshot analysis, document interpretation and image-assisted conversational systems are now incorporated into widely distributed AI platforms. This category has progressed well beyond research demonstration.

Audio-language interaction is similarly commercially available, although capabilities differ substantially across systems, languages and operating conditions. Native multimodal architectures have reduced some of the fragmentation of earlier pipelines in which separate speech recognition, language processing and speech synthesis components handled successive stages.

Video understanding is developing rapidly but is less dependable than ordinary image understanding. Video introduces time: systems must determine not only what appears but what happened before or after something else, whether objects remain the same across frames and how visual events relate to speech or sound. Video-MME found declining performance as duration increased, illustrating why apparently strong short-video demonstrations do not establish dependable understanding of lengthy recordings.

Broad omni-modal systems integrating text, images, audio and video represent a rapidly developing frontier. Major model families in North America and China demonstrate that this is no longer confined to laboratory research. Qwen3.5-Omni, ERNIE 5.0 and Google's Gemini developments are examples of efforts to bring several modalities into increasingly unified systems.

Open-weight multimodality has developed as well. Meta's Llama 4 Scout and Maverick are natively multimodal image-and-text models, while Baidu released multiple ERNIE 4.5 variants. This widens access to multimodal capability beyond proprietary hosted services and allows organisations and researchers greater flexibility in deployment and adaptation.

Benchmark performance has improved substantially, but benchmark design itself is evolving because older tests can cease to discriminate reliably between systems. MMMU-Pro, for example, strengthened the earlier MMMU benchmark by filtering questions answerable without genuine visual information and introducing settings intended to require combined visual and textual understanding.

The Status is consequently one of **commercial maturity for several bounded multimodal functions, rapid improvement in general-purpose visual and audio interaction, and continuing research and early deployment for more difficult integrated reasoning across long, complex or numerous multimodal inputs.**

5. What is driving its development and use?

The most fundamental driver is that **human information is inherently multimodal**. People communicate through language, voice, images, gestures, diagrams and demonstrations. Workplaces contain documents and screens. The physical world produces visual, acoustic and sensor information. An AI system restricted to text can therefore access only part of the information relevant to many tasks.

A second driver is the development of **foundation-model architectures capable of learning shared representations across modalities**. Rather than building completely separate systems for every task, developers increasingly train large models or integrated architectures to connect concepts appearing in different information forms. This allows knowledge acquired through one modality to support interpretation of another.

Third is the enormous expansion of **digital multimodal data**. Images, video, audio, scanned documents, presentations and sensor streams are now generated at extraordinary scale. This provides both training material and a large body of information that organisations and individuals want AI systems to interpret.

Fourth is improved **computing infrastructure**. Training and serving multimodal systems is computationally demanding because images, video and audio can represent much more input data than short text exchanges. Improvements in accelerators, model efficiency, compression, inference infrastructure and edge processing are making such workloads more practical.

Fifth is **user demand for more natural interaction**. Asking a system "What is this?" while pointing a camera at something can be easier than describing it. Spoken interaction can be

preferable while moving or working with one's hands. Showing a system a screenshot can be more efficient than explaining a software problem in words. Multimodality therefore changes the interface between people and AI as well as the underlying model capability.

Sixth is demand for **AI that can operate within real environments and workflows**. Agents interacting with graphical interfaces need visual understanding. Robots need perception. Autonomous vehicles require multiple sensors. Clinical systems need to combine different patient information sources. Multimodality is consequently an enabling technology for several other important parts of the AI landscape.

Finally, development is geographically broad. US-based developers have played a major role in commercial multimodal foundation models, but Chinese laboratories and companies are significant contributors to omni-modal and open-model development. European developers and research institutions contribute to multimodal models, evaluation and application research, while multimodal applications and research communities extend across Asia and other regions. The present landscape is therefore international, although access to compute, data, advanced hardware and deployment infrastructure remains highly uneven.

6. What are the present limitations?

The central problem is that **combining modalities does not guarantee that a system has understood them correctly**.

Visual hallucination remains material. A system may describe an object that is absent, miscount objects, confuse relationships or infer something from prior knowledge rather than from the image actually supplied. Research on multi-image hallucination has found that hallucination can increase as the number of images rises, while other work distinguishes failures caused by perception from failures arising during reasoning after an image has been perceived correctly.

Spatial reasoning is particularly difficult. Systems can recognise objects while still misunderstanding whether one object is behind, inside, above or interacting with another. A fluent verbal explanation can consequently conceal a mistaken visual interpretation.

Video adds temporal difficulty. Understanding a long recording requires retaining and relating events occurring at different times. Sampling only some frames can miss crucial events, while processing many frames increases computational demands. Video hallucination research and Video-MME both demonstrate that longer and more complex video remains challenging.

Multiple images create additional problems. A model must determine which image contains which evidence, preserve object identity between views and avoid merging unrelated visual information. Dedicated multi-image reasoning benchmarks remain an active research area, itself evidence that dependable performance is not solved.

Data quality is another constraint. Multimodal datasets are difficult to curate because relationships between modalities must be correct. An image and caption can be loosely associated rather than precisely aligned; subtitles may not describe everything visible in a

video; medical data collected at different times may not be directly comparable; and sensor streams require accurate temporal and spatial alignment.

Bias can also cross modalities. Images, speech and video contain demographic, cultural and geographic patterns that may be poorly represented in training data. Performance can differ across languages, accents, visual environments and cultural contexts. Multimodality therefore adds potential sources of bias rather than automatically reducing them.

Compute and latency remain important. High-resolution images, long audio recordings and video streams contain large amounts of information. Processing them can increase memory requirements, inference cost and response time. Real-time applications add stricter constraints because useful output may be required while events are still occurring.

Privacy is particularly significant because multimodal inputs can contain unusually rich personal information. A photograph may reveal faces, locations or documents in the background. Audio can capture other people's conversations. Video can expose homes or workplaces. Medical multimodal systems can combine several highly sensitive patient datasets. The capability to integrate information can therefore amplify privacy consequences when access or governance is inadequate.

Safety depends strongly on application. An inaccurate description of a holiday photograph may be inconvenient; a mistaken interpretation of medical imagery, traffic conditions or physical surroundings can be much more consequential. The appropriate standard of reliability therefore cannot be inferred merely from general multimodal benchmark performance.

Evaluation remains difficult for the same reason. There is no single score that captures visual perception, speech understanding, video comprehension, cross-modal reasoning, generation, latency, hallucination, robustness and performance across languages and environments. NIST's 2026 draft TEVV-Athlon framework explicitly treats multimodal models among AI systems requiring application-sensitive test, evaluation, verification and validation.

Finally, integration itself can fail. In autonomous driving, for example, multimodal sensor fusion is valuable partly because sensors provide complementary evidence, yet calibration errors, timing differences, sensor failure and environmental change can undermine fusion. Reviews continue to identify spatio-temporal misalignment, domain shift and interpretability as real-world challenges.

Multimodality therefore increases the information available to AI, but it also increases the number of ways information can be missing, inconsistent, misaligned or incorrectly interpreted.

7. What you should be aware of

The most important qualification is that **multimodal does not mean human-like perception**.

A system capable of discussing a photograph, answering questions about a chart and conversing through speech has crossed an important technological boundary. It can access forms of information that text-only systems cannot. But successful performance on these tasks does not establish that the system experiences or understands the world as a person does.

Nor does the presence of several modalities mean they are equally integrated. One system may have exceptionally strong image understanding but relatively limited video capability. Another may combine audio and language effectively but process visual information through a less integrated pathway. "Multimodal" therefore describes a broad capability class, not a uniform technical standard.

A second distinction is between **perception and reasoning**. A system may correctly recognise what is present in an image and then reason incorrectly about it. Alternatively, its reasoning process may be coherent but based on a visual detail it misperceived. This distinction matters because an apparently incorrect "reasoning" answer may actually originate in perception, while a correct description does not prove sound reasoning.

Third, **additional modalities can improve performance without guaranteeing it**. Video-MME showed that subtitles and audio could materially improve video understanding. In other circumstances, however, conflicting or irrelevant information can make interpretation harder. More information is useful only if the system can identify, align and appropriately weight it.

Fourth, **availability is not the same as dependable adoption**. Image and document understanding are already widely available in general-purpose AI products. Real-time audio-visual interaction is increasingly available. More ambitious uses in clinical decision-making, autonomous systems and scientific reasoning remain at materially different levels of validation and deployment.

Fifth, multimodality is increasingly becoming an **enabling layer for other AI developments**. Agents can use vision to operate software. Robots can combine language and perception. AI assistants can understand what a user sees and hears. Retrieval systems can search across different media. The significance of multimodal AI may therefore become less visible as a separate category precisely because multimodal capability is being incorporated into broader AI systems.

There is also an important global dimension. Multimodal development is not confined to one provider or country. Major proprietary, open-weight and research systems are being developed across the United States, China, Europe and wider international research communities. At the same time, access to the computing infrastructure, datasets, languages and devices needed to benefit from these systems remains uneven.

Most importantly, multimodal capability should be judged by the **whole task and operating environment**, not by an impressive demonstration. Recognising a scene once is different from interpreting thousands of images reliably. Answering a question about a short clip is different from monitoring hours of video. Producing a plausible medical interpretation is different from demonstrating safe clinical effectiveness. Describing an object is different from giving sufficiently reliable information for a physical system to act upon it.

As at 30 August 2026, Multimodal AI has therefore moved from a specialist research area into a central feature of the applied AI landscape. Image-and-language understanding is broadly available; document and visual interaction are increasingly routine; audio and video are becoming more deeply integrated; and systems capable of combining several modalities within increasingly unified architectures are developing rapidly.

The central current development is not simply that AI can process more kinds of files. It is that the boundaries between previously separate forms of machine perception, language processing, interaction and generation are progressively weakening. AI systems are becoming able to relate what is written, seen, heard and recorded within a common task. That materially expands what they can do—but reliable integration across complex real-world information remains an active technological challenge.

Evidence Notes

1. **OpenAI — GPT-4o, 13 May 2024; GPT-4o System Card, 8 August 2024.** Primary developer evidence that GPT-4o was designed and trained across text, vision and audio and accepts combinations of text, audio, images and video. Used as evidence of technical architecture and availability rather than independent proof of all performance claims.
2. **Google DeepMind / Google — Gemini materials, 2023–26.** Primary evidence concerning natively multimodal Gemini systems, native audio, multimodal embeddings and the 2026 Gemini Omni direction. Developer evidence; product availability and architectural claims do not by themselves establish reliability across applications.
3. **Qwen Team — Qwen3.5-Omni Technical Report, 17 April 2026, together with Alibaba Cloud model documentation.** Evidence concerning integrated text, image, audio and audio-visual capability and multilingual multimodal interaction. Benchmark results are principally developer-reported and should not be interpreted as independent evidence of equivalent real-world performance.
4. **Baidu — ERNIE 4.5, 30 June 2025; ERNIE 5.0, 6 February 2026.** Primary Chinese developer evidence concerning multimodal model architectures and the movement towards unified text, image, video and audio models. Used principally to establish the international technological landscape.
5. **Meta — Llama 4, 5 April 2025.** Primary evidence concerning open-weight natively multimodal models and long-context image-and-text capability. Comparative benchmark claims are developer claims.
6. **Fu et al. — Video-MME, CVPR 2025.** Peer-reviewed benchmark covering 900 videos and 2,700 question-answer pairs across short, medium and long videos, including audio and subtitles. Particularly useful evidence concerning the value of combining modalities and continuing difficulty with longer video.
7. **Yue et al. — MMMU-Pro, ACL 2025.** Peer-reviewed benchmark designed to test whether multimodal systems genuinely integrate visual and textual evidence rather than answering principally from text. Useful evidence concerning the continuing evolution and difficulty of multimodal evaluation.
8. **Li et al. — MIHBench, August 2025.** Research evidence concerning hallucination across multiple images, including object existence, counting and identity consistency. Research benchmark rather than operational deployment evidence.
9. **Dong et al. — MIRAGE, May 2025.** Research evidence distinguishing perception-related and reasoning-related hallucination in multimodal systems and identifying continuing weaknesses in spatial reasoning.
10. **Gao et al. — HAVEN, March 2025.** Research benchmark concerning hallucination in video understanding. Used as evidence that dynamic visual information creates reliability problems beyond those of static-image interpretation.
11. **Cheng et al. — Multimodal Multi-image Reasoning Benchmark, June 2025.** Research evidence concerning the continuing difficulty of structured reasoning across multiple images.

12. **Tariq et al. — Multimodal artificial intelligence models for radiology, BJR Artificial Intelligence, 17 January 2025.** Academic review explaining the clinical rationale for combining imaging with electronic health records and other information while identifying data and deployment considerations.
 13. **Schouten et al. — Navigating the landscape of multimodal AI in medicine, published online 30 May 2025.** Scoping review of technical challenges and clinical applications. Used to distinguish extensive medical research activity from uniform clinical maturity.
 14. **Review of vision-language models in diagnostic imaging, 2025.** Evidence concerning limited external validation, hallucination, generalisability and continuing requirements for human oversight in clinical applications.
 15. **Viktor & Kiss — Multimodal Sensor Fusion in Autonomous Vehicles, 2026.** Systematic review of 66 peer-reviewed studies covering cameras, lidar, radar and other sensors and multimodal fusion architectures. Useful evidence that multimodality extends beyond generative foundation models and is already integral to specialised perception systems.
 16. **Burgess et al. — MicroVQA, CVPR 2025.** Peer-reviewed evidence concerning multimodal reasoning in microscopy-based scientific research and the gap between ordinary visual question answering and specialist research-level interpretation.
 17. **Be My Eyes — 2025–26 product and deployment materials.** Evidence of practical visual-assistance applications for blind and low-vision users, including AI visual interpretation and integration with wearable devices. Company evidence, used to establish availability and application rather than independently verified accuracy or impact.
 18. **NIST — TEVV-Athlon Framework initial public draft, 7 August 2026.** Public-sector evidence concerning the need for application-sensitive testing, evaluation, verification and validation across AI systems including multimodal models. The document was an initial public draft at the reference date rather than a final standard.
 19. **Systematic review of vision-language models, July 2026.** Review of research published from 2021–25 covering vision-language understanding, vision-conditioned generation and multimodal synthesis. Useful synthesis of the field's architectural development and continuing challenges.
 20. **Survey of multimodal LLMs across domains, 31 July 2026.** Recent academic synthesis covering multimodal systems and applications in healthcare, education, industry, governance, security, assistive systems and scientific research. Used principally as broad corroborating evidence of the expansion of multimodal AI across application domains.
-

6. Conversational & Voice AI

Status as at 30 August 2026

Status snapshot

Conversational & Voice AI has moved from principally command-based assistants, scripted chatbots and separate speech-processing tools towards increasingly natural, open-ended spoken interaction with general-purpose AI. Speech recognition and conventional text-to-speech are established technologies with extensive operational use, while foundation-model-based conversational AI is widely available. Real-time speech-to-speech interaction, expressive synthetic speech, multilingual translation and interruption-aware dialogue are newer but increasingly deployed capabilities.

Current systems can sustain flexible conversations, transcribe and generate speech, translate spoken languages, support customer service, clinical documentation, workplace communication, education and accessibility, and increasingly connect spoken requests to retrieval systems and external tools. Maturity nevertheless varies substantially: transcription and synthesis are well established; generative voice interaction is rapidly maturing; full-duplex conversation and sophisticated tool-connected voice agents remain comparatively newer and less uniformly proven in operational settings.

The central direction of development is towards voice becoming a general-purpose interface to wider AI capabilities, with lower latency, more natural turn-taking, greater expressiveness and broader multilingual support.

The principal qualification is that conversational naturalness is not evidence of factual reliability, reasoning ability or autonomous agency. Language and accent performance remain uneven, synthetic voice creates impersonation and trust risks, and product availability does not establish widespread adoption or demonstrated effectiveness. **As at 30 August 2026, AI can converse increasingly naturally, but human-like conversation should not be mistaken for human-like understanding or dependable capability.**

1. What is it?

Conversational & Voice AI comprises AI systems that allow people to interact with machines through ordinary language, whether typed or spoken, together with the speech technologies that enable machines to recognise, interpret and generate human speech.

The category spans several related but distinct capabilities. Conversational AI interprets what a person says or writes, maintains enough context to conduct an exchange and generates an appropriate response. Automatic speech recognition converts spoken language into text or another machine-readable representation. Text-to-speech systems turn written or generated content into spoken audio. More recent speech-to-speech systems can process and produce audio more directly, reducing the dependence on a sequence of separate transcription, language-processing and speech-synthesis stages.

These technologies are not new. Automated telephone systems, dictation software, speech recognition, virtual assistants and customer-service chatbots have existed for years. The important current change is the convergence of increasingly capable language models with improved speech understanding and generation. This is turning conversational interfaces from systems principally designed around commands, scripted dialogue or predefined intents into interfaces capable of much more open-ended interaction.

Voice interaction is also becoming less mechanically turn-based. Earlier systems commonly waited for a person to finish speaking, converted the utterance into text, generated an answer and then synthesised that answer as speech. Increasingly, systems process streaming audio, respond with much lower latency, detect interruptions, interpret vocal characteristics and generate more expressive speech. Some can listen and speak simultaneously rather than enforcing rigid alternating turns. OpenAI's GPT-Live, introduced in July 2026, for example, uses a full-duplex architecture designed to listen and speak at the same time, while Google's live-audio systems similarly emphasise low-latency dialogue, interruption handling and natural vocal rhythm.

This distinction matters because conversation is more than exchanging sentences. Human speech conveys information through timing, pauses, emphasis, pace, pitch and other vocal cues. A capable voice system must therefore manage not only *what* is being said but aspects of *how* and *when* it is being said.

Conversational fluency should nevertheless not be confused with intelligence in every other sense. A system may speak confidently and naturally while providing an incorrect answer. It may conduct an excellent conversation without possessing strong reasoning capability. It may understand a request without being authorised or technically able to act on it. Conversely, a powerful reasoning or agentic system may use voice merely as its interface.

The Technology Landscape therefore contains overlapping layers: speech recognition and synthesis; conversational language understanding; real-time audio interaction; and connections to wider capabilities such as reasoning, retrieval, multimodal perception and tool use. Broader multimodal integration belongs principally within Multimodal AI, and autonomous execution within AI Agents & Agentic Systems. The significance here is that conversation and speech are becoming increasingly capable general interfaces through which people can access those wider AI capabilities.

2. What can it currently do?

Modern conversational systems can sustain substantially more flexible dialogue than earlier generations of chatbots and voice assistants.

They can answer questions, explain concepts, summarise information, help users formulate ideas, translate languages, provide instructions, conduct practice conversations and maintain context across multiple exchanges. When connected to retrieval systems or external tools, they can also search for information, access organisational knowledge, initiate transactions or support workflows.

Speech recognition is sufficiently mature for extensive practical use in dictation, transcription, captioning, customer service, meetings and professional documentation. Contemporary systems can identify speech under varied conditions, distinguish speakers, handle specialised vocabulary where appropriately configured and operate across substantial numbers of languages. Deep-learning and transformer-based approaches have materially improved automatic speech recognition, although performance continues to depend on language, accent, acoustic conditions, domain and training data.

Speech generation has changed equally substantially. Synthetic voices can now produce natural rhythm, emphasis and intonation rather than merely reading words in a recognisably artificial voice. Systems can vary speed, tone and delivery and, in some implementations, reproduce or create particular voice characteristics. The boundary between speech synthesis and generative audio has consequently become less distinct.

The more important development for conversational interaction is real-time speech-to-speech capability. Instead of treating speech simply as text with an audio layer attached, newer systems can work more directly with audio. OpenAI's production Realtime API, for example, supports speech-to-speech interaction and telephony integration, while Google's native-audio systems support low-latency dialogue and tool-connected voice agents. These are commercially available capabilities rather than solely laboratory demonstrations.

Turn-taking is also improving. Conversation historically presented a difficult engineering problem because a system had to decide whether a pause meant that the user had finished speaking or was merely thinking. Responding too early caused interruptions; waiting too long made conversation feel unnatural. Full-duplex and continuously streaming architectures are beginning to reduce this constraint. OpenAI reported in 2026 that GPT-Live removed the separate turn detector from the main audio path, allowing the model to listen while speaking and enabling more continuous interaction. This represents a significant technical development, although it should not be interpreted as evidence that machine conversation now reproduces the full social understanding involved in human dialogue.

Speech translation is another rapidly developing capability. Meta's Seamless research demonstrated streaming multilingual translation with approximately two seconds of latency, while Google subsequently moved live speech-to-speech translation into user-facing products. By June 2026, Google's Gemini 3.5 Live Translate was providing near-real-time translation across more than 70 languages while attempting to preserve features such as intonation, pacing and pitch.

This moves translation closer to simultaneous conversation rather than a sequence of speak–wait–translate–listen exchanges.

Language coverage is nevertheless uneven. Large multilingual speech models have considerably broadened recognition and generation beyond English and a small group of major languages, but lower-resource languages, dialects and accents remain more difficult because training data are less abundant and often less representative. Research published in 2025–26 continues to identify data scarcity, speaker variability, noise, dialect and low-resource adaptation as material challenges.

The current capability boundary is therefore not simply whether AI can “talk”. It clearly can. The more significant question is how reliably it can understand different people and

environments, sustain natural interaction, preserve meaning across languages and connect conversation to accurate information and dependable action.

3. Where is it currently being applied?

Conversational and voice technologies are being applied across both consumer and organisational environments.

General-purpose AI assistants are among the most visible applications. Voice increasingly provides an alternative to typing for interacting with general AI systems. Users can ask questions, brainstorm, learn, translate, practise languages or discuss information conversationally. In some systems, voice interaction is also being combined with visual input, web information and tools, making spoken conversation an interface to capabilities extending well beyond speech itself.

Customer service and contact centres constitute one of the largest organisational application areas. Conversational systems answer routine questions, identify customer intent, retrieve information, route calls, support self-service and increasingly interact with transaction systems. Human service representatives are also using conversational AI as an assistance layer for retrieving information, summarising calls and suggesting responses.

Adoption should not, however, be inferred from the intensity of industry activity. A Gartner survey reported in late 2024 that only 5% of surveyed customer-service leaders already had a customer-facing generative-AI voicebot deployed, while substantially larger proportions were exploring or piloting them. By 2026, investment pressure had risen considerably, but evidence remained mixed regarding realised returns. Gartner reported that service and support leaders invested a median 12% of their 2025 budgets in AI, yet only 24% demonstrated positive financial returns across their AI use cases.

Healthcare provides a different form of voice application. Speech recognition has long supported clinical dictation, but ambient AI systems increasingly listen to clinician-patient conversations—with appropriate consent and controls—and generate draft clinical documentation. Microsoft's Dragon Copilot, for example, combines voice dictation, ambient listening and generative AI for clinical workflows, and has been deployed in healthcare settings including the United Kingdom. These systems illustrate how voice AI can operate in the background rather than requiring users to speak explicit commands.

Meetings and workplace communication use speech AI for transcription, speaker identification, summaries, captions and translation. Here, conversational AI often acts on recorded or live speech rather than engaging directly in dialogue.

Education and language learning use conversational systems for spoken practice, pronunciation assistance, tutoring and role-play. Voice is particularly relevant because learners can practise interactive speech rather than merely read or write responses.

Accessibility is another significant application. Speech interfaces can reduce reliance on keyboards or screens, while speech synthesis can provide communication support for people unable to speak conventionally. Conversational interfaces can also make digital services more

accessible to people whose literacy, physical abilities or circumstances make text interfaces difficult.

Multilingual public services demonstrate why the global voice landscape extends beyond the major commercial AI platforms. India's government-backed BHASHINI infrastructure provides translation, speech-to-text, text-to-speech and related language services across Indian languages. Government reporting in 2026 stated that BHASHINI supported voice in 22 languages, hosted more than 350 AI models and datasets and had processed more than four billion language inferences. Its applications have included multilingual, voice-enabled public-information services.

This is significant because voice can address several access barriers simultaneously. A citizen who cannot comfortably type in a dominant digital language may still be able to speak naturally in a local language. Conversational AI can therefore function not merely as a convenience but as an interface layer between people and digital services.

The application landscape consequently ranges from mature speech-processing functions to newer generative voice agents. Transcription and synthesis are already deeply established. Open-ended, tool-connected conversational voice systems are moving into deployment, but their maturity varies considerably by task, organisation and language.

4. How developed is it?

Conversational & Voice AI cannot sensibly be assigned one maturity level.

At the most established end are **automatic speech recognition and conventional text-to-speech**. Both have long been commercially deployed at scale in smartphones, accessibility systems, dictation, automated telephone services, captioning and other applications. Their accuracy and naturalness have improved substantially, but their existence as operational technologies is no longer experimental.

Text-based conversational AI is also widely available and used at substantial scale. Foundation models have transformed the flexibility of these systems by allowing dialogue across far broader topics than traditional intent-based chatbots.

Generative voice interaction occupies a rapidly maturing middle ground. Natural speech generation, streaming audio input and output, interruption handling and real-time APIs are commercially available. OpenAI's Realtime API reached general availability in August 2025, while Google has made native-audio systems available through consumer products, developer environments and enterprise platforms.

By mid-2026, a further step was becoming visible: movement from alternating speech turns towards more continuous interaction. GPT-Live's full-duplex architecture and Google's development of increasingly low-latency live-audio models illustrate this transition. These developments make voice interfaces substantially more natural, but they are relatively recent compared with conventional speech recognition and should not be treated as equally mature across all deployment conditions.

Real-time speech translation similarly spans research, product availability and deployment. Streaming translation has been demonstrated for several years, while Google had moved near-real-time translation covering more than 70 languages into Google Translate and developer access by June 2026. Yet language coverage alone does not establish equal quality across all languages, accents, domains or conversational situations.

Enterprise voice agents capable of acting through tools are less uniformly mature. Technical platforms can connect speech interaction to telephony, databases, applications and workflow tools. This makes it possible for a conversational system to do more than answer a question—for example, retrieve an account record or initiate an authorised transaction. But dependable operation depends on the complete surrounding system, including identity, permissions, data quality, workflow design, error handling and human escalation.

Customer-service evidence illustrates this distinction. Interest and expenditure are substantial: Gartner reported in August 2026 that AI spending among surveyed customer-service leaders had increased by 38% while their overall budgets increased only 2%. Yet customer behaviour and realised value remained less straightforward; another 2026 Gartner survey found customers approximately three times more likely to use third-party generative-AI tools than company-provided chatbots during service interactions.

The current Status is therefore one of mature underlying speech technology, widespread conversational AI, rapidly advancing natural voice interaction and growing—but still uneven—deployment of conversational systems capable of completing more complex real-world tasks.

5. What is driving its development and use?

Several forces are reinforcing one another.

First is the improvement of **foundation models and conversational language capability**. Earlier voice systems were constrained by relatively narrow dialogue architectures. General-purpose language models can interpret much more varied instructions and maintain richer conversational context, making speech a substantially more useful interface.

Second is progress in **native audio modelling**. Systems increasingly learn directly from and generate audio rather than relying entirely on intermediate text. This can preserve information that transcription may discard, including aspects of tone, timing and vocal expression, while reducing the latency created by passing information through multiple separate models. Google's native-audio work explicitly focuses on these characteristics, while OpenAI's progression from cascaded systems to turn-based audio and then full-duplex interaction demonstrates the same broader technological direction.

Third is **lower latency**. Small delays matter greatly in spoken conversation. Improvements in model inference, streaming, networking and system architecture are allowing AI responses to arrive quickly enough for interaction to feel progressively less like operating software and more like conducting a conversation.

Fourth is the integration of voice with **tools and wider AI capabilities**. Speech becomes much more valuable when the conversational system can retrieve current information, access organisational systems or perform authorised actions. The growth of voice agents therefore partly reflects developments occurring in agentic systems, retrieval and application integration rather than speech technology alone.

Fifth is **economic demand for scalable interaction**. Organisations handle large volumes of customer enquiries, appointments, internal requests and documentation. Conversational systems offer the possibility of handling some of these interactions automatically or assisting human workers. Customer service has consequently become a major focus of AI investment.

Sixth is **multilingual demand**. Much of the world's population communicates most naturally in languages that have historically been poorly served by digital interfaces. Large multilingual models, transfer learning and public language infrastructure are expanding the economic and social case for voice interaction. India's BHASHINI demonstrates one national-scale approach, while global model developers are simultaneously extending multilingual speech and translation capabilities.

Finally, voice is a particularly **low-friction human interface**. People learn to speak before they learn to operate software. In situations involving driving, physical work, disability, small screens or limited literacy, speech can also be more practical than typing. The attraction of conversational AI therefore lies partly in reducing the amount of interface knowledge users require to access increasingly complex digital systems.

6. What are the present limitations?

The most important limitation is that **natural conversation can create an impression of reliability that the underlying system does not necessarily deserve**.

A fluent voice, appropriate pauses and convincing intonation can make an AI response feel socially credible. None of these characteristics guarantees factual accuracy. Conversational systems remain capable of misunderstanding requests, generating incorrect information or responding confidently when uncertain.

Speech recognition itself remains uneven. Noise, overlapping speakers, microphones, specialised vocabulary, speech impairments, accents and dialects can all affect performance. Language inequality remains particularly important. Research continues to find that lower-resource languages have less training data and weaker supporting resources, and bias evaluation for many African and other underrepresented languages remains limited.

Turn-taking remains difficult, despite substantial improvement. Natural conversation contains hesitation, interruption, incomplete sentences, backchannels, silence and simultaneous speech. Full-duplex systems represent an important advance, but determining conversational intention remains different from merely detecting sound.

Latency has not disappeared. Complex reasoning, external tool calls, network delays and retrieval can still interrupt conversational flow. Some newer architectures attempt to maintain

the voice interaction while deeper processing occurs elsewhere, but this introduces additional system complexity.

Integration introduces operational risk. A conversational agent answering a question has limited consequences if it makes a minor mistake. A system capable of altering a reservation, transferring a call, changing an account or initiating another transaction requires stronger authentication, permissions, validation and exception handling. Conversational competence does not itself establish transactional reliability.

Privacy is particularly important for voice. Spoken interactions may contain personal, commercial, medical or other sensitive information. Ambient systems can capture conversations involving multiple people, potentially including people who did not initiate the interaction. Recording, retention, consent and subsequent use of voice data therefore require attention independently of model capability.

Synthetic speech creates another set of problems. **Voice cloning and impersonation** can reproduce recognisable voices sufficiently convincingly to support fraud and deception. The US Federal Bureau of Investigation reported AI-generated voice messages being used to impersonate senior officials, while the Federal Trade Commission has specifically identified voice cloning as a consumer-fraud risk.

Detection is not a complete solution. The FTC has noted that as synthetic voices improve, signals used by detectors can become harder to identify and that no single technical measure provides a complete answer. Provenance, authentication, detection, platform controls and user awareness are consequently developing in parallel.

Regulation is also becoming more material. In the European Union, Article 50 transparency obligations applying from 2 August 2026 require, among other provisions, that people be informed when they are directly interacting with certain AI systems unless this is obvious, and require machine-readable marking of specified AI-generated or manipulated content. Synthetic audio and deepfakes therefore increasingly sit within explicit transparency regimes rather than solely voluntary practice.

Finally, **human preference and trust remain constraints on adoption.** In Gartner's 2026 customer-service survey, 87% of respondents considered access to a human agent essential when companies use generative AI for customer service. This illustrates a wider point: technical ability to conduct a conversation does not establish that people will accept an AI system as the appropriate participant in every conversation.

7. What you should be aware of

The most important development in Conversational & Voice AI is not simply that computers can speak more naturally. Speech is becoming a general-purpose interface to increasingly capable AI systems.

This changes the significance of voice technology. Traditional voice assistants were often command interfaces: ask for the weather, set a timer or choose from a predefined menu. Contemporary systems can engage in open-ended dialogue, interpret more varied language and

connect the conversation to retrieval, reasoning and tools. The interface is moving from **voice command** towards **voice collaboration**.

But several distinctions remain essential.

Fluency is not factual reliability. A system that speaks smoothly and confidently may still be wrong.

A human-like voice does not imply human-like understanding. Expressive timing, laughter, hesitation or empathy-like delivery can be generated computationally. They should not be treated as evidence of consciousness, emotion or human social understanding.

Conversation is not agency. A voice system may explain what should be done without being able to do it. Conversely, when conversational systems are connected to tools, the consequential capability comes partly from those external systems and permissions rather than from speech itself.

Availability is not adoption. Real-time voice APIs and sophisticated conversational products are now widely available, but deployment varies substantially between industries, languages and organisations. Customer-service evidence in particular shows substantial investment and experimentation alongside mixed evidence of realised value.

Multilingual does not mean linguistically equal. A system may technically support many languages while performing considerably better in some than others. Accent, dialect, code-switching and culturally specific expression add further variation. The expansion of national and regional language initiatives such as BHASHINI is therefore an important part of the global landscape rather than a peripheral development.

Voice also changes the trust problem. Text visibly appears as generated information on a screen. Spoken AI enters a communication channel people associate strongly with human identity. As synthetic voices become increasingly natural and replicable, recognising *who or what is speaking* becomes an important part of digital trust. Regulatory transparency requirements, watermarking, provenance mechanisms and fraud countermeasures are developing partly in response to this change.

Most importantly, maturity varies across the category. Speech recognition and synthesis are established technologies. General conversational AI is widely available. Low-latency speech-to-speech interaction has moved rapidly from research into consumer and developer products. Full-duplex conversation and highly capable tool-connected voice systems represent a newer frontier. Reliable service across the world's languages and conversational environments remains uneven.

As at 30 August 2026, Conversational & Voice AI is therefore moving beyond the era in which speech was principally an input or output mechanism attached to software. Increasingly, conversation itself is becoming the interface through which people encounter AI.

The central current development is not that machines have acquired human conversation in its full social and cognitive sense. It is that the technological barriers between speaking naturally to an AI system, having it understand the exchange and connecting that conversation to useful

digital capabilities have become substantially smaller.

Evidence Notes

1. **OpenAI — Introducing GPT-Live, 8 July 2026; engineering description, 3 August 2026.** Primary developer evidence for full-duplex voice interaction, simultaneous listening and speaking, and architectural movement beyond rigid turn-taking. Used as evidence that the capability existed in a user-facing system, not as independent validation of all performance claims.
2. **OpenAI — gpt-realtime and Realtime API, 28 August 2025.** Primary evidence of production availability of speech-to-speech models, telephony integration, tool connectivity and enterprise/developer voice-agent infrastructure.
3. **Google — Gemini 3.1 Flash Live, 26 March 2026; Google DeepMind Gemini Audio / Live Dialogue.** Primary developer evidence concerning low-latency native-audio conversation, vocal characteristics, tool-connected voice agents and product/developer availability. Performance statements remain principally developer evidence.
4. **Google — Gemini 3.5 Live Translate, 9 June 2026.** Primary evidence for near-real-time speech-to-speech translation across more than 70 languages and preservation of vocal characteristics such as pacing, pitch and intonation. Language coverage should not be interpreted as equal accuracy across every language or setting.
5. **Google DeepMind — Gemini 2.5 native audio, 3 June 2025.** Technical/developer evidence explaining the significance of direct audio processing and vocal information beyond textual transcription.
6. **Meta AI — Seamless Communication.** Research evidence concerning multilingual streaming speech translation, expressive translation and approximately two-second streaming latency. Useful as evidence of technical development rather than current mass adoption.
7. **Automatic Speech Recognition: A survey of deep learning techniques and approaches, International Journal of Cognitive Computing in Engineering, 2025.** Research overview of modern automatic speech recognition, multilingual systems, transfer learning and continuing challenges involving data dependence, speaker variability and noise.
8. **Speech Dataset Creation in Low-Resource Contexts: A Systematic Review, Procedia Computer Science, 2026.** Evidence concerning persistent data, representation, infrastructure, annotation and governance constraints affecting lower-resource speech technologies.
9. **Nakatumba-Nabende et al. — systematic review of bias evaluation and mitigation in automatic speech recognition for low-resource African languages, ACM Computing Surveys, 2025.** Evidence of continuing accent, dialect, demographic and evidence-coverage issues in underrepresented languages.
10. **Digital India BHASHINI Division / Government of India, 2026.** Primary public-sector evidence concerning national-scale multilingual speech and language infrastructure, including speech-to-text, text-to-speech, translation, language coverage and reported inference volumes. Government-reported scale figures are treated as programme evidence rather than independently audited measures of effectiveness.
11. **Microsoft — Dragon Copilot, 3 March 2025, and UK deployment material, 24 February 2026.** Company and deployment evidence concerning ambient voice AI, clinical documentation and integration of speech recognition with generative AI in healthcare workflows. Used principally to establish application and deployment rather than independently validated clinical outcomes.
12. **Gartner — customer-service conversational GenAI surveys, December 2024–August 2026.** Evidence distinguishing strong organisational interest and investment from deployment, customer acceptance and

demonstrated financial return. Particularly useful for preventing product availability or expenditure from being treated as equivalent to successful adoption.

13. **European Commission — Guidelines on transparency obligations under Article 50 of the EU AI Act, 20 July 2026.** Primary regulatory evidence. Relevant transparency obligations applying from 2 August 2026 include disclosure of direct interaction with certain AI systems and requirements concerning machine-readable marking and disclosure of specified AI-generated or manipulated content.
 14. **US Federal Bureau of Investigation — Senior U.S. Officials Impersonated in Malicious Messaging Campaign, 2025.** Public-sector evidence that AI-generated voice messages were being used operationally for impersonation and social engineering, demonstrating that synthetic-voice misuse is a realised rather than solely hypothetical risk.
 15. **US Federal Trade Commission — Voice Cloning Challenge and technical discussion of AI-enabled voice cloning.** Public-sector evidence concerning fraud, impersonation and biometric/creative-rights risks and the limitations of relying upon any single detection mechanism.
-

7. AI Search & Research Systems

Status as at 30 August 2026

Status snapshot

AI Search & Research Systems have moved beyond conventional ranked retrieval towards systems that can interpret questions, retrieve and inspect multiple sources, synthesise information and, increasingly, conduct multi-stage research processes. As at 30 August 2026, AI-enhanced and conversational search is already a scaled consumer technology, while deeper research systems that formulate successive searches, analyse documents and construct extended evidence-based reports are commercially available but less mature.

Current systems can substantially reduce the work involved in information discovery, comparison, literature searching and initial synthesis. They are being applied across general web search, professional and enterprise research, market and technical investigation, and scholarly discovery. Specialist scientific systems are also extending AI assistance into screening, evidence extraction and parts of systematic review.

The principal direction of development is from one-step generated answers towards iterative research: systems that decide what information is required, search repeatedly, inspect different information environments and revise their inquiry as evidence emerges.

The central qualification is that retrieval and citations do not make an answer inherently dependable. Systems can select weak sources, miss important evidence, misinterpret material or attach citations that do not fully support their claims. Research completeness, source quality and citation reliability therefore remain material limitations. As at 30 August 2026, these systems are increasingly capable research assistants, but fluent synthesis should not be equated with verified or comprehensive research.

1. What is it?

AI Search & Research Systems are technologies that use artificial intelligence not merely to locate information, but increasingly to interpret a question, decide what information is needed, retrieve material from multiple sources, assess and combine what has been found, and present a synthesized response.

Traditional web search principally retrieves and ranks documents. A user enters words or a question, the search engine identifies potentially relevant pages and normally leaves the user to open those pages, compare them and draw conclusions. AI-enabled search changes this interaction. Systems can accept much longer and more conversational questions, interpret their meaning rather than relying mainly on keyword matching, generate direct answers from retrieved information, support follow-up questions and provide links or citations intended to show where the information came from.

This is now visible at very large scale. Google has progressively incorporated generated answers into Search through AI Overviews and AI Mode. By May 2026, Google reported that

AI Mode alone had surpassed one billion monthly active users globally. In China, Baidu reported that AI-generated content appeared on more than half of its mobile search-results pages by June 2025 and 64% by July. Alibaba similarly transformed Quark from a search and information service into an AI assistant incorporating search, reasoning and research. These company figures establish substantial product availability and usage, although they do not independently establish the accuracy or effectiveness of the generated answers.

A more significant technological step is **multi-stage research**. Instead of performing one retrieval operation and answering immediately, a research system can break a question into sub-questions, formulate multiple searches, inspect results, follow promising sources, revise its search direction as evidence emerges and finally construct a longer synthesis. OpenAI describes its deep-research system as using reasoning to search, interpret and analyse text, images and PDFs while changing course in response to what it finds. Google's Deep Research and Anthropic's Research similarly conduct successive searches rather than relying on a single retrieval operation.

The landscape also includes **specialist research systems**. Scientific tools such as Semantic Scholar use AI to improve discovery across scholarly literature, while systems such as Elicit increasingly assist with literature searching, screening, extraction and evidence synthesis. These differ from general web-research assistants because their information universe, workflows and evidence requirements can be more deliberately constrained. Semantic Scholar, for example, operates over a very large structured scholarly corpus, while Elicit has developed workflows corresponding to stages of systematic review.

Retrieval-augmented generation is an important enabling approach across this landscape. Instead of requiring a language model to answer entirely from information embedded during its training, the system retrieves external material and supplies relevant information to the model when an answer is being produced. The external source may be the public web, a scientific database, an organisation's internal documents or a combination of them.

This distinction is crucial. The underlying language model may be fluent but outdated, incomplete or wrong. Retrieval can provide fresher and more specialised information, but it does not automatically make the resulting answer correct. Information must still be found, selected, interpreted, attributed and synthesized accurately.

AI Search & Research Systems therefore sit between conventional information retrieval, generative AI and agentic systems. Their defining activity is **information discovery and evidence synthesis**. When the same architecture proceeds beyond research to undertake transactions or other actions, the activity increasingly belongs within AI Agents & Agentic Systems.

2. What can it currently do?

Current systems can answer many ordinary information questions by finding web sources and producing a concise synthesized response rather than requiring the user to inspect a page of links. They can support conversational follow-up, allowing a user to refine a question, ask for comparison or request more detail without beginning a new search from scratch.

They are particularly useful where a question requires information scattered across several sources. Current systems can compare products or organisations, summarize competing explanations, identify themes across documents, assemble background material, create an initial market or technical overview and locate sources that a user may not otherwise have found. Microsoft, for example, introduced Copilot Search in Bing as a combination of traditional and generative search that cross-checks information across sites and provides cited responses. Anthropic's web search generates targeted queries, examines retrieved material and supplies citations to source material.

Multi-stage systems can go considerably further. They can decompose a complex task into separate lines of inquiry, conduct successive searches, examine dozens or potentially hundreds of sources and produce multi-page reports. OpenAI's deep research can analyse web pages, uploaded files and PDFs and use code for data analysis. Anthropic's advanced Research was designed to investigate hundreds of internal and external sources and may spend substantially longer on difficult investigations than an ordinary chatbot response.

Research is also moving beyond the open web. Some systems can combine public information with user files, enterprise repositories and connected services. This matters because much professionally useful research depends on both external evidence and information held inside an organisation. The technological direction is consequently towards research across multiple information environments rather than web browsing alone.

Scientific research systems demonstrate a more structured form of capability. Elicit's systematic-review workflow can assist with identifying papers, defining and applying screening criteria, extracting information from full texts and synthesizing evidence, while retaining supporting passages intended to make individual AI-generated conclusions inspectable. Semantic Scholar provides AI-supported search and discovery across a corpus exceeding 200 million scholarly papers.

Research evidence also shows that well-designed retrieval pipelines can materially improve scientific synthesis. A 2026 *Nature* paper on OpenScholar found that retrieval, reranking, iterative feedback and citation verification contributed to stronger literature synthesis and citation accuracy; removing these components degraded performance. This is important because it demonstrates that dependable research is not simply a property of having a powerful language model. Retrieval architecture and verification processes materially affect the result.

Yet current capability has a sharp boundary. These systems can **perform research-like activities**, but they do not thereby guarantee that the research is complete, representative or correct. Their strongest present capability is reducing the human effort required to discover, organise and initially synthesize information. Their ability to establish what should ultimately be believed remains materially less dependable.

3. Where is it currently being applied?

The broadest application is ordinary internet search. AI-generated answers increasingly appear where users previously encountered principally ranked links. Google AI Overviews reached more than a billion users before AI Mode itself subsequently surpassed one billion monthly active users globally. In a Pew Research Center analysis of the browsing activity of 900 US

adults, 58% encountered at least one Google search producing an AI-generated summary during March 2025. The study is US-specific and should not be treated as a global adoption measure, but it demonstrates that generative search had already become part of ordinary information behaviour rather than a specialist experiment.

China demonstrates that this transition is not confined to Western search platforms. Baidu has reworked its search experience around longer prompts, multimodal input and generated answers. Alibaba's Quark incorporated deep search using Qwen models to combine reasoning with online information retrieval, while Tencent, ByteDance, DeepSeek and other Chinese technology groups have also contributed to an increasingly competitive AI-information ecosystem. Baidu's reported hundreds of millions of monthly app users and Alibaba's reported Quark user base indicate that AI-assisted information retrieval is occurring within major Chinese consumer platforms at substantial scale.

General-purpose research assistants are being applied to market research, competitive analysis, technical background work, purchasing comparisons, travel and planning questions, policy investigation, financial and business research and preparation of reports. Their attraction is strongest when the relevant information is dispersed and the alternative requires many separate searches and substantial reading.

Professional use increasingly incorporates internal information. Research systems can inspect connected documents or enterprise data alongside the public web, creating applications in organisational knowledge discovery, due diligence, sales research, strategic intelligence and preparation of internal reports. Anthropic's Research, for example, was introduced with combined access to the web and connected workplace information, illustrating the convergence between internet research and enterprise knowledge retrieval.

Scientific and evidence-intensive fields form another important application. AI-supported literature discovery has existed for considerably longer than current generative search, but newer systems increasingly move beyond identifying papers into screening, extraction and synthesis. Elicit's systematic-review system illustrates this development, while Semantic Scholar represents the more established AI-assisted scholarly discovery layer.

The applications nevertheless occupy different evidential levels. A generated shopping comparison, a corporate background brief and a clinical systematic review may all involve “research”, but they do not require the same completeness, methodological rigour or consequences for error. Current technology spans all of these contexts; its appropriateness and reliability cannot therefore be inferred merely from the shared label.

4. How developed is it?

AI Search & Research Systems contain several maturity levels simultaneously.

AI-enhanced search is already a **scaled consumer technology**. Google's deployment, Baidu's transformation of search and the availability of AI answers from Microsoft, Brave and other providers establish that generative synthesis is no longer an experimental addition to information retrieval. Google reported in 2026 that AI Mode had surpassed one billion monthly

active users, while Baidu had incorporated generated content into a majority of mobile search-result pages during 2025.

Conversational search is similarly established as a product form. Users can ask natural-language questions, receive direct responses with links and continue with follow-up questions. Multimodal input, longer queries and personalized or contextual search are also moving into operational products.

Multi-stage deep research is less mature. It is commercially available from several leading AI providers and is capable of undertaking substantial real research processes, but it emerged primarily from 2024–25 onwards and continues to change rapidly. Google initially presented Deep Research in December 2024 as an agentic research feature; OpenAI released its version in February 2025; Anthropic introduced Research in April 2025. Availability subsequently broadened, but product availability should not be confused with settled research reliability.

Specialist scientific systems occupy a mixed position. AI-assisted scholarly search is established, while AI-supported systematic review and automated evidence synthesis remain developing technologies. Elicit's progression from paper discovery towards structured systematic-review workflows illustrates how domain-specific research tools are moving from assistance with isolated stages towards support for a larger part of the research process. Its published performance figures remain partly developer-generated and should therefore be interpreted accordingly.

The key maturity gap is consequently between **ability to produce an impressive research output** and **demonstrated ability to conduct research dependably enough that the output can be accepted without checking**. Current evidence strongly supports the former. It does not support the latter across general research tasks.

5. What is driving its development and use?

The first driver is the natural fit between large language models and information retrieval. People frequently search because they want an answer rather than a list of documents. Language models can turn retrieved material into an explanation, comparison or summary and maintain conversational context through subsequent questions.

A second driver is improved reasoning and tool use. More capable models can determine that additional information is required, generate searches, evaluate what has been retrieved and repeat the process. This is what transforms a one-step answer engine into a multi-stage research system.

A third is the enormous volume of available information. The difficulty facing many users is no longer merely locating any information, but identifying relevant material among large quantities of web pages, reports, research papers and organisational documents. Scientific literature illustrates the problem particularly clearly: Semantic Scholar's corpus contains hundreds of millions of papers. AI systems offer a way of reducing the human reading and screening burden.

Freshness is another major driver. Conventional language models contain knowledge derived largely from training data and therefore cannot by themselves reliably represent current events or newly published information. Search and retrieval connect generative systems to sources available at the time of the question.

Competition is accelerating development. AI companies have entered search, incumbent search providers have incorporated generative interfaces, and Chinese technology groups are undertaking a similar transformation. Baidu's integration of AI deeply into search and Alibaba's repositioning of Quark demonstrate that this is an international platform shift rather than simply competition among US chatbot providers.

Finally, user convenience is a powerful adoption force. Synthesized answers reduce the effort involved in opening, reading and comparing sources. Pew's observed behaviour illustrates the effect: when a Google AI summary was present, US users clicked conventional search results on 8% of visits, compared with 15% where no AI summary appeared; links contained within AI summaries themselves were clicked in only about 1% of visits. The study does not establish why every user behaved this way, but it demonstrates how answer-oriented search can change the path between a question and its underlying information sources.

6. What are the present limitations?

The central limitation is that **retrieval does not eliminate hallucination**. A system can search the web and still misunderstand a page, choose a weak source, combine incompatible claims, misstate what a source says or generate an assertion that is not supported by the cited material.

Citation presence therefore differs from citation reliability. A hyperlink beside a sentence can create a strong appearance of evidential grounding even where the linked source supports only part of the statement—or something different altogether.

Independent testing has exposed this problem. In 2025 the Tow Center for Digital Journalism tested eight generative search systems using 1,600 queries requiring identification of known news articles. Collectively, the systems produced incorrect answers in more than 60% of the tests and sometimes fabricated or misattributed links. The experiment addressed a particular retrieval task involving news articles rather than the entire range of AI search, so its percentage should not be generalized to every research use. It nevertheless demonstrates that fluent source-linked output is not sufficient evidence of correct retrieval or attribution.

Research completeness is another major constraint. Search systems decide which queries to issue, which results to inspect and when to stop. Relevant information can therefore be absent from the final report even where everything actually cited is correct. Search ranking, language, geography, accessibility, paywalls, indexing and the system's own query strategy all influence what it sees.

This creates particular problems for literature reviews. One 2025 investigation comparing AI research tools with reference lists from four human-authored academic literature reviews found little overlap for most tools, although the test design itself does not establish that every human reference was uniquely correct or that alternative relevant literature was necessarily inferior.

The result nevertheless illustrates the difficulty of equating plausible literature discovery with comprehensive scholarly search.

Source quality is separate from retrieval accuracy. A system may faithfully summarize a low-quality page. The open web contains misinformation, outdated material, marketing content, duplicated reporting, AI-generated pages and sources with undisclosed conflicts. Counting sources does not resolve these differences.

Freshness also remains imperfect. A system technically capable of web search may not necessarily search on every query, may retrieve cached or indexed material, or may fail to locate a newly published development. Testing reported in early 2026 continued to show problems where AI systems responded to rapidly developing news without successfully grounding themselves in current information.

Access restrictions create another boundary. Paywalls, subscription databases, robots controls, authentication requirements and poorly indexed material can prevent systems from retrieving important evidence. Enterprise research partly addresses this by connecting authorised internal repositories, but then introduces privacy, security and access-control requirements.

Multi-stage browsing also creates a security problem not present in ordinary static search to the same degree: a research agent can encounter malicious instructions embedded within websites or documents. OpenAI's deep-research system card specifically identified prompt injection as a risk associated with autonomous web browsing and described mitigations intended to reduce it.

There are economic and infrastructure constraints as well. An ordinary search may require one retrieval operation and a short response. Deep research can involve repeated model inference, many searches, document processing and long-context reasoning. Greater depth therefore consumes more time and computation and may remain subject to usage restrictions or differential service levels.

Finally, research quality is difficult to evaluate automatically. Factual correctness, source quality, completeness, appropriate weighting of contradictory evidence and accurate interpretation are different dimensions. A report can perform well on some while failing badly on another.

7. What you should be aware of

The most important current change is that search is moving from **finding information** towards **interpreting information for the user**.

That can be genuinely useful. AI systems can reduce the mechanical burden of searching, opening documents, extracting relevant passages and assembling an initial synthesis. Multi-stage research systems extend this assistance from seconds-long answers into investigations involving many sources. Specialist systems demonstrate that retrieval, reranking, evidence extraction and citation verification can be deliberately engineered to improve research quality.

But the same development changes where judgement occurs. In conventional search, the user can see multiple competing results before choosing what to read. In generative search, the system increasingly selects evidence and constructs the interpretation before the user sees it. Convenience therefore comes partly from delegating source selection and synthesis to the AI.

This matters because citations should not be mistaken for verification. Three different questions remain: **does the cited source exist; does it contain the claimed information; and is it a sufficiently good source from which to draw that conclusion?** Current systems can fail at each stage.

Nor is a long research report necessarily better grounded than a short search answer. Multi-stage research can examine much more material, but additional searches can also introduce weak evidence, duplicated information or errors that propagate into the final synthesis. The important measure is not how many sources were consulted but how appropriately evidence was identified, interpreted and reconciled.

Scientific research provides both evidence of progress and a warning against generalisation. Structured retrieval and verification can achieve strong results in bounded scholarly environments, while hallucinated references remain sufficiently serious that *Nature* reported in 2026 on evidence of large numbers of invalid citations appearing in scientific publications. AI can therefore simultaneously improve literature synthesis and create new routes by which unreliable references enter the information environment.

The information ecosystem is also being affected. Pew's evidence that users click external links less frequently when AI summaries appear suggests that generative search can alter how readers reach original publishers and information producers. This is not simply a commercial issue: if fewer users inspect underlying material, the practical importance of accurate attribution and source representation increases.

Geography matters. Google's AI search has achieved very broad international reach, while China has developed a substantial parallel ecosystem centred on Baidu, Alibaba, Tencent, ByteDance, DeepSeek and others. Language availability, local content ecosystems, regulation, search indexes and access to information differ between countries. "Searching the web" therefore does not describe a completely uniform global information environment.

Most importantly, **search, synthesis and research should not be treated as synonymous with truth**. Retrieval gives a model access to evidence it might otherwise lack. Multi-stage research allows it to search more extensively. Citations make its conclusions more inspectable. None of these guarantees that the evidence set is complete, the sources are authoritative or the interpretation is correct.

As at 30 August 2026, AI Search & Research Systems have therefore progressed well beyond conventional ranked retrieval. AI-generated search is deployed at very substantial scale; conversational and multimodal information discovery are established product forms; and multi-stage research systems can undertake increasingly sophisticated investigations across public and private information sources. Specialist research systems are beginning to automate parts of evidence synthesis that previously required extensive human effort.

The central Status qualification is equally important: these systems are becoming increasingly capable **research assistants**, but dependable research still requires distinguishing what was

retrieved, what the sources actually establish and what conclusions the evidence justifies.

Evidence Notes

1. **Google — AI Mode and AI Search materials, March 2025–May 2026.** Principal evidence concerning large-scale integration of generated answers, conversational search and multi-stage search into a major global search platform. Google's reported usage figures are company evidence and establish reported scale rather than independent measures of accuracy or user benefit.
2. **Pew Research Center — web-browsing analyses, 23 May and 22 July 2025.** Independent observational evidence from 900 US adults concerning exposure to Google AI summaries and subsequent clicking behaviour. The sample is US-specific and cannot establish global behaviour.
3. **Baidu — Q2 2025 Results, 20 August 2025.** Company evidence that AI-generated content appeared in more than half of Baidu mobile search-result pages by June 2025 and 64% by July, together with reported Baidu App usage. Useful evidence of substantial Chinese deployment; performance claims are not independently verified.
4. **Reuters — Baidu Search overhaul, 2 July 2025.** Independent reporting concerning Baidu's redesign of search around longer, voice and image queries and AI-generated information.
5. **Alibaba Group — Quark AI super-assistant, 13 March 2025; Quark deep-search reporting, May 2025.** Evidence of the development of an important Chinese AI search and research platform integrating Qwen reasoning, online retrieval and reference links. User numbers and capability descriptions originating from Alibaba are company claims.
6. **OpenAI — Introducing deep research, 2 February 2025; Deep Research System Card, 25 February 2025.** Primary technical and product evidence concerning multi-step browsing, source interpretation, PDF and file analysis, and risks including hallucination and prompt injection. These establish system design and developer evaluations rather than independent proof of general research reliability.
7. **Google — Gemini Deep Research, December 2024 onward.** Primary evidence concerning agentic research that plans searches, explores web sources and creates synthesized reports.
8. **Anthropic — Research and web-search announcements, March–May 2025.** Primary evidence concerning iterative research across the web and connected organisational information and cited web-search responses.
9. **Microsoft — Copilot Search and Deep Research, 4 April 2025.** Company evidence concerning the integration of generative synthesis, cited web sources and multi-stage research within Microsoft's consumer AI and Bing ecosystem.
10. **Brave Search — Answer with AI.** Product evidence illustrating AI-generated cited answers built on an independent search index and showing that the transition extends beyond the largest search providers.
11. **Semantic Scholar / Ai2.** Evidence of an established specialist AI-powered scholarly discovery system operating over a corpus containing hundreds of millions of research papers and billions of citation relationships. Corpus-size figures are provider statistics.
12. **Elicit — Systematic Review, 20 February 2025, and subsequent evaluation materials.** Evidence concerning AI-supported scholarly searching, screening, data extraction and synthesis with source-linked supporting passages. Reported accuracy and time-saving figures are substantially developer-generated and should not be treated as independent proof across all research disciplines.

13. **OpenScholar — *Nature*, 2026, “Synthesizing scientific literature with retrieval-augmented language models.”** Research evidence demonstrating the importance of retrieval, reranking, iterative feedback and citation verification to scientific literature synthesis and showing that retrieval-system design materially influences correctness and citation quality.
 14. **Tow Center for Digital Journalism / Columbia Journalism Review — “AI Search Has a Citation Problem”, 6 March 2025.** Independent controlled testing of eight generative-search products using 1,600 news-retrieval queries. Particularly useful evidence that source-linked AI answers can contain incorrect retrieval, attribution and URLs. The task was deliberately specific and its error rates should not be generalized to all forms of AI research.
 15. **Columbia Journalism Review — research-tool evaluation, 19 August 2025.** Evidence concerning limitations of automated scholarly discovery when compared with reference sets used in selected human-authored literature reviews. The test involved only four papers and therefore illustrates retrieval variability rather than establishing general performance rates.
 16. ***Nature* — hallucination and citation reporting, 2025–2026.** Independent reporting concerning fabricated references and the persistence of citation hallucination, including evidence that invalid AI-generated citations were entering scientific publication.
-

8. AI Coding & Software Development

Status as at 30 August 2026

Status snapshot

AI Coding & Software Development has moved from primarily completing or generating code towards participating across a wider software-development process. AI systems can now explain and generate code, debug, refactor, write tests and documentation, review changes and work across repositories. More capable coding agents can undertake bounded multi-stage tasks by inspecting files, editing code, running tools and tests, responding to results and preparing changes for human review.

Maturity varies materially across this landscape. Code completion, explanation and conversational assistance are established and widely used. Repository-aware assistance is also well established as a product capability. Agentic software engineering is commercially available and increasingly deployed for bounded tasks, but dependable autonomous responsibility for complex production software remains considerably less mature.

The principal current direction is from AI that **suggests code** towards AI that **participates in and can be delegated parts of the development process**, including longer-running and increasingly agentic work. High benchmark performance and substantial developer adoption demonstrate significant capability and real use, but neither establishes consistently improved software outcomes.

The central qualification is that functioning code is not necessarily secure, maintainable or contextually appropriate software. Human judgement, verification and oversight therefore remain material. **As at 30 August 2026, AI assistance is mainstream and bounded agentic development is operational, while general autonomous software engineering remains developing rather than mature.**

1. What is it?

AI Coding & Software Development describes the use of artificial intelligence directly within the process of creating, understanding, testing, changing and maintaining software.

Its most familiar form is the coding assistant: a system that predicts or generates code from what a developer has already written or from a natural-language instruction. The category has, however, expanded considerably. Current systems can explain unfamiliar code, locate relevant files, generate functions or entire components, identify possible defects, propose fixes, write tests and documentation, refactor existing software, review proposed changes and translate between programming languages.

Increasingly, they can also work across a software repository rather than responding only to the code visible on screen. A developer may give a system an issue such as “find the cause of this failing test and fix it”; the system can inspect multiple files, search the repository, reason about dependencies, edit code, run commands and tests, observe failures, revise its approach

and prepare a proposed change for human review. This is commonly described as **agentic coding** or an **AI coding agent**.

The distinction matters. Earlier AI coding systems principally assisted a developer while the developer remained continuously engaged. Newer systems can be delegated a bounded software-engineering task and work on it for minutes or hours with varying degrees of independence. By 2026, some environments also allow developers to supervise several coding agents working on different tasks in parallel.

Coding agents therefore overlap with the wider field of AI agents, but their defining environment is software engineering. Their tools are repositories, editors, terminals, build systems, test suites, issue trackers and development infrastructure; their outputs are software changes.

Nor is AI coding synonymous with simply asking a general-purpose chatbot to write a program. Increasingly capable systems combine foundation models with software-development tools, repository retrieval, long context windows, execution environments and feedback from compilers and tests. Specialist coding models and open-weight alternatives have also developed in Europe and China alongside systems originating in the United States.

The most important boundary is between **generating code that appears to work and engineering dependable software**. Production software has requirements extending beyond immediate functional correctness. It must often be secure, maintainable, understandable, appropriately architected, compatible with existing systems and capable of continuing to operate as circumstances change.

AI Coding & Software Development in August 2026 is therefore best understood as a continuum: from highly established code completion and conversational assistance, through repository-aware development support, to increasingly capable but less dependable agents undertaking multi-stage software-engineering work.

2. What can it currently do?

Current AI coding systems are capable of useful work across much of the software-development lifecycle.

At the simpler end, they can generate routine code, complete partially written functions, explain syntax and unfamiliar codebases, convert between programming languages, create documentation and comments, produce regular expressions and database queries, generate test cases and suggest straightforward refactorings. These capabilities are sufficiently established that AI assistance has become an ordinary part of many developers' working environments.

The more significant development is the movement beyond individual snippets. Repository-aware systems can inspect numerous files, identify relationships between components and make coordinated changes. Agentic systems can use terminals and development tools, execute tests, interpret the results and iterate before presenting a proposed solution.

This makes tasks such as fixing bugs, updating dependencies, adding tests, implementing bounded features, migrating interfaces and performing repetitive changes across many files increasingly suitable for AI assistance. Google's Jules, for example, was released from beta in August 2025 after its beta users had produced more than 140,000 publicly shared code improvements. The system could work asynchronously against repositories and return changes for review.

The development of coding-specific agents from [OpenAI](#), [Anthropic](#), [Google](#) and [GitHub](#), together with open and commercially deployable systems from organisations including China's Qwen team and France's [Mistral AI](#), illustrates that repository-scale agentic coding has become an important international direction rather than an isolated research experiment.

Benchmark performance confirms substantial technical progress. SWE-bench tests whether systems can resolve real software issues drawn from GitHub repositories rather than merely answer programming questions. Stanford's 2026 AI Index reported that, by February 2026, leading systems were resolving roughly 70–77% of the human-validated SWE-bench Verified tasks under the benchmark conditions.

That represents a major capability advance, but it should not be interpreted as meaning that an AI can independently complete a similar proportion of ordinary professional software work. Benchmark tasks have defined starting conditions and automated evaluation. Real projects involve incomplete requirements, organisational conventions, architectural trade-offs, undocumented assumptions, security constraints and consequences that may not be captured by whether a test suite passes.

Evidence from real developers reinforces the distinction. METR's controlled study of early-2025 tools involved 16 experienced open-source developers completing 246 tasks in repositories they knew well. Contrary to both their own expectations and expert forecasts, allowing AI increased completion time by approximately 19%. By 2026, METR reported that later agents appeared to provide small productivity benefits in a follow-on experiment, but also concluded that participant-selection effects made the magnitude difficult to estimate reliably.

The current capability is consequently real but uneven. AI can generate substantial amounts of functioning software and can complete some repository-level tasks with remarkably little intervention. Dependable performance across arbitrary complex engineering work remains a materially higher threshold.

3. Where is it currently being applied?

AI coding is now applied across professional software development, open-source development, education and increasingly software creation by people who would not traditionally describe themselves as programmers.

Within professional development teams, common uses include code completion, generating routine implementation, explaining unfamiliar code, searching repositories, producing tests and documentation, debugging, refactoring and reviewing proposed changes. More capable teams are beginning to delegate complete issues or bounded engineering tasks to coding agents.

The scale of general AI-tool use is substantial. Stack Overflow's 2025 Developer Survey, covering more than 49,000 respondents from 177 countries, found that 84% of respondents were using or planning to use AI tools in their development process, while 51% of professional developers reported daily use.

A separate JetBrains survey of 24,534 developers across 194 countries and territories found that 85% regularly used AI tools for coding and development in 2025 and 62% used at least one dedicated coding assistant, agent or AI-enabled editor. By January 2026, a later JetBrains survey reported that 90% of developers surveyed regularly used at least one AI tool at work for coding and development and 74% had adopted specialised developer AI tools. These are survey populations rather than measurements of every developer worldwide, but they establish that AI-assisted development has moved well beyond experimental adoption.

Use is geographically broad. JetBrains deliberately sampled developers across regions including China, India, Japan, South Korea, Brazil, Europe, Africa, the Middle East, Central and South America, and Southeast Asia and Oceania. GitHub's platform evidence similarly shows a rapidly internationalising developer population: India became its largest public and open-source contributor base in 2025, while Brazil and Indonesia were among the other major developer communities.

The technology itself is also geographically distributed. US-developed coding systems coexist with Qwen's Qwen3-Coder and Qwen Code in China and Mistral's Devstral family in Europe. Qwen3-Coder was specifically developed for coding, tool use and agentic interaction, while Mistral's Devstral family was designed for repository-level software-engineering tasks and includes models that can be deployed locally. This is important where organisations require greater control over code, data location or infrastructure.

AI is also lowering the technical threshold for software creation. Natural-language instructions can now produce websites, prototypes, scripts, interfaces and small applications. People with limited conventional programming expertise can therefore create software that previously required more specialised knowledge.

That development has popularised the term “**vibe coding**” for software creation driven heavily by natural-language prompting. It should not be confused with mainstream professional engineering practice. Stack Overflow's 2025 survey found that most respondents did not identify vibe coding as part of their professional work. The distinction is important because generating a plausible application and maintaining production software under security, reliability and governance requirements are different activities.

4. How developed is it?

AI Coding & Software Development cannot sensibly be assigned one maturity level.

Code completion, explanation and conversational assistance are established capabilities. They are embedded in widely available development environments and used at substantial scale. The principal question is no longer whether AI can produce useful code suggestions, but how those suggestions affect quality, productivity and engineering practice.

Repository-aware assistance is also well established as a product capability. Systems can search codebases, reason across multiple files, propose coordinated edits and use information from development environments.

Agentic software engineering is commercially available and increasingly deployed, but remains less mature operationally. Coding agents can take an issue, work asynchronously, execute tools and produce a pull request. GitHub reported more than one million pull requests created by its coding agent between May and September 2025, providing evidence of meaningful real-world use rather than demonstration alone.

By 2026 the interaction model was developing further towards multiple agents and longer-running delegated tasks. OpenAI reported more than four million weekly Codex users in April 2026 and more than five million by June, while also describing growing enterprise use. These figures are developer evidence from the provider rather than independent measurements of effectiveness, but they demonstrate considerable service use.

Broadly autonomous software engineering remains developing rather than mature. Current agents still operate most dependably when objectives can be bounded, relevant context can be made available and results can be checked through tests, review or other verification. The harder the task depends on ambiguous requirements, architectural judgement, organisational knowledge or consequences not captured by automated tests, the less informative benchmark success becomes.

The overall Status is therefore asymmetric: AI assistance is already mainstream in software development; substantial coding can be delegated; bounded agentic development is operational; but reliable autonomous responsibility for complex production software remains considerably less mature.

5. What is driving its development and use?

The first driver is the unusually favourable nature of software as an AI working environment. Programming languages are text-based and highly structured, while enormous quantities of code, documentation and technical discussion exist in digital form. Software can often be compiled, executed and tested automatically, giving an AI system rapid feedback on whether an attempted solution works.

Second, foundation models have become substantially better at reasoning over long contexts and using tools. This enables a coding system to move beyond predicting the next few lines of code towards inspecting repositories, deciding which files matter, invoking terminals and iterating after failures.

Third, software development contains large amounts of repetitive or time-consuming work. Boilerplate generation, documentation, test creation, code migration, searching unfamiliar repositories and routine maintenance provide natural opportunities for assistance. JetBrains' global survey found developers particularly willing to delegate repetitive code, information search, language conversion, documentation and summaries of code changes.

Fourth is strong user demand. Software developers were among the earliest professional groups able to integrate generative AI directly into their normal working environment. High adoption has encouraged AI developers, cloud providers, development-platform companies and open-source communities to devote substantial resources to coding capability.

Fifth is economic pressure to increase software-development capacity. Organisations continually face backlogs of features, maintenance, security remediation, legacy-system work and technical debt. If AI reduces the effort required for even part of that workload, the potential economic value is considerable.

Evidence of realised productivity is nevertheless mixed. Google's 2025 DORA study, based on nearly 5,000 technology professionals worldwide, concluded that AI acted principally as an **amplifier**: it could strengthen effective development environments but could also magnify weaknesses in poorly functioning ones. GitLab's 2026 survey of 1,528 developers and technology buyers across six countries similarly found respondents reporting faster code production while also reporting that the overall software-delivery process had not accelerated to the same extent.

Finally, competition is broadening access. Proprietary frontier systems coexist with open-weight coding models and tools that can be deployed locally or within organisational infrastructure. This matters for organisations that cannot readily send proprietary source code to external services and contributes to the international development of the field.

6. What are the present limitations?

The central limitation is that **functional code is not necessarily good software**.

An AI can produce code that compiles and passes the available tests while still misunderstanding the underlying requirement, introducing an architectural inconsistency, duplicating existing functionality, mishandling an unusual condition or creating a maintenance problem that appears only later.

Context remains a major challenge. Large repositories contain years of design decisions, conventions, dependencies and undocumented organisational knowledge. Although context windows and repository-search systems have improved substantially, providing an AI with more code does not guarantee that it will identify which information is important or interpret it correctly.

Security is particularly significant. Veracode's 2025 evaluation tested more than 100 language models across 80 tasks in Java, Python, JavaScript and C#. It reported that 45% of generated samples failed its security tests. The study used deliberately bounded tasks and Veracode's own static-analysis methodology, so it should not be treated as a measurement of the vulnerability rate of all AI-written production code. It nevertheless demonstrates the important distinction between code that performs the requested function and code that performs it securely.

Subsequent results showed improvement from some newer models rather than a uniform solution to the problem. Security performance therefore needs to be assessed separately from general coding performance.

Agentic coding introduces additional risks because the system does more than produce text. An agent may be permitted to read private repositories, execute shell commands, install dependencies, access credentials or interact with development infrastructure. The security question consequently expands from “is the generated code safe?” to “what can the agent itself access and do?” OpenAI's description of its own Codex deployment emphasises sandboxing, access controls, approval boundaries and telemetry for precisely this reason.

Software supply chains add another complication. AI systems may suggest outdated, unsuitable or nonexistent dependencies; agents that can install packages can turn an incorrect suggestion into an operational security exposure. OWASP's current guidance therefore treats AI-generated code, agent permissions, dependencies, prompt injection and supply-chain integrity as related software-security concerns.

Verification itself can become a bottleneck. If AI generates code faster than humans can understand, test and review it, apparent productivity at the coding stage can shift effort downstream. GitLab's 2026 survey found 85% of respondents agreeing that AI had shifted a bottleneck from writing code towards reviewing and validating it, while 82% expressed concern that AI-generated code could create technical debt their organisations were not prepared to manage. These are survey perceptions, not direct measurements of software outcomes, but they illustrate an emerging operational issue.

Human skills create another uncertainty. AI can make experienced developers faster in some circumstances, but it can also generate convincing errors that require expertise to detect. For less experienced programmers, the ability to produce software without understanding every line may accelerate learning and experimentation, but it can also weaken the connection between code generation and comprehension.

Developer attitudes reflect this tension. Stack Overflow found that 46% of respondents distrusted the accuracy of AI outputs compared with 33% who trusted them, despite widespread usage. Sixty-six per cent identified “almost right” solutions as a frustration and 45% reported that debugging AI-generated code could take more time.

Finally, productivity remains difficult to measure. Lines of code produced are a poor proxy for useful software. Faster completion of an individual coding task does not necessarily mean faster product delivery if additional review, integration, testing, remediation or maintenance follows. The contrasting results from controlled experiments, developer surveys and organisational studies show why claims of universal productivity multipliers should be treated cautiously.

7. What you should be aware of

The most important development is not that AI has “learned to program” in a single definitive sense. It is that progressively larger parts of software-development work can now be expressed as natural-language objectives and delegated to AI systems.

This changes the developer's interaction with software. The progression has been from **writing code**, to **writing with AI assistance**, to **asking AI to make changes**, and increasingly to **supervising agents that perform bounded development work**.

But assistance, delegation and autonomy should not be conflated.

A code-completion system proposes text while the developer remains continuously involved. A repository-aware assistant can understand more of the surrounding project. A coding agent can plan, edit, execute and test with less intervention. None of these facts by themselves establish that the system should be trusted with independent responsibility for production software.

Benchmark performance should likewise be interpreted carefully. Resolving a high proportion of software issues on SWE-bench demonstrates significant capability. It does not measure every aspect of engineering quality, and benchmark tasks cannot reproduce the full combination of ambiguous requirements, organisational context, cybersecurity, maintainability and accountability present in real systems.

The same qualification applies to adoption. High developer usage demonstrates that the technology is useful enough to have become part of ordinary software work. It does not demonstrate that every use increases productivity or that every organisation achieves the same benefit.

AI may also change who creates software. Natural-language development lowers barriers for non-specialists and enables experienced developers to work at a higher level of abstraction. Yet easier code generation can increase the amount of software requiring testing, security assessment and long-term maintenance. More software is not automatically better software.

Human oversight consequently remains material, although its form is changing. Developers increasingly review plans, diffs, test results and agent actions rather than manually typing every implementation detail. The skill emphasis can therefore move towards problem definition, architecture, verification, security, debugging and judgement.

Capability, deployment and maturity should not be inferred automatically from one another. A benchmark result establishes performance under particular test conditions; a released coding agent establishes availability; repeated use establishes adoption; dependable improvement to software outcomes requires stronger evidence. This distinction is central to interpreting rapidly developing AI technologies.

As at 30 August 2026, AI Coding & Software Development is therefore simultaneously mature and immature. AI assistance has become a mainstream part of development. Repository-scale understanding and generation have advanced substantially. Coding agents can already undertake meaningful multi-stage engineering work and are moving into operational use at significant scale.

What has not been established is dependable general autonomy across software engineering.

The central current development is consequently the transition from **AI that suggests code to AI that participates in the software-development process**. The practical frontier is no longer merely whether a model can write functioning code. It is whether increasingly autonomous systems can produce software that remains correct, secure, maintainable and appropriate within the much larger technical and organisational environment in which software actually operates.

Evidence Notes

1. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Technical Performance, 2026.** Provides independent synthesis of coding benchmark performance, including SWE-bench Verified. Particularly useful for demonstrating rapid improvement in repository-level software tasks. Benchmark performance should not be interpreted as equivalent to autonomous production engineering.
2. **Stack Overflow — 2025 Developer Survey, July 2025.** Large international developer survey covering AI adoption, trust, agent use and workflow experience. More than 49,000 respondents from 177 countries participated in the overall survey. Particularly useful for demonstrating simultaneous high use and comparatively low trust in AI-generated results. Survey evidence reflects respondent experience rather than independently measured productivity.
3. **JetBrains — State of Developer Ecosystem 2025.** Survey of 24,534 developers across 194 countries and territories, conducted April–June 2025. Provides unusually broad geographic evidence concerning AI-tool use and developer attitudes. JetBrains notes that some product-user sampling bias may remain.
4. **JetBrains — AI Pulse / developer-tool adoption evidence, April 2026 publication.** Reports January 2026 survey evidence showing widespread use of AI and specialised coding tools among surveyed developers. Useful as more recent adoption evidence; remains survey-based.
5. **DORA / Google — State of AI-assisted Software Development 2025.** Based on nearly 5,000 technology professionals and more than 100 hours of qualitative research. Its finding that AI acts as an organisational “amplifier” is useful in distinguishing individual coding assistance from overall software-delivery effectiveness.
6. **METR — Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, July 2025.** Randomised controlled study involving 16 experienced developers, 246 tasks and mature open-source repositories. Found approximately 19% longer task-completion times when early-2025 AI tools were allowed. Valuable controlled evidence, but deliberately narrow and not representative of all developers, tools or tasks.
7. **METR — developer productivity experiment update, 24 February 2026, and Frontier Risk Report, May 2026.** Follow-on evidence suggests later agents may provide modest productivity improvements, but METR concluded that selection effects prevented a reliable estimate of the contemporary effect. Particularly useful evidence of how quickly capability changes can make earlier productivity measurements obsolete.
8. **GitHub — Octoverse 2025, October 2025, updated February 2026.** Platform evidence concerning the international developer ecosystem and early coding-agent use, including more than one million agent-created pull requests between May and September 2025. Provider/platform evidence rather than an independent measure of resulting code quality.
9. **Google — Jules, May and August 2025.** Primary product evidence concerning asynchronous repository-level coding, including planning, bug fixing, testing and preparation of changes for review. Google's August release reported more than 140,000 publicly shared code improvements during beta. Establishes product capability and usage, not independent effectiveness.
10. **Anthropic — Claude Code, February 2025 onwards; 2026 product record.** Primary developer evidence concerning terminal-based agentic coding and the subsequent evolution of coding capability. Used principally as evidence that delegated repository-level software engineering has become an established product form.
11. **OpenAI — Codex materials, February–June 2026.** Primary developer evidence concerning coding agents, parallel agent operation, repository and terminal access, enterprise deployment and reported service adoption. User numbers and productivity examples are provider-reported and should not be treated as independent measures of effectiveness.
12. **OpenAI — Running Codex safely at OpenAI, 8 May 2026.** Primary evidence concerning operational controls for coding agents, including permissions, technical boundaries, approval requirements and telemetry.

Useful for establishing that increased coding autonomy creates security and governance requirements beyond conventional code generation.

13. **Qwen Team — Qwen3-Coder, 22 July 2025.** Primary Chinese developer evidence concerning an open coding model designed for agentic coding, tool use and long-context software tasks, together with the Qwen Code command-line tool. Benchmark claims are developer-reported.
 14. **Mistral AI — Devstral, Devstral updates and Devstral 2, May–December 2025.** European developer evidence concerning coding models designed specifically for repository-level and agentic software engineering, including open-weight and locally deployable variants. Benchmark and cost-performance claims are developer-reported.
 15. **Veracode — 2025 GenAI Code Security Report and 2026 update.** Evaluation of more than 100 models initially, later expanded beyond 150, using 80 coding tasks across four programming languages and four vulnerability categories. Demonstrates that functional coding performance and secure coding performance are distinct. Results are based on Veracode's selected tasks and static-analysis methodology and should not be projected directly to all production code.
 16. **GitLab — 2026 AI Accountability Report, 23 June 2026.** Survey conducted by The Harris Poll involving 1,528 developers and technology buyers across six countries. Provides evidence concerning faster code generation, downstream review bottlenecks, maintainability concerns and governance. Vendor-sponsored survey evidence; reported productivity and quality effects are respondent assessments.
 17. **OWASP — Top 10:2025 and Secure Coding with AI guidance.** Current security-practice evidence concerning conventional software vulnerabilities and additional issues created by AI-assisted and agentic coding, including dependency, prompt-injection, permissions and supply-chain risks.
 18. **NIST — Secure Software Development Framework and SP 800-218A, updated June 2025.** Public-sector standards evidence reinforcing the continuing importance of secure software-development lifecycle practices as AI becomes integrated into software and AI-system development.
-

9. Image, Audio & Video AI

Status as at 30 August 2026

Status snapshot

Image, Audio & Video AI has become an increasingly established layer of digital media production, spanning generative imagery and editing, synthetic speech and voice, music and sound generation, and rapidly advancing video and audiovisual creation. Its maturity is uneven. AI-assisted image editing, generative still imagery and synthetic speech are commercially well established; music generation is in growing commercial use; while generative video and integrated audiovisual generation are advancing rapidly but remain less consistently dependable.

Current systems can create and iteratively edit high-quality images, generate natural and expressive speech, produce complete music tracks, and generate increasingly realistic short video sequences with greater control over characters, camera movement and sound. These capabilities are already being applied in creative production, marketing, advertising, localisation, social media and other content workflows, although product availability and reported use should not be equated with universal adoption or demonstrated effectiveness.

The principal direction of development is from one-shot generation towards controllable, iterative and integrated production: preserving identities and visual elements, editing selected features, using reference media, and combining imagery, movement, dialogue, music and environmental sound within connected workflows.

The main qualification is reliability and control. Visually or audibly convincing output may still contain physical, temporal, compositional or factual errors, and sustained professional production remains substantially harder than producing an impressive individual asset or short clip. Provenance, rights, consent and authenticity have consequently become increasingly important alongside generation capability.

As at 30 August 2026, the technology is already materially embedded in media creation, but its most advanced audiovisual capabilities remain less mature than their most striking demonstrations can suggest.

1. What is it?

Image, Audio & Video AI describes artificial-intelligence systems that create, modify, transform or help interpret visual and auditory media. The category includes image generation and editing; video generation and editing; synthetic speech and voice cloning; sound-effects and music generation; dubbing and translation; restoration and enhancement; and tools that combine several of these functions within a creative workflow.

The important development is not simply that computers can now generate pictures, voices or moving images. Digital tools have manipulated media for decades. What has changed is the extent to which generative models can produce new media from ordinary language, reference

material or other media, while increasingly allowing users to control composition, identity, style, movement, camera behaviour, speech, music and editing through instructions rather than specialist manual production.

Image generation is the most established generative part of this landscape. Current systems can create photorealistic, illustrative, graphic and stylised imagery from text, modify uploaded images, add or remove objects, extend scenes and preserve important features through successive edits. Systems such as [ChatGPT Images](#) and [Adobe Firefly](#) illustrate the movement from one-shot image generation towards iterative, instruction-based visual editing integrated into wider creative workflows.

Video generation is developing particularly rapidly. Earlier systems were principally demonstrations of short synthetic clips. By 2026, leading systems can generate comparatively realistic moving scenes from text or reference images, maintain characters and settings with greater consistency, control camera movement, extend or edit clips and, increasingly, generate accompanying speech, music and environmental sound. [Google Veo 3.1](#) and [ByteDance Seedance 2.0](#) exemplify this shift towards controllable audiovisual generation rather than silent moving imagery.

Audio AI covers several distinct technologies. Text-to-speech systems synthesise spoken language; voice-cloning or voice-design systems reproduce or construct voices; audio models generate sound effects and ambient sound; and music models can create complete instrumental or vocal tracks from descriptions or reference material. By the reference date, systems can produce expressive multi-speaker dialogue in many languages and increasingly allow control over emotion, pacing and vocal delivery. Music systems similarly allow users to specify genre, instrumentation, structure and lyrics, and to edit or extend generated material.

These categories increasingly converge. A video-generation system may accept images, audio and video as references and generate synchronized pictures and sound; an image system may operate inside a conversational assistant; and creative applications may combine generation, editing, translation and asset management. The present element nevertheless concentrates on the media-generation and editing capability itself. General cross-modal understanding belongs principally in **Multimodal AI**, while low-latency spoken human–AI dialogue belongs principally in **Conversational & Voice AI**.

The landscape is therefore best understood as a continuum: from mature AI-assisted editing and synthetic speech, through widely available generative imagery and music, to rapidly improving but less consistently dependable generative video and integrated audiovisual production. Capability, availability, deployment and maturity need to be distinguished rather than treated as equivalent.

2. What can it currently do?

For still images, current systems can generate high-quality material across a wide range of visual forms. Users can describe subjects, environments, lighting, perspective, typography, artistic treatment and composition; provide one or more reference images; or start from an existing photograph or design and request changes. Editing has become particularly important because it changes generative imagery from a system for producing candidate pictures into a

tool that can participate in iterative production. Objects can be inserted, removed or altered; backgrounds replaced; canvases extended; images restyled; and selected characteristics preserved while other characteristics change.

Image models have also improved at following complex instructions and rendering text within images, although neither capability is uniformly reliable. Detailed spatial relationships, exact layouts, numerous specified objects, factual diagrams and repeated fine edits can still expose inconsistencies. The ability to produce an attractive image should therefore not be confused with the ability to execute every element of a design specification exactly.

Video capability has advanced from short, visibly artificial sequences towards footage that can, under favourable conditions, appear convincingly photographed. Models can generate scenes from text, animate reference images, preserve elements from supplied visual references, specify camera movements and produce vertical as well as conventional landscape formats. Some services offer 1080p or 4K output or upscaling suitable for incorporation into professional production processes.

A particularly important development is **joint audiovisual generation**. Video systems increasingly generate dialogue, environmental sounds, effects and music alongside the visual sequence rather than requiring these components to be assembled separately. Seedance 2.0, for example, was released with text, image, audio and video reference inputs and synchronized stereo audio generation, including voice, ambience and music. Veo 3.1 similarly incorporated generated audio into several video-generation workflows. These are developer descriptions of system capability, not independent evidence that every generated sequence achieves professional production quality.

Benchmark evidence confirms substantial improvement while also illustrating the remaining gap. Stanford's 2026 AI Index reported marked progress on VBench-2.0, a benchmark examining aspects of video quality and controllability. Yet VBench-2.0 itself was developed partly because visually convincing video can still violate physics, commonsense, anatomical correctness or compositional integrity. Separate physical-consistency research found substantial failures across tested video systems on basic physical laws.

Synthetic speech is already capable of highly natural output. Systems can narrate material, generate different speakers, change emotional delivery and synthesize speech across many languages. [ElevenLabs](#), for example, describes its Eleven v3 system as supporting more than 70 languages, multi-speaker dialogue and controls for emotional and non-verbal expression. Such developer demonstrations establish that expressive multilingual synthesis is available; naturalness and consistency still vary by voice, language, context and required performance.

Voice technologies can also support dubbing and localisation. AI can translate speech, synthesize translated dialogue and, in some workflows, modify mouth movements so that translated performances appear more visually consistent with the new language. A Swedish film released in US cinemas in 2025 used AI-assisted visual localisation across thousands of shots while retaining human actors within the production process, illustrating how synthetic-media technology can augment rather than wholly replace conventional dubbing.

Music generation has similarly moved well beyond brief experimental audio samples. Current systems can generate structured songs with vocals and instrumentation, respond to genre and arrangement instructions, use audio references, edit portions of tracks and produce longer

compositions section by section. Systems including [Suno](#) and ElevenLabs' Music v2 demonstrate increasingly controllable song and soundtrack generation. These are commercial-system capability claims; artistic quality remains subjective and does not establish equivalent performance across musical forms.

The overall capability boundary is therefore uneven. AI can already generate media that is usable in many real workflows. It can sometimes produce material that is difficult for an ordinary viewer or listener to distinguish from conventionally produced media. But demanding sequences, exact continuity, complex physical interaction, sustained character performance and fine creative direction can still require repeated generation, manual editing and professional judgement.

3. Where is it currently being applied?

The broadest application is **content creation and editing**. Designers, photographers, marketers, social-media creators and other media producers use generative AI for ideation, image generation, retouching, compositing, background replacement, resizing, variations and visual effects. The technology is increasingly incorporated into applications people already use rather than operating only as a separate experimental service.

Evidence of meaningful adoption exists, although available surveys frequently come from technology suppliers and should be interpreted accordingly. Adobe's June 2026 survey of more than 16,000 creators across the United States, United Kingdom, France, Germany, South Korea, Japan, India and Australia found that 75% described creative AI as integrated or essential to their workflow. A separate Adobe study of professional creative work found substantial use for photo finishing and video visual effects, while physical filming and camera setup remained much less AI-intensive. These are supplier-sponsored surveys rather than comprehensive measures of the global creative economy, but they indicate that creative AI has progressed substantially beyond demonstration among the populations surveyed.

Marketing and advertising are major organisational applications because they require large quantities of visual material in different sizes, languages, channels and variants. Generative systems can create initial concepts, campaign imagery, product backgrounds, localised variations and short video assets, while conventional creative software provides subsequent refinement. Adobe's 2026 global enterprise research reported that nearly half of surveyed organisations had integrated generative AI organisation-wide or across multiple functions for marketing content creation and activation. Again, this is vendor-sponsored survey evidence, but it provides stronger evidence of organisational use than product availability alone.

Film, television and audiovisual production are using AI more selectively. Current applications include concept visualisation, pre-visualisation, visual effects, synthetic backgrounds, footage extension, localisation, dubbing and lip modification. Fully generated video is increasingly credible for short-form material, advertisements, social content and particular effects, but long-form narrative production presents substantially greater continuity and creative-control requirements.

Online and social-media creation is another significant domain. The falling technical threshold for producing images, songs and short videos allows individuals and small

organisations to generate media that previously required specialised equipment or skills. This does not mean that professional production has become unnecessary: selection, art direction, editing, sound design, legal clearance and quality control remain important, particularly for commercial material.

Speech and audio production includes audiobooks, games, education, accessibility, narration, corporate media and multilingual localisation. Synthetic voices allow content to be produced across languages or revised without completely rerecording a performance. European dubbing provides evidence of both active experimentation and resistance: studios and technology providers are exploring AI-supported localisation while actors and professional associations have raised concerns about consent, compensation, performance quality and the use of recorded voices.

Music and sound generation are being used for songs, demos, background music, advertising, games, video soundtracks and creator content. The increasing availability of editing, reference-audio and structural controls means these systems are becoming production tools rather than merely generators of novel songs.

Application is geographically broad, but development and access are uneven. US-based technology companies remain important in image, creative-software and voice systems; European firms contribute significantly in audio and creative tools; and China has become a major centre of generative-video development. ByteDance's February 2026 release of Seedance 2.0, for example, attracted substantial international attention and targeted applications including film, e-commerce and advertising.

The application landscape is consequently broader than the phrase “AI art” suggests. The technology increasingly sits inside ordinary media workflows—sometimes generating an entire asset, but often performing one stage within a larger human-directed process.

4. How developed is it?

Image, audio and video AI cannot sensibly be assigned a single maturity level.

AI-assisted image editing is highly developed. Machine-learning-based enhancement, masking, removal, retouching and other image operations have been incorporated into mainstream software for years. Generative filling and instruction-based editing extend this established base. Commercial availability and substantial user adoption are clear.

Generative still imagery is also commercially mature as a capability, although this does not imply perfect reliability or settled legal treatment. High-quality systems are widely accessible to consumers, developers and businesses. They can produce useful outputs rapidly and participate in iterative workflows. The frontier is increasingly less about whether an image can be generated at all and more about controllability, consistency, editing precision, provenance and integration into production.

Synthetic speech is similarly well established for many uses. Natural-sounding narration and multilingual speech are commercially available at scale. More demanding expressive performance, exact preservation of an individual's identity and low-latency conversational

speech introduce additional challenges, but synthetic voice is no longer principally a research technology.

Music generation has reached growing commercial use. Complete songs, vocals and instrumental material can be generated through widely accessible services, and editing and structural controls are improving. Its technological maturity, however, coexists with unresolved questions about training rights, artist consent, ownership and industry treatment. Commercial availability should therefore not be mistaken for a settled production or legal environment.

Generative video occupies a less mature but rapidly advancing position. Leading systems now produce impressive short audiovisual sequences and offer increasing control over references, characters, camera movement, sound and editing. Google reported that its Flow filmmaking environment had generated more than 275 million videos by October 2025, demonstrating considerable product use, although a generation count does not establish how much material entered professional or commercial production.

Research evaluation still finds meaningful weaknesses. VBench-2.0 explicitly evaluates physics, commonsense, human fidelity and controllability because visually plausible outputs can fail on these dimensions. The Stanford AI Index's 2026 results similarly show that leading video models remain far from saturating this broader evaluation.

Integrated audiovisual generation is earlier still. Systems capable of jointly generating convincing visuals, dialogue, music and environmental sound represent an important advance because they reduce the need to assemble each medium independently. Yet synchronized audiovisual quality is itself becoming a new evaluation problem: research benchmarks introduced in 2025 examine whether generated sound actually corresponds correctly to events, speech and timing in the video.

The Status at 30 August 2026 is therefore one of mature commercial capability in several forms of image editing, image generation and synthetic speech; growing commercial maturity in music generation; and rapid but still uneven maturation in video and integrated audiovisual generation.

5. What is driving its development and use?

Several forces are reinforcing one another.

First is **improvement in generative models themselves**. Better instruction following, reference conditioning, image consistency, temporal modelling, speech synthesis and audio generation have expanded the range of material that can be created. Video models have progressed particularly quickly from producing visually unstable clips towards more coherent scenes with controllable camera behaviour, reference-based characters and synchronized sound.

Second is **greater controllability**. Creative users generally need more than a surprising first result. They need to preserve a character, modify one element without changing everything else, match an established visual identity, control timing or camera movement, revise a section

of music or reproduce a voice consistently. Improvements in reference inputs, inpainting, generative editing and iterative conversational control are therefore as important to practical adoption as raw generation quality.

Third is **integration into existing workflows**. Generative capabilities are increasingly embedded within image editors, video applications, workplace tools, APIs and creator platforms. Adobe's integration of image and video generation into its wider creative environment and Google's incorporation of Veo into applications including Vids illustrate how generative media is moving from standalone demonstrations into production environments.

Fourth is **demand for greater quantities of digital content**. Organisations increasingly need media for multiple platforms, audiences, languages, formats and personalised variants. AI can reduce the time required to create first drafts, variations, adaptations and routine edits. Adobe's 2026 enterprise survey reported that 76% of surveyed organisations believed generative AI had moderately or significantly improved the volume and speed of content ideation and production. This is self-reported, supplier-sponsored evidence rather than an independent productivity measurement, but it helps explain organisational interest.

Fifth is **lowering of production barriers**. A person without specialist illustration, animation, composition or recording skills can now describe an intended result and receive a usable starting point. Professional expertise remains important for demanding output, but the cost of experimentation and initial production has fallen substantially.

Sixth is **global technological competition**. Generative media development is no longer concentrated in one country or one set of laboratories. Chinese video systems have become prominent alongside US-developed systems, while European companies have developed significant voice and audio technologies. This competition is accelerating improvements in quality, cost, speed and access.

Finally, **provenance, rights and regulation are themselves becoming technology drivers**. As synthetic media becomes harder to distinguish visually or audibly from recorded media, developers, standards organisations and regulators are pushing for machine-readable provenance and watermarking. The Coalition for Content Provenance and Authenticity's C2PA standard supports cryptographically signed information about the origin and editing history of media, while Google has extended SynthID watermarking across generated images, audio and video.

6. What are the present limitations?

The first limitation is **reliable control**. Generative systems remain probabilistic: the same instruction can produce materially different results. An attractive output may omit requested objects, alter identities, misplace details or introduce unwanted changes during editing. This matters much more in professional production than in casual experimentation, because consistency across assets and revisions is often essential.

For video, **temporal and physical consistency remain central challenges**. Objects and people must persist through time, interact correctly and obey plausible physical relationships. Hands, bodies, clothing, reflections, object permanence and complex interactions can still

behave incorrectly. VBench-2.0 was specifically designed to move evaluation beyond superficial visual quality towards human fidelity, physics, commonsense and compositional integrity. Research testing physical laws has likewise found substantial shortcomings even in advanced systems.

Long-duration consistency is harder than producing an impressive short clip. Maintaining the same character, environment, narrative logic, lighting, sound, performance and visual details across many shots creates a different problem from generating one successful sequence. Consequently, short-form demonstrations should not be interpreted as evidence that an entire feature film, advertisement or programme can be generated to professional specification without substantial human direction and post-production.

Audio has its own limitations. Synthetic voices can sound extremely natural but may vary in pronunciation, emotion, pacing and consistency. Some highly expressive systems have historically involved trade-offs between quality and latency or predictability. Music generation can produce coherent and impressive tracks while still providing less precise control than an experienced composer, performer or producer may require.

Rights and consent are major environmental constraints. Training data, copyrighted imagery and music, performer likenesses, voice replicas and stylistic imitation have generated continuing legal and contractual disputes. These questions differ between jurisdictions and cannot be reduced to one universal rule. They affect not only whether a model can technically generate an output but whether particular material can safely be distributed or commercially exploited.

Authenticity is another fundamental problem. The same capability that enables legitimate synthetic narration, visual effects or localisation can generate convincing fabricated photographs, voices and video. Fraud, impersonation and deceptive deepfakes consequently form part of the technological environment even though misinformation itself is outside the principal scope of this element.

Detection alone is not a complete solution. NIST work has highlighted the difficulty of generalising synthetic-image detectors to previously unseen generation methods; reported experimental performance fell sharply when detectors encountered a new generator type. This helps explain the increasing emphasis on provenance information and watermarking rather than relying exclusively on classifiers attempting to infer whether media “looks AI-generated”.

Provenance systems also have limitations. C2PA Content Credentials can provide tamper-evident information about creation and editing, and Adobe automatically attaches them to certain Firefly outputs. Google's SynthID embeds signals directly within generated media and is designed to survive common transformations. Neither approach establishes that all unmarked material is authentic, nor can provenance by itself establish whether the events depicted actually occurred.

Regulation is becoming more material. From 2 August 2026, Article 50 transparency obligations under the European Union AI Act began applying to relevant AI-generated and manipulated content. Among other requirements, the regime addresses machine-readable marking of synthetic content and disclosure of deepfakes, subject to its defined scope and transitional arrangements. This represents a significant move from voluntary provenance practices towards legally required transparency in a major market.

Finally, **generation is not the same as creative judgement**. Systems can produce many alternatives rapidly, but deciding what communicates effectively, fits a cultural context, respects rights, meets a brief and should ultimately be published remains a broader human and organisational process. The technology can compress parts of media production without automatically assuming responsibility for the finished work.

7. What you should be aware of

The most important point is that Image, Audio & Video AI has already moved substantially beyond experimental media generation. Still-image generation and editing, synthetic speech and several forms of creative assistance are widely available and increasingly incorporated into real workflows. Music generation is commercially established, while video is advancing particularly quickly.

But “**AI-generated media**” describes very different levels of maturity. Producing one convincing photograph, a short voiceover or a few seconds of striking video is not equivalent to reliably executing a complete professional production. Image, speech, music and video should not be treated as if progress in one establishes equivalent maturity in the others.

A second distinction is between **generation and control**. The technological frontier is increasingly about whether creators can direct, preserve and revise what is generated. Reference images, selective editing, consistent identities, camera instructions, controllable voices and editable music are therefore central developments. A system that occasionally creates an exceptional result may be less operationally useful than one that produces predictable, revisable results.

Third, **visual realism does not establish physical or factual truth**. A generated video may look like photography while depicting physically impossible events. A synthetic voice may be indistinguishable from a real speaker while conveying words that person never said. Photorealism has therefore become a weaker indicator of authenticity.

This is why provenance is becoming part of the media technology landscape itself. C2PA Content Credentials, embedded watermarking such as SynthID and regulatory marking requirements attempt to preserve information about how content was created or modified. By July 2026, C2PA was publishing implementation guidance specifically addressing synthetic and non-synthetic content, while the EU's new AI transparency regime began applying the following month.

Fourth, **absence of a detected AI signal is not proof that media is genuine**. Detection systems may fail on new generators or altered files; metadata may be removed; watermarking is not universal; and much synthetic media is created outside systems participating in provenance schemes. Authentication increasingly requires multiple signals rather than a single detector.

Fifth, creative-media AI is genuinely global. US companies remain prominent, but China's rapid progress in video generation is materially important, European developers are significant in synthetic voice and audio, and creative users worldwide increasingly have access to these

capabilities. At the same time, language coverage, cultural representation, licensing conditions, compute access and regulatory environments remain uneven.

Finally, the current transformation is not adequately described as machines simply replacing artists, photographers, musicians or filmmakers. Evidence from creative workflows shows extensive use of AI for ideation, variations, editing, finishing and effects alongside continued human activity in direction, physical capture, collaboration, selection and accountability. Adobe's 2026 research, for example, found much heavier AI use in particular finishing tasks than in filming and physical camera setup. Because this evidence comes from a major creative-software supplier, it should not be treated as a neutral census of the entire creative workforce, but the task-level pattern is instructive.

As at 30 August 2026, Image, Audio & Video AI is therefore best characterised as a rapidly consolidating **synthetic-media production layer**. AI can already create and transform high-quality images, realistic speech, music and increasingly convincing audiovisual sequences. The most mature capabilities are becoming ordinary components of creative software and content production. Video and integrated audiovisual generation are moving quickly towards greater consistency and control, but remain less dependable for sustained, precisely directed production.

The central current development is consequently not simply that AI can generate media. It is that generation, editing, localisation, sound and visual production are increasingly becoming **instruction-driven, iterative and integrated**—while the parallel problem of establishing the origin, rights and authenticity of digital media becomes more important at the same time.

Evidence Notes

1. **Stanford Institute for Human-Centered AI — AI Index Report 2026, Technical Performance, 2026.** Material independent synthesis of progress in image and video AI, including VBench-2.0 results. Benchmark performance establishes capability under specified evaluations rather than production reliability. [AI Index Report 2026](#)
2. **Zheng et al. — “VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness”, March 2025.** Research benchmark evaluating human fidelity, controllability, creativity, physics and commonsense; particularly material for distinguishing visually convincing video from intrinsically faithful video. [VBench-2.0 paper](#)
3. **Guo et al. — “T2VPhysBench”, May 2025.** Research evidence concerning continuing violations of basic physical laws in generated video. Benchmark evidence rather than production-deployment evidence. [T2VPhysBench](#)
4. **Google — Veo 3.1 and Flow materials, October 2025–January 2026.** Primary developer evidence concerning generated audio, reference-image control, editing, character and setting consistency, vertical video and higher-resolution output. Used as evidence of available capability; performance claims are developer evidence. [Veo 3.1 and Flow](#)
5. **ByteDance Seed — Seedance 2.0, 12 February 2026.** Primary technical/developer evidence concerning joint audio-video generation, multimodal references, editing and controllability. Internal benchmark claims are not treated as independent comparative evidence. [Seedance 2.0 technical release](#)

6. **Reuters — ByteDance's new AI video model goes viral as China looks for second DeepSeek moment, 12 February 2026.** Independent reporting supporting the significance of Chinese generative-video development and its intended applications in film, advertising and e-commerce. [Reuters reporting on Seedance 2.0](#)
7. **OpenAI — ChatGPT Images materials, December 2025–April 2026.** Primary evidence concerning instruction-based image generation, editing, text rendering and iterative modification. Used principally to establish product capability and availability. [ChatGPT Images](#)
8. **Adobe — Firefly materials, 2025–2026.** Primary evidence concerning image generation and editing, video generation and integration into established creative workflows. Claims concerning commercial safety reflect Adobe's own position and applicable product terms rather than an independent determination of legal risk. [Adobe Firefly Video Model](#)
9. **Adobe — Creators' Toolkit Report, 16 June 2026.** Survey of more than 16,000 creators across eight countries, providing evidence of substantial reported creative-AI adoption. Supplier-sponsored survey and not representative of all creators or the global population. [2026 Creators' Toolkit Report](#)
10. **Adobe — “How AI is redistributing creative work”, June 2026.** Mixed evidence from job postings, surveys and qualitative research; particularly useful for task-level distinctions between AI-assisted finishing, ideation, physical production and collaboration. Adobe-sponsored research and partly US/UK focused. [Future of Creative Work research](#)
11. **Adobe / Oxford Economics — 2026 AI and Digital Trends in Content Creation and Management, May 2026.** Global survey evidence concerning organisational deployment of generative AI in marketing and content workflows. Self-reported organisational evidence rather than independently measured productivity. [2026 AI and Digital Trends report](#)
12. **ElevenLabs — Eleven v3 and Music v2, 2025–2026.** Developer evidence concerning expressive multilingual synthetic speech, multi-speaker dialogue, music generation, reference audio and editing. Used to establish commercially available capability rather than independent superiority. [Eleven v3 Music v2](#)
13. **Suno — v5, 23 September 2025.** Developer evidence concerning commercially available song generation, vocals, composition and increased creative control. Qualitative performance claims are vendor claims. [Suno v5](#)
14. **Reuters — “Voice actors push back as AI threatens dubbing industry”, 30 July 2025.** Independent reporting concerning deployment and experimentation in European dubbing, alongside professional concerns over quality, consent, rights and employment. [Reuters dubbing report](#)
15. **NIST — Open Media Forensics / deepfake evaluation materials, 2025–2026.** Public-sector evidence concerning synthetic-media detection and the continuing problems of robustness and generalisation to unseen generators and post-processing. [NIST media-forensics programme](#)
16. **Coalition for Content Provenance and Authenticity — C2PA implementation guidance, July 2026.** Standards evidence concerning tamper-evident Content Credentials for communicating whether media is synthetic, modified or non-synthetic. Provenance information should not itself be interpreted as proof that depicted content is factually true. [C2PA implementation guidance](#)
17. **Google DeepMind — SynthID and SynthID Detector, 2025–2026.** Primary evidence concerning imperceptible watermarking across images, audio and video and subsequent detection. Google itself notes that watermarking is one component of a broader authenticity approach rather than a universal solution. [SynthID](#)
18. **European Commission — AI Act Article 50 transparency guidance and Code of Practice, June–August 2026.** Primary regulatory evidence. Transparency obligations applying from 2 August 2026 address machine-readable marking of relevant AI-generated or manipulated content and disclosure of deepfakes, subject to scope and transitional provisions. [European Commission Article 50 guidance](#) [Transparency Code of Practice](#)

10. AI Memory, Context & Personalisation

Status as at 30 August 2026

Status snapshot

AI Memory, Context & Personalisation is becoming an important enabling layer of applied AI, allowing systems to work with larger bodies of information, maintain continuity across interactions and adapt responses using relevant knowledge about users, tasks and previous activity. Its components are at different stages of maturity. Long-context processing and retrieval-supported context are established product capabilities, while persistent personal memory is operational in major consumer systems but remains less mature as a dependable long-duration capability. More adaptive, structured and agent-oriented memory remains an active area of research and development.

Current systems can analyse extensive documents and conversations, retrieve relevant information from external stores, retain selected preferences or prior information across interactions and use connected personal or organisational data to provide more contextualised assistance. These capabilities are already being applied in general-purpose assistants, enterprise knowledge systems, project and task continuity, and some personalised digital services.

The principal direction of development is shifting from simply increasing how much information AI can retain or access towards managing memory more intelligently: deciding what to preserve, retrieve, update, supersede, ignore or delete.

The central qualification is that context capacity, stored information, persistent memory and effective personalisation are not equivalent. Larger or longer-lasting memory does not guarantee accurate retrieval, current information or better decisions. As at 30 August 2026, persistent and personalised context is a significant deployed capability, but reliable lifelong memory with consistent accuracy, appropriate forgetting and effective user control remains developing.

1. What is it?

AI Memory, Context & Personalisation describes the technologies that allow an AI system to use information beyond the user's immediate instruction: information already present in a conversation, retrieved from earlier interactions or documents, deliberately retained about a user or task, or used to adapt the system's responses to an individual.

The distinction between these mechanisms is important because they are often described collectively as “memory” even though they work differently.

A **context window** is the amount of information a model can consider during a particular inference or interaction. It can contain the current conversation, documents, images, tool results, instructions and other material supplied to the model. Context windows have expanded substantially: commercially available systems by the reference date supported hundreds of thousands of tokens, while some models supported approximately one million tokens. This

allows an AI system to work directly with much larger documents, conversations and collections of material than earlier systems could.

A large context window is not the same as persistent memory. Unless information is preserved elsewhere and subsequently reintroduced, it need not survive into a later interaction.

Retrieval-supported memory addresses this problem by storing information outside the model—such as conversation records, files, summaries, database entries or numerical representations of their meaning—and retrieving relevant portions when required. This resembles retrieval-augmented generation used for document search, but the retrieved material may concern a particular user, task or continuing relationship rather than general information.

Persistent memory carries selected information across interactions. It can include preferences, previous decisions, recurring activities, project information or other facts considered useful later. Some systems explicitly store individual memories; others synthesise information from conversation history or maintain structured summaries. OpenAI, for example, described a 2026 memory architecture that synthesises information from conversations in an attempt to preserve relevant and current user context over long periods. Google's Gemini had already introduced the ability to reference previous chats and learn user preferences, while China's Kimi provides a user-facing memory store.

Personalisation is the use of such context to alter what the system does. It may be simple—remembering a preferred writing style—or more consequential, such as adapting recommendations, retrieving information from personal accounts, continuing a long-running project or resolving an ambiguous instruction according to previously observed preferences.

A further distinction is necessary between system memory and **model memorisation**. Foundation models can retain patterns and sometimes information encountered during training within their learned parameters. That is not equivalent to a controllable user memory. A memory system ordinarily involves information that can be stored, retrieved, updated or deleted separately from the underlying model.

The category therefore spans several layers: increasingly long working context; external retrieval and storage; persistent records across interactions; mechanisms for deciding what should be remembered or forgotten; and systems that use these capabilities to personalise subsequent behaviour.

This is becoming an important enabling layer of contemporary AI rather than simply a convenience feature. An AI assistant that remembers nothing must repeatedly reconstruct its understanding of a user or task. An AI system with dependable memory can potentially maintain continuity across days, months or projects. For agentic systems, memory can also preserve goals, decisions, previous actions and task state between stages of extended work.

2. What can it currently do?

The most mature capability is maintaining substantial **within-session context**.

Long-context models can analyse large documents, lengthy conversations, code repositories and collections of information without requiring every relevant item to be manually reduced to a short prompt. Million-token context windows are commercially available in some model families, while others support several hundred thousand tokens. This represents a major expansion from the comparatively short contexts of earlier generative AI.

Large capacity does not mean that every item within that capacity is used equally reliably. Long-context research continues to find that performance can deteriorate as inputs become longer or more distracting, particularly where reasoning rather than simple retrieval is required. The earlier “lost in the middle” phenomenon—difficulty using information positioned within long inputs—has improved substantially in some newer systems, but controlled 2026 evaluations continued to find positional and length-related failures under more demanding conditions.

Persistent systems can now do something different: **carry selected knowledge between conversations**.

Current consumer assistants can remember facts or preferences, retrieve relevant previous conversations and use them when producing later responses. ChatGPT's memory architecture, Gemini's past-chat personalisation and Kimi's Memory Space provide different examples of this broader movement from stateless conversation towards continuing user context.

Memory can also be maintained through external files or stores. Anthropic demonstrated that models given access to local files could create and update their own memory records during extended tasks. Research systems increasingly use combinations of summaries, vector retrieval, knowledge graphs, episodic records and structured user profiles rather than attempting to place an entire interaction history into every prompt.

These systems can support **personalisation**. An assistant can remember preferred formats, interests, recurring constraints or information about an ongoing project and use those details without the user repeating them. More advanced implementations can combine memory with information retrieved from connected services. Google's Personal Intelligence, introduced in 2026, for example, can—with user permission—draw on information from Gmail, Photos, YouTube and Search to provide answers based on personal context.

Research is also demonstrating more sophisticated possibilities. Systems can distinguish relatively durable user preferences from individual events, maintain structured representations that change as new evidence arrives, and use remembered preferences to interpret underspecified instructions. Research presented at ACL 2026 demonstrated approaches involving evolving user-memory structures, preference updating and personalised agent actions.

The qualification is that these capabilities remain imperfect. Remembering information is only the first problem. A useful memory system must decide **what to retain, what to retrieve, what has changed, what is no longer relevant and when memory should not influence the answer**. Those remain active research and engineering problems.

3. Where is it currently being applied?

The most visible application is in **general-purpose AI assistants**.

Persistent memory allows conversations to continue without users repeatedly restating their preferences, background or earlier work. ChatGPT and Gemini illustrate the movement towards assistants that can use information accumulated across conversations, while Kimi demonstrates that similar development is occurring within China's large AI ecosystem.

A second important application is **knowledge and project continuity**. AI systems used for research, coding, document production and other extended tasks increasingly need to retain information beyond a single model invocation. Project instructions, files, previous outputs, decisions and task state can be retrieved when work resumes. Anthropic's work on memory files illustrates one implementation; its enterprise document cataloguing similarly uses retrieval to make organisational information available as working context.

Enterprise AI provides a broader form of contextualisation. Rather than remembering principally who an individual is, systems can retrieve authorised information from organisational documents, databases and connected applications. This allows an assistant to answer within the context of an organisation's own information rather than relying only on the model's general training knowledge. This overlaps with AI Search & Research Systems, but becomes part of memory and context when retrieved information maintains continuity around particular users, projects or tasks.

Personal context is also being incorporated into **connected digital assistants**. Information from email, calendars, photos, search histories and other applications can provide much richer context than conversation history alone. Google's Personal Intelligence is a prominent current example, although availability has expanded geographically in stages rather than appearing everywhere simultaneously.

Conversational and companion systems are another important application because sustained interaction depends heavily on remembering previous exchanges, relationships and preferences. Alibaba's Qwen role-playing system, for example, provides a long-term-memory mechanism that periodically summarises conversation history so that extended interactions can continue beyond its normal context limit.

Research is increasingly extending memory into **personalised agents**. Here memory does not merely influence wording: it can affect what action an agent chooses. ACL 2026 research demonstrated systems using remembered user preferences and routines to resolve vague instructions or tailor subsequent actions. These remain principally research demonstrations rather than evidence that reliable personalised autonomous agents have reached broad deployment.

The geographical landscape is therefore broader than a single cluster of US developers. Major US systems have pushed persistent personal context into consumer and enterprise assistants; Chinese systems provide their own long-term-memory and personalisation mechanisms; and research groups internationally are developing new memory architectures. Deployment remains uneven, however, because privacy rules, languages, product availability and access to

connected personal data differ between jurisdictions.

4. How developed is it?

AI Memory, Context & Personalisation contains technologies at materially different stages of maturity.

Long-context processing is established product capability. Context windows measured in hundreds of thousands of tokens are no longer unusual among leading models, and million-token context is commercially available. The engineering capability to supply very large amounts of material to a model has therefore progressed substantially.

However, **dependable use of long context is less mature than context capacity itself.** The maximum advertised window describes how much material a system can accept, not how accurately it can identify, combine and reason over everything within that material. Research continues to expose degradation as context becomes long, complex or filled with competing information.

Retrieval-supported context is comparatively mature. Retrieval-augmented systems are widely incorporated into enterprise and assistant architectures. Searching files, organisational knowledge or earlier conversations and inserting relevant results into the current context is now a practical system-design pattern rather than primarily a research concept.

Persistent personal memory has moved into operational consumer products but is still developing. Major assistants can retain or retrieve information across conversations, and user controls for memory are increasingly visible. OpenAI's June 2026 description of its revised architecture is revealing: the company identified freshness, correctness and scalability over multi-year horizons as continuing memory challenges even while deploying the technology to very large user populations.

Adaptive long-term memory is less mature. Research continues to explore how memories should be selected, compressed, organised, updated and forgotten. AAAI and ACL research in 2026 proposed hierarchical memory, preference-aware updating, graph-based retrieval and evolving user-profile structures. The existence of several competing approaches indicates both substantial progress and the absence of a settled architecture.

The same applies to **lifelong personalisation.** MemoryCD, a 2026 benchmark built around long-term cross-domain user behaviour, found that existing memory methods remained well short of satisfactory performance across its evaluated personalisation tasks. This is benchmark evidence rather than a direct measure of every commercial system, but it demonstrates that maintaining a coherent understanding of a person across years and different domains remains an open problem.

The current Status is consequently not “AI now remembers everything.” It is that persistent and personalised context has become a real product capability, while dependable long-duration memory management remains an active engineering and research frontier.

5. What is driving its development and use?

The first driver is straightforward: **continuity makes AI more useful**.

Repeatedly explaining the same project, preferences or circumstances imposes friction. As AI systems become recurring assistants rather than occasional question-answering tools, users increasingly expect them to retain relevant context. Memory allows an interaction to begin from an accumulated understanding rather than from zero.

Second is the growth of **longer and more complex tasks**. Coding, research, document analysis and agentic workflows may extend across many steps or sessions. Such systems need to preserve plans, observations, previous actions and unresolved issues. Research on long-horizon agents increasingly treats memory management as an active process rather than passive storage.

Third is the rapid expansion of **context-window capacity**. Larger windows allow models to work directly with more information and provide the working space into which retrieved memories, documents and tool outputs can be assembled. Improvements in inference infrastructure and context-handling methods have therefore enabled richer memory architectures.

Fourth is **retrieval technology**. Vector search, document indexing, structured databases, knowledge graphs and summarisation provide mechanisms for selecting small amounts of relevant information from stores far larger than a model's immediate context. This makes persistent memory economically and computationally more practical than continually replaying complete histories.

Fifth is **competition to build personalised assistants**. Major AI developers increasingly present personal context as part of the transition from generic chatbot to continuing assistant. The relevant technological competition is therefore no longer only over what a model knows or how well it reasons, but how effectively the surrounding system can understand the particular user and task.

Finally, agentic AI strengthens the requirement. An agent that acts over time needs some representation of what it has already done and why. Research in 2026 increasingly connects memory with action, demonstrating that persistent context is becoming part of the infrastructure required for longer-running AI systems rather than merely a personalisation feature.

6. What are the present limitations?

The first limitation is **retrieval accuracy**. Storing information does not guarantee that the system will retrieve the right information at the right time. A relevant memory can be missed; an irrelevant one can be retrieved; or several individually accurate memories can be combined incorrectly.

Second is **staleness**. People, organisations and projects change. A remembered preference that was correct six months ago may now be wrong. Persistent systems therefore need mechanisms

for updating, superseding or forgetting information rather than simply accumulating it. Research on preference updating and evolving memory structures demonstrates that this remains an active technical problem.

Third is **memory selection**. Retaining everything is neither technically efficient nor necessarily desirable. Systems must distinguish durable information from temporary remarks, hypothetical statements and context-specific preferences. Research published in 2026 found that determining what is worth storing for different users remains a significant unresolved challenge.

Fourth, **more context can itself impair performance**. Very large contexts contain more potentially distracting information. Research continues to find that reasoning can deteriorate when relevant evidence is surrounded by large amounts of competing material. Long context therefore creates a context-management problem rather than eliminating one.

Fifth is **over-personalisation**. Previous interactions can anchor future responses too strongly. An AI that continually interprets a user through earlier behaviour may narrow alternatives, reinforce outdated assumptions or reproduce earlier errors. ACL 2026 research explicitly identified “memory anchoring”, in which accumulated history constrains later responses excessively.

Privacy is an unusually important limitation because useful personal memory may contain exactly the information that requires strongest protection: relationships, communications, work activities, interests, documents, location information or data retrieved from personal accounts.

This creates questions about consent, storage, access, retention, deletion and the distinction between information used for personalisation and information used for model training. Product developers increasingly provide controls to disable memory, use temporary conversations, review retained information or delete it. Google introduced Temporary Chats alongside its past-chat personalisation controls, while OpenAI provides controls over memory and temporary interactions.

Regulation reinforces these concerns internationally. European data-protection authorities have examined how GDPR principles apply to personal data used in AI models. Singapore's Personal Data Protection Commission has issued guidance covering personal data in AI recommendation and decision systems. China applies personal-information requirements to generative AI and, in 2026, introduced specific rules for anthropomorphic AI interaction services that include protections for interaction data and deletion rights.

Deletion is technically more complicated than a simple “forget” command may imply. Information can exist in chat histories, extracted memories, summaries, files, indexes, connected services or model-training datasets under different arrangements. Effective user control therefore depends on the architecture of the whole system, not merely on whether a visible memory entry can be removed.

Finally, personalisation can interact with **safety**. Research presented at ACL 2026 found that benign personal memories could alter how models interpreted harmful requests and reduce safety performance in tested memory-augmented systems. This does not establish that all personalised systems become less safe, but it demonstrates that safety evaluations performed

on stateless models may not fully represent systems operating with extensive personal context.

7. What you should be aware of

The most important distinction is that **context, memory and personalisation are related but not interchangeable**.

A million-token context window is not a million-token permanent memory. It describes the amount of information a model can receive at one time. Persistent memory requires another mechanism to preserve information between interactions.

Likewise, **stored information is not necessarily usable memory**. A system may possess years of conversation history but still fail to retrieve the particular fact needed for the current task. The quality of memory depends on selection, organisation, retrieval, updating and use.

More memory is therefore not automatically better. Relevant, current and appropriately weighted memory can improve continuity. Excessive, outdated or incorrectly inferred memory can make an AI less useful. The emerging technological challenge is increasingly **memory management**, not simply memory capacity.

Personalisation also changes the nature of an AI system. A generic model largely responds to the information supplied in the present interaction. A personalised system may combine the same model with accumulated knowledge about the user, retrieved documents, previous conversations and connected services. Two people giving the same instruction may consequently receive different responses because the surrounding system supplies different context.

This distinction becomes particularly important for agents. Memory can allow an agent to preserve goals, plans, previous actions and user preferences across extended activity. It therefore supports greater continuity and potentially greater autonomy. But memory does not itself establish reliable agency: remembering what happened previously does not guarantee correct reasoning or correct action next.

User control is correspondingly central. Persistent AI introduces a continuing question that largely stateless software did not: **what should the system be allowed to remember?** Product architectures increasingly include mechanisms for viewing, editing, disabling or deleting memory, but implementations differ and regulatory expectations are developing across jurisdictions.

There is also a difference between **personalisation and accuracy**. Knowing that a user prefers concise answers can improve usefulness. Remembering that the user once stated an incorrect fact does not make that fact correct. A personalised system must therefore distinguish information about the user's preferences or circumstances from independently verifiable claims about the world.

Finally, apparent familiarity should not be confused with human memory or understanding. Current systems construct continuity through context windows, stored representations, retrieval, summarisation and learned inference. They can produce the experience of an assistant

that “knows” a user, but that experience is generated by technical mechanisms with imperfect selection, retrieval and updating.

As at 30 August 2026, AI Memory, Context & Personalisation has moved decisively from an experimental feature towards a significant layer of the applied AI landscape. Large context windows are established capabilities. Retrieval-supported context is widely used. Persistent memory and past-interaction personalisation are present in major consumer systems. Research is progressing towards structured, adaptive and agent-oriented memory.

The frontier has consequently shifted. The principal question is no longer simply whether an AI system can retain information. It is whether it can **retain the right information, retrieve it when relevant, update it when circumstances change, ignore it when inappropriate and give users meaningful control over what persists.**

Evidence Notes

1. **OpenAI — “Dreaming: Better memory for a more helpful ChatGPT”, 4 June 2026.** Primary developer evidence concerning the evolution from explicit saved memories and past-chat context towards synthesised persistent memory, including the problems of freshness, correctness and scalability over long time horizons. Used as evidence of product architecture and deployment, not as independent proof of memory accuracy.
2. **Google — Gemini past-chat personalisation, 13 August 2025; Personal Intelligence, 14 January 2026; Australian expansion, 13 May 2026.** Primary evidence that a major consumer assistant can use previous conversations and, with permission, information from connected personal services to personalise responses. Geographic availability has varied during rollout.
3. **Kimi — Memory Space documentation, current by August 2026.** Developer evidence of persistent user memory within a major Chinese AI assistant, including explicit storage limits. Used to demonstrate implementation rather than comparative performance.
4. **Alibaba Cloud / Qwen — long-term memory documentation.** Developer evidence of conversation summarisation and persistent context for extended Chinese-language role-playing interactions beyond the underlying context window.
5. **Anthropic — Claude 4, 22 May 2025.** Developer evidence showing models using externally available files to create and maintain memory records during long-running tasks. This demonstrates a memory mechanism rather than independent evidence of general long-term reliability.
6. **Google AI for Developers — Gemini model documentation, August 2026; OpenAI — GPT-5 documentation.** Primary technical evidence of commercially available long-context systems, including approximately one-million-token context in the cited Gemini model and 400,000-token context in GPT-5. Maximum context capacity should not be interpreted as uniform reasoning performance across the entire window.
7. **Liu et al. — “Lost in the Middle: How Language Models Use Long Contexts”, 2023.** Foundational research demonstrating positional weaknesses in long-context information use. Later models have improved materially on some retrieval tests, so the original result should not be treated as a fixed characteristic of every 2026 model.
8. **Zhang et al. — “Positional Failures in Long-Context LLMs”, May 2026.** Recent controlled evaluation showing that positional and length-related degradation can persist on reasoning tasks even where simpler long-context retrieval has improved. Preprint evidence rather than a universal evaluation of deployed systems.

9. **Huang et al. — Mem-PAL, AAAI 2026, 14 March 2026.** Peer-reviewed evidence concerning hierarchical retrieval-supported memory for personalised multi-session dialogue; includes a Chinese multi-session dataset and benchmark.
 10. **Zhang et al. — MemoryCD, March 2026.** Benchmark of long-context, cross-domain user memory using longitudinal behaviour, finding substantial remaining limitations in existing memory methods. Benchmark results should not be interpreted as direct measurements of every commercial assistant.
 11. **ACL 2026 research — Memory as Action; Preference-Aware Memory Update; Inside Out; MemORAI; PersonaAgent; PersonalAlign.** Collectively demonstrates active research into autonomous context curation, changing preferences, structured user memory, provenance-aware retrieval and the connection between memory and personalised agent behaviour. These principally establish research capability rather than scaled commercial deployment.
 12. **Huang et al. — “Controllable Memory Usage”, ACL 2026.** Research evidence concerning “memory anchoring” and the possibility that excessive reliance on previous interactions can constrain later outputs.
 13. **Guo et al. — “When Personalization Legitimizes Risks”, ACL 2026.** Research evidence that personalised memory can interact with model safety in ways not visible in stateless evaluation. Results concern tested systems and should not be generalised automatically to all deployed assistants.
 14. **European Data Protection Board — Opinion on AI models, 18 December 2024.** Regulatory evidence concerning application of GDPR principles to personal data in AI development and deployment.
 15. **Singapore Personal Data Protection Commission — Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems, 1 March 2024.** Regional regulatory guidance illustrating the relevance of consent, transparency and personal-data obligations to AI personalisation.
 16. **China — Interim Measures for the Administration of Anthropomorphic AI Interaction Services, announced 10 April 2026 and effective 15 July 2026.** Primary regulatory evidence concerning interaction-data protection, user deletion options and safeguards surrounding persistent AI relationships in China. Its scope is specifically anthropomorphic emotional-interaction services rather than all AI memory systems.
-

11. Edge & On-Device AI

Status as at 30 August 2026

Status snapshot

Edge & On-Device AI is the increasingly important part of the AI landscape in which computation occurs on, or close to, the device generating the data rather than depending continuously on remote cloud infrastructure. It spans very different levels of maturity. Specialised local inference—including image processing, speech functions, biometrics, sensor classification and vehicle perception—is already mature and widely deployed. Dedicated AI processors are also established across smartphones and increasingly personal computers, vehicles and embedded equipment. More capable generative and multimodal AI running locally is commercially real but remains less mature and more constrained.

Current systems can analyse images and sensor data, recognise speech, translate and summarise text, support personal-computing functions, assist vehicle and robotic perception, and perform industrial monitoring and control with reduced latency and connectivity dependence. The principal development is the growing amount and sophistication of AI that can operate locally as hardware improves and models become more computationally efficient.

The important qualification is that on-device AI does not mean the end of cloud AI. Memory, processing capacity, energy, heat and software constraints still limit local systems, particularly for larger generative models. Many current architectures are therefore hybrid: routine, private or latency-sensitive processing occurs locally while more demanding workloads use remote computing.

As at 30 August 2026, edge inference is already deeply established, while increasingly capable local generative and multimodal AI represents a significant but still developing extension of that foundation.

1. What is it?

Edge & On-Device AI refers to artificial-intelligence processing performed on, or close to, the device where data are generated rather than requiring every input to be sent to a distant cloud data centre.

On-device AI is the clearest form: a smartphone, personal computer, vehicle, camera, robot, appliance or sensor runs an AI model using its own processor and memory. **Edge AI** is somewhat broader. Processing may occur on the originating device, on an industrial computer beside a machine, on a local gateway serving several sensors, or on other computing infrastructure located near the activity rather than in a remote central cloud.

The distinction is principally about **where computation occurs**, not how large or sophisticated the model is. A tiny anomaly-detection model running inside a vibration sensor is on-device AI. So is a substantially larger language or vision model running on a laptop, smartphone or

robot. Conversely, a relatively small model executed in a distant data centre is not on-device AI merely because the model itself is small.

Much of the established edge landscape consists of **inference**: a model has usually been developed or trained elsewhere and the finished model is then placed on a device to recognise an image, detect an event, interpret speech or determine an action. Some local adaptation and learning are possible, but large-scale model training still usually depends on substantially greater computing resources.

Running AI locally has long been technically possible. Smartphones have used machine learning for photography, speech, biometrics and other functions for years; cars process camera and sensor information locally; industrial systems perform machine vision and anomaly detection; and very small neural networks can operate on microcontrollers and even within intelligent sensors.

What has changed is the amount and type of AI that devices can perform. Dedicated **neural processing units (NPUs)** and other AI accelerators have become common across phones, PCs and embedded systems. Software increasingly compresses and optimises neural networks through techniques such as quantisation, which represents model calculations with fewer bits, making them smaller and less demanding. Hardware and software improvements are consequently allowing language, image and multimodal models that once required substantial server infrastructure to run locally on increasingly ordinary devices. Microsoft, for example, defines its current Copilot+ PC category around devices containing NPUs able to perform more than 40 trillion operations per second, while current mobile processors from Samsung, MediaTek and others explicitly target generative-AI workloads on the handset.

Edge AI should nevertheless not be understood as the displacement of cloud AI. Increasingly, the two form a **hybrid architecture**. A device may perform routine, private or time-sensitive work locally and use remote computing for requests requiring larger models, more memory, current external information or substantially greater processing capacity. Apple Intelligence, for example, combines models running on devices with larger server-based models for workloads that exceed what the device is intended to handle. Automotive developers similarly describe architectures combining immediate on-device responses with richer cloud capabilities.

The current landscape is therefore best understood as a distribution of AI computation across a continuum: from tiny always-on intelligence inside sensors, through smartphones, PCs and machines, to local industrial computers and powerful edge systems, with the cloud remaining an important complementary layer.

2. What can it currently do?

Edge AI can already perform a large range of perception, language, prediction and control tasks without continuously transmitting raw data to a remote AI service.

At the smallest scale, microcontrollers and intelligent sensors can perform activities such as **wake-word detection, vibration and acoustic analysis, anomaly detection, gesture recognition, simple classification, occupancy sensing and basic computer vision**. These are

normally specialised models rather than general-purpose AI systems. Their significance lies in their ability to operate continuously using very little power and to react immediately to sensor readings.

Computer vision is among the most mature edge applications. Cameras and local processors can recognise objects, inspect manufactured components, analyse production activity, monitor equipment and extract information from video. Some intelligent image sensors can perform AI processing inside the sensor itself and output selected metadata rather than transmitting the original images. Sony, for example, integrates AI processing and model memory within its intelligent vision sensors, explicitly reducing the amount of image data that must be transferred elsewhere.

Smartphones can perform considerably more sophisticated inference. Current high-end devices support combinations of speech recognition, translation, text generation and rewriting, notification summarisation, image enhancement and generation, image understanding and personalised assistance. Google describes Gemini Nano as operating directly on phones for functions including notification summarisation and text proofreading, while Samsung's current mobile AI hardware supports local text, image, audio and video models.

Personal computers increasingly form another important edge-AI platform. Dedicated NPUs can run persistent or frequent AI workloads without placing all computation on the main CPU or GPU. Available applications include background effects in video calls, speech transcription and translation, image processing, local search and retrieval, text assistance and increasingly local generative models. Microsoft's commercially available Copilot+ PC ecosystem spans Qualcomm, AMD and Intel processors and provides a concrete example of AI acceleration becoming an ordinary characteristic of consumer and business computers rather than a specialist research configuration.

Vehicles constitute a different and particularly important category because decisions may need to be made faster than a cloud connection can reliably support. Local systems already process camera, radar and other sensor information for driver assistance, perception, driver monitoring, navigation and cockpit functions. Newer automotive processors are also being designed to support local generative and conversational AI. Qualcomm reported in January 2026 that its automotive platforms incorporating edge AI were present across more than 75 million vehicles, although that company figure covers a broad range of vehicle intelligence rather than establishing that all such vehicles contain the newest generative capabilities.

Industrial equipment, robots and autonomous machines similarly depend on local inference when sensing and action must occur with very low latency. NVIDIA's Jetson Thor platform, commercially available since August 2025, illustrates the upper end of on-device computing: an embedded system capable of running multiple vision, language and robotics models while processing sensor streams locally. The increasing importance of local processing in physical systems follows directly from the need to respond rapidly without assuming continuous cloud connectivity. Physical-AI systems frequently need to respond in fractions of a second, making faster local processing particularly important.

The frontier is now moving from narrow inference towards **generative and multimodal AI at the edge**. Current processors can run language models, vision-language models and some image-generation systems locally, and developers are demonstrating increasingly sophisticated assistants and robotic reasoning systems. MediaTek's Dimensity 9500, for example, supports

local large-model processing and image generation, while NVIDIA's embedded platforms can run language, vision and robotics foundation models.

These capabilities remain materially below unconstrained data-centre AI, however. Local models usually face stricter limits on memory, processing, energy and thermal load. A demonstration that a model can execute on a phone or embedded processor therefore does not establish that it matches the quality, context capacity or speed of a much larger cloud system.

3. Where is it currently being applied?

The broadest deployment is in **consumer devices**. Smartphones already contain substantial AI processing for cameras, speech, biometrics, predictive interfaces and communications. Generative AI has broadened this to summarisation, writing, translation, image creation and assistant functions. Apple, Google, Samsung, OPPO, MediaTek and other participants across the global mobile ecosystem are increasingly designing hardware, operating systems and applications around combinations of local and cloud inference. OPPO and MediaTek, for example, were demonstrating and rolling out additional local mobile-AI capabilities during 2026, reflecting development beyond the US technology ecosystem.

Personal computers are undergoing a similar transition. AI accelerators are now incorporated into devices from multiple processor architectures and manufacturers. The applications remain uneven—some features perform locally while others require cloud systems—but hardware capable of meaningful local AI inference has entered mainstream commercial product ranges rather than remaining confined to specialist workstations.

Manufacturing and industrial operations use edge AI for visual quality inspection, process monitoring, predictive maintenance, safety monitoring and analysis of machinery and worker activity. Local processing can reduce the need to stream large volumes of factory video or sensor data externally and can continue operating where connectivity is unreliable. One reported deployment using NVIDIA Jetson systems processes manufacturing camera feeds at the edge and extracts operational metadata rather than continuously treating the cloud as the primary inference location. This is supplier-associated case evidence rather than a representative measure of industrial adoption, but it illustrates the architecture in actual use.

Automotive systems constitute another large installed domain. Much perception and control in advanced driver-assistance systems must occur within the vehicle. Digital cockpits additionally employ local AI for voice, driver recognition, personalisation and multimedia functions. Current development programmes from global vehicle manufacturers increasingly combine these established functions with more capable conversational and generative models.

Robotics and autonomous machinery use edge processors to integrate cameras and other sensors with navigation, manipulation and control. NVIDIA reported more than 7,000 customers using its Jetson edge ecosystem by August 2025 and subsequently expanded the platform in 2026. These are developer figures and should not be read as measures of final-production deployment, but they demonstrate a substantial development ecosystem spanning manufacturing, healthcare, logistics, agriculture and robotics.

At the smaller-device end, **embedded and Internet-of-Things systems** are incorporating AI into microcontrollers, appliances, cameras, wearables and sensors. STMicroelectronics' STM32N6 family, for example, is in volume production with an integrated neural accelerator intended for real-time vision and audio inference, showing that dedicated AI hardware is reaching microcontroller-class systems rather than being confined to phones or computers.

The geographical pattern is consequently distributed. US companies are particularly prominent in software platforms and high-performance AI accelerators; Taiwan and South Korea are major mobile semiconductor centres; Japan has important sensor and electronics capabilities; European semiconductor companies are prominent in embedded and industrial edge hardware; China has a large smartphone, automotive and device-manufacturing ecosystem. Because much of the underlying semiconductor intellectual property and software is used across international supply chains, the edge landscape cannot readily be divided into isolated national systems.

4. How developed is it?

Edge & On-Device AI contains several maturity levels simultaneously.

Conventional specialised edge inference is mature. Image processing, speech functions, sensor classification, biometric recognition, vehicle perception and other bounded models have operated locally at commercial scale for years. Their presence is so routine that many users do not identify them as “edge AI”.

AI-accelerated consumer hardware is now established and scaling. Dedicated neural processors are present across major smartphone platforms and increasingly across personal computers. Commercial availability is extensive. The more important uncertainty is not whether appropriate hardware exists but how extensively software applications make effective use of it.

Industrial and embedded AI is established but highly heterogeneous. Machine vision and specialised sensor inference are mature in suitable applications; newer accelerator-equipped microcontrollers are extending what can be done under much tighter power and memory limits. Arm, STMicroelectronics and other suppliers now provide hardware and software specifically intended to deploy neural networks across devices ranging from microcontrollers to more capable embedded processors.

On-device generative AI is commercially real but less mature. Language models now operate on mainstream phones and PCs, while more powerful edge hardware can execute multimodal and robotics models. Commercial processors launched during 2025–26 were explicitly designed for transformer and generative-AI workloads. This moves local generative AI beyond research demonstration. It does not mean that frontier cloud capability has been reproduced on ordinary endpoints.

At the high-performance end, embedded systems are becoming remarkably capable. NVIDIA's commercially available Jetson Thor provides up to 128 GB of memory and substantially more AI compute than its predecessor, allowing larger generative and robotics models to run close to machines and sensors. Its August 2026 launch of a lower-cost Orin Nano 2 further illustrates

expansion across performance tiers. These specifications establish technical availability, not the extent to which the newest generation has achieved scaled operational deployment.

The Status as at 30 August 2026 is therefore not that AI has suddenly “moved from the cloud to the edge”. Specialised edge AI is already deeply established; dedicated AI hardware is becoming pervasive across device categories; and significantly more capable generative and multimodal inference is now moving onto devices. The newest forms remain constrained and frequently operate as one component of hybrid cloud-and-edge systems.

5. What is driving its development and use?

The first driver is **latency**. Sending data to a distant server, waiting for processing and returning the answer introduces delay and depends on network conditions. For interactive speech, industrial control, robotics, vehicles and other real-time systems, local inference can provide much faster and more predictable responses.

The second is **connectivity independence**. Devices in factories, vehicles, farms, mines, homes and remote locations cannot always assume continuous high-quality network access. Local processing allows at least some AI capability to continue during weak connectivity or complete disconnection.

Third is **privacy and data localisation**. Keeping a recording, image, document or sensor stream on the originating device can reduce how much sensitive raw data is transmitted elsewhere. This is particularly attractive for personal computing, cameras, industrial systems and healthcare-related equipment. Apple's architecture explicitly uses local models where possible, while Sony's intelligent image sensor can output analysed metadata instead of the underlying image.

Fourth is **bandwidth and cloud-computing cost**. High-resolution cameras and large numbers of sensors can generate enormous data streams. Analysing data locally and sending only relevant events or metadata can reduce network traffic and remote processing requirements. This becomes increasingly important as organisations deploy many connected devices.

Fifth is rapid improvement in **specialised hardware**. Dedicated NPUs allow many neural-network operations to run more efficiently than if the same workloads relied entirely on general-purpose processors. At one extreme, modern PC NPUs exceed 40 trillion operations per second; at another, microcontrollers now incorporate neural accelerators capable of real-time vision and audio inference.

Sixth is **model optimisation**. Quantisation, compression, distillation, memory-management techniques and hardware-specific software toolchains allow increasingly capable models to operate within limited memory, power and storage. Google's 2026 work to accelerate Gemini Nano on Pixel devices illustrates continuing optimisation of production on-device language models, while embedded toolchains from Arm and ST support conversion and quantisation for much smaller processors.

Finally, demand is increasingly self-reinforcing. Once large installed bases of phones, PCs, vehicles and embedded systems contain AI accelerators, developers have stronger incentives

to build software that uses them. Major frameworks—including Android's device AI infrastructure, ONNX Runtime and PyTorch ExecuTorch—are increasingly supporting local deployment across different hardware platforms.

6. What are the present limitations?

The fundamental limitation is **resource constraint**. Edge devices have finite memory, storage, processing capacity and power. A data centre can distribute a model across large clusters of accelerators; a phone, sensor or robot must work within a far smaller physical and electrical envelope.

Energy and heat are therefore central constraints. Running substantial models continuously can drain batteries or generate heat that forces processors to reduce performance. Research on energy-efficient edge AI continues precisely because model inference, training and data movement can place significant demands on resource-constrained devices.

Model capability can also be lower. Compressing or quantising a model may make it small and fast enough to run locally but can involve trade-offs in accuracy or flexibility. Device models may have less memory for long context, fewer parameters or fewer tools than cloud systems. Some sophisticated requests therefore still need to be handed to larger remote models.

Hardware specifications are themselves easy to misinterpret. Measures such as **TOPS—trillions of operations per second—do not provide a simple cross-platform measure of useful AI intelligence**. Results depend on numerical precision, memory bandwidth, model architecture, software optimisation, utilisation and the workload being performed. Even hardware suppliers caution that system optimisation affects actual performance.

Fragmentation creates another difficulty. Phones, PCs, vehicles and embedded devices contain different processors, operating systems, AI accelerators and software frameworks. A model that executes efficiently on one NPU may require conversion or optimisation for another. Toolchains are improving, but deploying and maintaining models across heterogeneous fleets remains materially harder than hosting one service centrally.

Updating models and information is also more complicated. Cloud providers can alter a central model once. Locally deployed models may need secure distribution to many devices operating different software versions or connectivity conditions. Large model files also consume storage and network bandwidth when updates are delivered.

Privacy advantages require qualification. **Local processing is not synonymous with privacy.** A device can still collect or transmit data; applications can have broad permissions; telemetry may leave the endpoint; and locally stored information can be exposed if the device or software is compromised. Whether an edge system protects privacy depends on the complete architecture and governance, not solely on the physical location of inference.

Local AI also creates distinctive **security risks**. Models and intellectual property placed on devices can be examined or extracted by attackers. Compromised firmware or model updates can affect distributed fleets. Adversarial inputs can be especially consequential where inference

controls vehicles, machinery or security systems. Physical access to an endpoint also creates attack possibilities not present in a protected data centre.

Finally, local availability does not establish operational effectiveness. A processor may technically execute a language or vision model while providing inadequate speed, battery life, accuracy or reliability for a particular application. Running a model on a device and running it well enough to create a dependable product are different achievements.

7. What you should be aware of

The most important qualification is that **edge AI is not an alternative universe separate from cloud AI**. The emerging architecture is often hybrid.

A smartphone may perform speech processing, text summarisation or image analysis locally but send a more difficult request to a larger server model. A vehicle may make safety-critical perception decisions locally while receiving maps, updates or richer services through network connections. An industrial camera may detect events at the edge while sending selected results to central systems for wider analysis.

Consequently, “runs on device” requires careful interpretation. It may mean that one particular model or feature operates locally, not that the entire AI service is independent of cloud infrastructure. Apple explicitly combines on-device foundation models with Private Cloud Compute for more complex workloads, while automotive systems increasingly describe complementary edge-and-cloud architectures.

The privacy argument should be interpreted similarly. Keeping raw data local can materially reduce exposure and can be especially valuable for cameras, personal information or proprietary industrial data. But architecture, permissions, encryption, software updates, telemetry and organisational practices determine whether that potential privacy advantage is realised.

Nor does local inference necessarily mean **lower total energy consumption**. Avoiding communications and remote computation may save energy for some workloads, but performing intensive inference repeatedly on millions or billions of devices also consumes electricity and can shorten battery life. The appropriate comparison depends on the model, hardware, workload frequency, network and data-centre alternative.

Geography also matters. Edge capability depends partly on access to modern semiconductor manufacturing, processor designs, embedded software ecosystems and device supply chains. Its development is unusually international: mobile chips, processor intellectual property, sensors, device manufacturing, operating systems and AI software frequently originate in different countries. The spread of edge AI therefore reflects an interconnected global technology system rather than the activity of any single national market.

Most importantly, the current shift should not be reduced to the claim that ever-larger AI models are being squeezed onto smaller machines. The more significant development is that **AI computation is becoming distributed across the technology environment**. Some

intelligence operates in data centres, some near networks and factories, some inside vehicles and robots, and some within phones, PCs, cameras and individual sensors.

As at 30 August 2026, specialised edge inference is mature and widespread; dedicated AI acceleration is becoming a normal component of major device categories; and generative and multimodal models are increasingly able to operate locally. Yet computing, memory, energy and software constraints remain substantial, and the most capable AI continues to depend heavily on larger infrastructure.

The present Status is therefore not the end of cloud dependence. It is the emergence of a more distributed AI architecture in which devices themselves are capable of performing a growing share of perception, interpretation, generation and decision-making close to the people, machines and environments producing the data.

Evidence Notes

1. **Apple — Apple Intelligence materials, June 2024–June 2026.** Primary evidence for hybrid on-device/server architecture, local foundation models and the use of Private Cloud Compute for more demanding requests. Company evidence establishes architecture and product availability; privacy properties are principally Apple claims, although the architecture permits external inspection of server software.
2. **Google Research — Accelerating Gemini Nano models on Pixel with frozen Multi-Token Prediction, 26 June 2026.** Technical evidence that production language models operate locally on smartphones and that inference efficiency remains an active engineering constraint.
3. **Microsoft — Copilot+ PC developer guidance and current Windows product information.** Evidence of commercial PC platforms containing NPUs above the 40-TOPS hardware threshold and of locally accelerated AI functions across Qualcomm, AMD and Intel systems. TOPS is a hardware throughput measure, not a direct measure of model intelligence or user effectiveness.
4. **MediaTek — Dimensity 9500, 22 September 2025.** Taiwan-based semiconductor evidence concerning commercially available smartphone hardware designed for local language, image-generation and transformer workloads. Performance figures are supplier benchmarks and should be treated accordingly.
5. **Samsung Semiconductor — Exynos on-device AI materials and Exynos 2500.** South Korean evidence concerning integrated mobile NPUs, local generative-AI processing and software support for frameworks including Android AICore, Gemini Nano and ExecuTorch. Hardware-performance figures are supplier evidence.
6. **OPPO / MediaTek — MWC 2026 on-device AI demonstrations, 4 March 2026.** Evidence of continuing on-device AI development within the Chinese and Taiwanese smartphone ecosystem. Demonstrations and announced roll-outs do not by themselves establish adoption across the installed smartphone base.
7. **Arm — Edge AI developer and processor materials, including Lumex and Ethos.** Evidence concerning AI inference from microcontroller-class devices through mobile processors, the role of NPUs and the widening software ecosystem for local model deployment. Vendor technical evidence rather than independent deployment measurement.
8. **STMicroelectronics — STM32N6 and ST Edge AI Suite.** European semiconductor evidence that dedicated neural acceleration and model optimisation have reached volume-production microcontrollers, supporting computer-vision, audio and other embedded inference without conventional server hardware.

9. **Sony Semiconductor Solutions — Intelligent Vision Sensor.** Japanese evidence of AI processing integrated within image sensors, illustrating the extreme edge of the landscape and the ability to transmit derived metadata instead of raw imagery.
 10. **NVIDIA — Jetson Thor, 25 August 2025; Jetson Orin Nano 2, 25 August 2026; Jetson ecosystem materials.** Evidence concerning high-performance embedded AI for robotics, vision and other real-time systems. Product specifications establish available local computing capability; developer, customer and adoption figures are company-reported rather than independent measures of deployed production systems.
 11. **Qualcomm — automotive platform materials, January–March 2026.** Evidence of the scale of edge-AI-capable automotive computing and of movement towards combined local generative AI and cloud services. The reported 75-million-plus vehicle figure is company evidence and covers a broad installed automotive platform base rather than equivalent use of the newest AI features in every vehicle.
 12. **Proceedings of the IEEE — Green Edge AI: A Contemporary Survey, 2025.** Research evidence concerning the tension between resource-intensive neural-network computation and the limited energy budgets of edge devices, and the continuing importance of energy-efficient edge inference and training.
-

12. Specialised, Small & Domain AI

Status as at 30 August 2026

Status snapshot

Specialised, Small & Domain AI represents the substantial part of the applied AI landscape in which capability is deliberately concentrated around particular tasks, industries, languages, environments or resource constraints rather than maximised for general-purpose breadth. It includes long-established task-specific AI, increasingly capable small foundation models and newer domain-adapted generative systems.

Maturity varies considerably. Traditional specialised systems for activities such as classification, forecasting, recommendation, anomaly detection and computer vision are established production technologies. Small foundation models are now an established and rapidly improving technical category, capable of language, reasoning, coding and increasingly multimodal tasks with substantially lower resource requirements than frontier-scale models. Domain-adapted generative AI is less uniformly mature: strong technical results exist in areas including medicine, cybersecurity, engineering and lower-resource languages, but operational deployment and demonstrated effectiveness remain uneven.

The principal current direction is a widening separation between model size and useful capability. Better training data, fine-tuning, knowledge distillation, quantisation and other efficiency techniques are enabling smaller systems to become highly capable within defined boundaries, while supporting advantages in cost, latency, control and deployment flexibility.

The central qualification is that specialised performance does not establish general capability, and benchmark success does not establish dependable real-world effectiveness. As at 30 August 2026, specialised and small AI is therefore not simply a lesser alternative to large general-purpose models, but an established and increasingly important complementary part of the applied AI landscape.

1. What is it?

Specialised, Small & Domain AI describes artificial-intelligence systems whose usefulness comes substantially from being designed, trained, adapted or optimised for a particular task, field, language, operating environment or resource constraint rather than principally from offering the broadest possible general-purpose capability.

The category contains several related but distinct ideas.

Small AI models use materially fewer computational resources than the largest general-purpose models. There is no universally accepted parameter threshold separating a “small language model” from a large one. Depending on context, the term can encompass models ranging from hundreds of millions to several billion parameters and sometimes considerably more. The important practical distinction is usually not a particular parameter count but

whether the model can perform useful work with substantially lower memory, computing, energy or inference requirements.

Specialised AI is optimised for a bounded purpose. This can include language models adapted for medicine or cybersecurity, vision models trained to identify particular industrial defects, speech systems designed for particular languages, models that classify documents or extract structured information, and longstanding predictive systems developed for specific activities such as fraud detection, forecasting or equipment monitoring.

Domain AI incorporates the terminology, data, conventions or operating requirements of a particular field, organisation, language community or environment. A domain model may be created by training specifically on domain data, continuing the training of a general model, fine-tuning it, attaching specialised components or combining it with external domain knowledge. Research literature increasingly treats domain specialisation as a family of approaches rather than a single model architecture.

These categories overlap, but they are not interchangeable. A small model can remain general-purpose. A domain model can be large. A specialised system can run in a cloud data centre rather than locally. Conversely, a model running on a phone is not necessarily domain-specialised. Deployment location therefore belongs principally in **Edge & On-Device AI**, while this element concerns the scale and specialisation of the intelligence itself.

The distinction has become more important because the rapid development of large foundation models has not eliminated the case for narrower systems. Instead, advances in training, data curation, knowledge distillation and model compression are allowing considerably smaller models to perform tasks that previously required much larger systems. Stanford's 2025 AI Index noted that the smallest model exceeding 60% on the MMLU knowledge benchmark fell from 540 billion parameters in 2022 to 3.8 billion in 2024—a 142-fold reduction. It also recorded a more than 280-fold fall between November 2022 and October 2024 in the inference cost of obtaining performance comparable with GPT-3.5 on that benchmark.

Model families illustrate the resulting range. Microsoft's Phi-4-mini has 3.8 billion parameters; Google's Gemma family includes variants from hundreds of millions of parameters upwards; Alibaba's Qwen3 family included dense models from 0.6 billion to 32 billion parameters; and Mistral's Ministral 3 family includes 3, 8 and 14 billion-parameter models. These examples are useful not as a model ranking but as evidence that substantial current AI development is occurring below frontier-model scale.

Specialisation is simultaneously becoming more geographically diverse. AI Singapore's SEA-LION programme adapts models to Southeast Asian languages and cultural contexts, while African initiatives are developing compact models for language identification and translation across languages poorly represented in mainstream training corpora.

Specialised, Small & Domain AI in August 2026 is therefore best understood not as a lesser alternative to general-purpose AI, but as a parallel part of the technology landscape: AI capability deliberately concentrated where breadth, scale or maximum generality is unnecessary or undesirable.

2. What can it currently do?

Current specialised and smaller AI systems can perform a wide range of bounded tasks effectively, and in some circumstances can match or exceed much larger general-purpose models on the activity for which they have been adapted.

For language tasks, small models can summarise, classify and rewrite text, answer questions, extract information, generate structured outputs, call software functions and perform selected reasoning and coding tasks. Microsoft's 3.8-billion-parameter Phi-4-mini, for example, was developed for reasoning-intensive language applications and supports a 128,000-token context window and function calling. Meta's earlier 1- and 3-billion-parameter Llama 3.2 models demonstrated summarisation, instruction-following and rewriting capabilities in models small enough for selected mobile and edge environments.

Small models are also increasingly multimodal. Gemma 3 variants combine text and image understanding at several model sizes, while Mistral's 3-, 8- and 14-billion-parameter Ministral 3 models include image understanding as well as multilingual and reasoning variants.

The more important capability, however, is not simply performing general tasks with fewer parameters. It is **concentrating capability**.

A model trained or adapted to a narrow domain can devote proportionately more of its effective capacity to the vocabulary, patterns and decisions that matter within that domain. Research in construction, for example, found that a small domain-adapted language model outperformed larger counterparts on a construction-specific question-answering dataset while providing faster inference. An industrial study involving extraction of technical requirements similarly found that a fine-tuned small language model could perform competitively on its defined task, although results from individual datasets should not be generalised to unrelated industrial settings.

Cybersecurity provides another example. A 2026 study fine-tuning models below four billion parameters found that some achieved parity with or better results than larger comparison models on the cybersecurity datasets tested. The evidence demonstrates the potential value of task-specific adaptation; it does not establish that small models generally outperform larger systems across cybersecurity.

Medicine is attracting particularly substantial interest. Google's MedGemma family adapts Gemma models for medical text and image comprehension, with variants intended as foundations for downstream healthcare applications. The 4-billion-parameter version supports medical text and several forms of medical imagery. Google explicitly states, however, that the models require application-specific validation and are not intended to be used directly for clinical diagnosis or patient-management decisions.

This distinction between technical capability and dependable application is important. Independent evaluation of medical models continues to identify shortcomings that ordinary benchmark scores may obscure. A 2026 evaluation of MedGemma 1.5 4B, for example, reported substantial weaknesses in recognising when additional clinical context was required. That study was a preprint rather than completed peer-reviewed evidence, but it illustrates why

strong performance on medical question answering should not be equated with clinical readiness.

Language and cultural specialisation show another form of capability. AI Singapore's SEA-LION family includes models adapted for Southeast Asian languages including Burmese, Indonesian, Malay, Tamil, Thai and Vietnamese. Its model family ranges across specialised text, vision-language, embedding and safety systems, including relatively compact variants.

Research involving low-resource languages provides evidence that scale alone does not determine performance. A 2026 evaluation across 200 languages found substantial weaknesses in unadapted small models for low-resource translation, but also found that targeted knowledge distillation could allow 2–3-billion-parameter models to match or surpass much larger systems on the tasks studied. African-language research similarly reported that targeted data and fine-tuning enabled modest-scale models to achieve substantial improvements, with a one-billion-parameter model matching or surpassing a commercial translation system in several evaluated languages.

The present capability picture is consequently conditional. Smaller systems are not miniature copies of frontier models with identical abilities. Their strength is increasingly that high-quality training, compression and domain adaptation can make relatively modest models highly competent within selected boundaries.

3. Where is it currently being applied?

Specialised AI is already deeply embedded across the economy, although much of it does not resemble the conversational generative AI that dominates public attention.

Established specialised machine-learning systems classify transactions, detect anomalies, recommend products, forecast demand, inspect manufactured goods, recognise speech, identify objects and support industrial control. Many such systems pre-date the current foundation-model era and remain appropriate precisely because their tasks are narrow.

The newer development is the extension of this principle into **generative and foundation-model technology**.

In enterprises, smaller and adapted language models are being explored or deployed for document classification, information extraction, customer support, internal knowledge tasks, software functions and structured workflows. Their appeal is strongest where the required activity is sufficiently defined that broad general knowledge offers limited additional value.

Healthcare is an important development area. Domain models are being investigated for medical documentation, medical-image interpretation, information extraction, clinical research and question answering. Peer-reviewed commentary published in *Nature Biomedical Engineering* in July 2026 described small language models as potentially more scalable, practical and efficient than large models for some medical applications, while emphasising continuing implementation challenges and limitations.

Law and finance similarly contain highly specialised terminology, document structures and regulatory contexts. Research has explored adaptation of relatively small models to legal subdomains: one 2025 study, for example, fine-tuned an eight-billion-parameter model across ten areas of Turkish law. Such studies establish technical feasibility and growing research activity rather than widespread replacement of professional legal judgement.

Manufacturing and engineering applications include extraction of technical requirements, analysis of engineering documentation, knowledge-graph construction and specialised question answering. Some research has progressed into industrial collaboration or prototype deployment, but evidence of broad adoption remains much thinner than evidence of technical feasibility.

Language-specific AI is particularly significant outside the most heavily represented global languages. AI Singapore's SEA-LION programme is designed around Southeast Asian languages and contexts. In Africa, the GSMA and Pleias released CommonLingua in April 2026, a compact language-identification model supporting 334 languages including 61 African languages. AfricaNLP research is also producing lightweight translation models for African languages. These developments illustrate a form of specialisation whose objective is not principally industrial expertise but linguistic inclusion.

Small models also appear in privacy-sensitive and resource-constrained environments. They can be operated on organisational infrastructure, consumer hardware or devices that cannot economically host frontier-scale models. That can reduce the amount of information sent to external services and provide more control over data handling. Whether processing actually occurs locally, however, depends on the deployment architecture and should not be inferred merely from model size.

The application landscape is therefore broad but uneven. Traditional specialised AI is already pervasive. Domain-adapted generative AI has moved substantially beyond laboratory research but varies from production systems to prototypes and experiments. Small foundation models are widely available and technically deployable, while reliable evidence of adoption and impact remains much less comprehensive than evidence of model availability.

4. How developed is it?

Specialised, Small & Domain AI cannot sensibly be assigned one maturity level.

Traditional task-specific AI is highly mature. Classification, forecasting, recommendation, computer vision, speech recognition, anomaly detection and optimisation systems have been deployed commercially for many years. In many bounded applications, specialised models are ordinary production technology.

Small foundation models are also established as a technical category. Major model ecosystems now routinely include smaller variants. Google, Microsoft, Meta, Alibaba, Mistral and IBM have all released models designed around different combinations of compactness, efficiency and adaptability. This is no longer an experimental question of whether useful language models can exist below frontier scale.

Their capabilities have improved materially. The Stanford AI Index's evidence of a 142-fold reduction between 2022 and 2024 in the model size required to cross one established benchmark threshold captures an important structural change: a given level of benchmark capability increasingly requires less model scale.

Research is also showing that reasoning ability is not determined by parameter count alone. An EMNLP 2025 study covering 72 small language models found that training method and data quality materially affected reasoning performance; some smaller models approached or exceeded larger comparison models on selected benchmarks, although larger models generally retained advantages in robustness and difficult intermediate reasoning.

Domain-adapted generative AI is less uniformly mature. The technical methods—fine-tuning, continued pre-training, retrieval augmentation, adapters, synthetic-data generation and knowledge injection—are established and extensively researched. What remains uneven is evidence that individual domain systems remain dependable when confronted with the full variability of real professional environments.

Healthcare illustrates this difference particularly clearly. Domain models exist and can demonstrate strong medical benchmark performance, yet their developers and independent researchers continue to emphasise validation, context sensitivity and clinical oversight.

Regional and low-resource-language models occupy another mixed maturity category. Working model families now exist, and targeted adaptation has produced strong benchmark improvements. Yet language coverage, datasets and evaluation resources remain extremely uneven. The existence of dedicated 2026 research workshops for low-resource language models and translation demonstrates both the progress of the field and the amount of work still required.

The present Status is consequently one of mature specialised AI at one end, established and rapidly improving small foundation models in the middle, and a diverse frontier of domain-, language- and organisation-specific generative systems whose operational maturity depends heavily on the application.

5. What is driving its development and use?

The first driver is **efficiency**.

Frontier AI models can be expensive to train and operate. Many applications do not require their full capability. Smaller models reduce memory, computing and often energy requirements, and can support higher-volume or lower-latency processing. Falling inference costs across AI reflect both hardware improvements and increasingly efficient models.

The second driver is **improved training efficiency**. Knowledge distillation allows a smaller model to learn from the outputs or reasoning of a stronger model. Quantisation represents model weights using lower numerical precision, reducing memory and computational requirements. Pruning removes less important model components. Research reviews identify these as central techniques for making capable models usable under tighter resource constraints.

The third is **better data rather than simply more parameters**. Carefully curated domain data, synthetic training material and targeted post-training can concentrate model capability. Microsoft's Phi research, for example, has deliberately emphasised high-quality and synthetic training data as a way of producing strong capability in compact models.

The fourth is **domain performance**. General-purpose models must distribute capability across an enormous range of subjects. Organisations dealing with specialised vocabulary, formats and procedures may instead benefit from concentrating capability around a narrower task. Research on domain specialisation increasingly documents this trade-off across medicine, law, finance, science, engineering and other fields.

The fifth is **privacy and control**. Smaller models can make private-cloud, on-premises and local processing more technically practical. This does not automatically make a system private or secure, but it expands the range of deployment architectures available to organisations that do not wish to transmit all sensitive information to externally operated foundation models.

The sixth is **latency and connectivity**. Some applications require rapid responses or operation without dependable network connectivity. Smaller models can be easier to place close to the application, although the detailed significance of local deployment belongs to Edge & On-Device AI.

The seventh is **linguistic and cultural coverage**. Global foundation models remain uneven in their treatment of languages and local contexts. Southeast Asian and African initiatives demonstrate that targeted datasets and model adaptation can address needs that global model development may not prioritise.

Finally, open-weight model ecosystems have lowered the technical barrier to specialisation. Developers can begin with an existing capable model and adapt it rather than training a foundation model from the beginning. This has widened participation beyond the small number of organisations able to finance frontier-scale training.

6. What are the present limitations?

The most important limitation is that **specialisation creates boundaries**.

A model that performs extremely well inside its intended domain may deteriorate sharply outside it. This is not necessarily a defect—the narrower competence may be exactly why the system is efficient—but users must understand where the boundary lies.

Small models also generally retain less broad knowledge and capability than leading frontier models. Although selected compact systems perform impressively on individual benchmarks, larger models continue to show advantages on many difficult, unfamiliar and multi-stage tasks. The ThinkSLM evaluation, for example, found stronger robustness in larger models despite competitive results from some smaller ones.

Data quality is critical. Domain models require representative, accurate and sufficiently comprehensive specialist data. Such data can be scarce, proprietary, sensitive or expensive to label. Medical, legal, engineering and low-resource-language domains all illustrate this

constraint. Poor or incomplete specialist data can produce a model that appears expert because it uses the correct vocabulary while still making substantive errors.

Specialisation can also **freeze assumptions into the model**. Laws, regulations, products, medical knowledge and organisational procedures change. A domain model trained on a static dataset may become outdated unless knowledge and system components are maintained.

Evaluation remains difficult. General benchmarks may not measure the precise behaviour required in an operational domain, while bespoke domain benchmarks can be narrow or susceptible to contamination. A model that performs well on examination-style medical questions, for example, has not thereby demonstrated safe clinical decision-making. Similar distinctions apply to legal reasoning, financial analysis and engineering work.

Compression introduces trade-offs. Quantisation, pruning and distillation can substantially reduce resource requirements, but preserving capability is not guaranteed. Research indicates that different techniques affect performance and robustness differently.

Privacy is conditional rather than inherent. A small model may be capable of local processing, but it can still leak information, be insecurely integrated, retain sensitive data or interact with external services. Model size alone is not a privacy control.

Domain expertise does not eliminate hallucination. Specialised generative models can still produce fluent but incorrect answers. In high-stakes domains, the apparent authority produced by specialised vocabulary may make such errors particularly difficult for non-experts to recognise.

Governance requirements also do not disappear merely because a model is small. Regulation increasingly distinguishes systems according to purpose, capability and risk rather than treating parameter count as the sole determinant. Under the European Union's AI Act, for example, obligations for general-purpose models and downstream AI systems depend on the nature and use of the model, with additional requirements for models meeting systemic-risk criteria.

Finally, the economics are application-specific. A smaller model may cost less per inference, but developing specialist datasets, evaluating the model, integrating it with operational systems and maintaining it can itself be expensive. A technically efficient model is not automatically the cheapest complete system.

7. What you should be aware of

The most important point is that **model size and useful capability are no longer reliably interchangeable**.

Larger models generally provide greater breadth and often stronger performance on difficult or unfamiliar tasks. But parameter count alone does not determine whether a model is the most capable system for a particular activity. Training data, architecture, post-training, domain adaptation, compression, retrieval and the surrounding application all matter.

This is why specialised AI has not disappeared as general-purpose models have improved. In many real applications the objective is not to answer every possible question. It is to perform one defined activity accurately, quickly, economically and consistently.

“Small” should also not be confused with “simple”. A model containing several billion parameters remains a substantial machine-learning system. Nor should “domain-specific” be interpreted as automatically trustworthy. A medical model can still make medical errors; a legal model can misunderstand law; a language-specific model can reproduce cultural biases; and an industrial model can fail when operating conditions differ from its training data.

Benchmark comparisons require particular care. A small model outperforming a much larger model on a specialised benchmark establishes something important: scale is not the only route to capability. It does not establish that the smaller model is generally more capable. The reverse is also true. A frontier model winning a broad benchmark does not prove that it is the most accurate, economical or appropriate system for every narrow operational task.

The distinction between **model capability** and **system capability** is equally important. A specialised model may operate alongside retrieval systems, databases, deterministic software, rules, sensors, larger models and human experts. In some architectures a small model handles routine cases while a more capable model or a person receives difficult exceptions. The useful unit of analysis is therefore often the complete system rather than the isolated model.

There is also a significant global dimension. Specialisation provides one route by which countries, research communities and language groups can develop AI capabilities without reproducing the enormous computing expenditure associated with frontier foundation models. Southeast Asian and African language initiatives show how targeted data can shift capability towards populations poorly represented in global training datasets. Yet these initiatives also expose the continuing inequality of available language data, computing infrastructure, evaluation resources and funding.

As at 30 August 2026, Specialised, Small & Domain AI is therefore not a transitional category waiting to be displaced by ever-larger general-purpose systems. It is an established and increasingly important part of the applied AI landscape.

The central development is a widening separation between **how large a model is** and **how useful it can be for a defined purpose**. Improvements in training, data quality, adaptation and compression are allowing substantial capability to be concentrated into smaller systems. At the same time, increasingly capable foundation models provide starting points from which domain-specific systems can be created.

The resulting landscape is not simply a contest between “small” and “large” AI. It is an increasingly varied ecosystem in which general-purpose models, specialised models and conventional task-specific systems occupy different roles—and where the appropriate measure of technological maturity is not parameter count, but demonstrated performance, reliability and effectiveness in the task and environment for which the system is actually used.

Evidence Notes

1. **Stanford Institute for Human-Centered Artificial Intelligence — AI Index Report 2025, April 2025.** Principal independent evidence concerning improving capability of smaller models and falling inference costs. Reports the 142-fold reduction between 2022 and 2024 in model size required to exceed 60% on MMLU and the more than 280-fold decline in cost for GPT-3.5-equivalent MMLU performance. Benchmark thresholds are indicators of technical progress rather than measures of general real-world effectiveness.
2. **Wang et al. — “A Comprehensive Survey of Small Language Models in the Era of Large Language Models”, ACM Transactions on Intelligent Systems and Technology, 24 November 2025.** Extensive peer-reviewed review of small-language-model architectures, training, compression, applications, LLM–SLM collaboration and trustworthiness. Particularly useful for distinguishing small-model development from simple parameter reduction.
3. **Srivastava, Cao & Wang — “ThinkSLM: Towards Reasoning in Small Language Models”, EMNLP 2025.** Evaluation of 72 small language models across reasoning benchmarks. Evidence that training methods and data quality materially affect reasoning capability and that some smaller models can compete with larger systems on selected tasks; also finds continuing robustness advantages for larger models. Benchmark evidence rather than operational-deployment evidence.
4. **Ling et al. — “Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey”, ACM Computing Surveys, 6 October 2025.** Peer-reviewed survey of domain specialisation across medicine, law, finance, science, engineering and other fields. Used principally to establish the breadth of domain-adaptation methods and applications.
5. **Song et al. — “Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey”, Findings of EMNLP 2025.** Research synthesis covering dynamic knowledge injection, static embedding, modular adapters and prompt-based methods. Evidence that domain specialisation is a family of techniques rather than a single architecture.
6. **Microsoft Research — Phi-4-Mini Technical Report, March 2025.** Primary technical evidence concerning the 3.8-billion-parameter Phi-4-mini and the role of curated and synthetic training data in compact-model performance. Developer evidence; performance claims should be interpreted in the context of reported evaluations rather than as independent evidence of deployment effectiveness.
7. **Google — Gemma 3, 12 March 2025, together with subsequent Gemma releases.** Primary evidence of lightweight general and task-adaptable models at multiple sizes, multimodal capability and model compression. Developer source, used principally as evidence of model capability and availability.
8. **Meta — Llama 3.2, 25 September 2024, and quantised Llama releases, 24 October 2024.** Primary evidence concerning 1B and 3B lightweight models and quantisation for reduced model size and memory requirements. Developer measurements of speed and memory improvements should not be interpreted as independent deployment studies.
9. **Alibaba — Qwen3, 29 April 2025.** Primary evidence of a globally available model family spanning dense models from 0.6B to 32B parameters and larger mixture-of-experts systems. Used to demonstrate the breadth of current model scaling strategies rather than comparative superiority.
10. **Mistral AI — Mistral 3 / Ministral 3, 2 December 2025.** Primary evidence of 3B, 8B and 14B multimodal, multilingual and reasoning-capable compact models. Performance and cost claims are developer evidence.
11. **IBM — Granite 3.0 and subsequent Granite family materials.** Primary evidence of enterprise-oriented models at 2B and 8B scale and smaller mixture-of-experts variants, with emphasis on adaptation and fit-for-purpose deployment. Developer evidence.
12. **Google Health AI Developer Foundations — MedGemma, from 20 May 2025; updated through 2026.** Primary evidence of domain adaptation for medical text and imaging. Google explicitly states that MedGemma is a development foundation requiring validation and is not intended for unmodified direct clinical decision-making.

13. **Qin et al. — “Small language models in medicine”, Nature Biomedical Engineering, 13 July 2026.** Peer-reviewed commentary assessing the potential, limitations and implementation considerations of small models in healthcare. Useful evidence that medical SLMs had become a substantive research and implementation field by the reference date.
 14. **Luginbuhl — “MedGemma on HealthBench”, March 2026.** Independent evaluation highlighting context-seeking limitations in a compact medical model. This was a preprint undergoing review at the reference date and is therefore treated as indicative rather than settled evidence.
 15. **Müller et al. — “Data-Centric Fine-Tuning of Small Language Models for Automatic Extraction of Technical Requirements”, IEEE Access, 22 July 2025.** Industry-linked evaluation of small-model adaptation for technical-document information extraction. Evidence relates to a defined task and dataset and should not be extrapolated to general industrial intelligence.
 16. **Wang, Fu & Kim — “Toward construction-specialized, small language models”, Advanced Engineering Informatics, 2025/26.** Domain-specific comparison demonstrating that a small adapted model could outperform larger models on the construction QA task tested. Task-specific research evidence rather than evidence of general superiority of smaller models.
 17. **TinySecGPT study, Computers, Materials & Continua, March 2026.** Evidence that sub-4B models fine-tuned for cybersecurity achieved competitive results against larger models on the evaluated cybersecurity datasets. The finding is domain- and benchmark-specific.
 18. **Song et al. — “Are Small Language Models the Silver Bullet to Low-Resource Languages Machine Translation?”, LoResMT 2026, March 2026.** Large-scale evaluation across 200 languages. Particularly useful because it shows both sides of the Status: unadapted SLMs underperformed systematically in low-resource languages, while targeted distillation substantially improved results.
 19. **African Languages Lab, ACL 2026.** Research evidence covering data collection and model development across African languages. Reports substantial improvements from targeted fine-tuning and instances where a 1B model matched or exceeded a commercial translation system on particular languages. Results concern the evaluated datasets and languages, not translation performance generally.
 20. **GSMA / Pleias — CommonLingua, 28 April 2026.** Evidence of a purpose-built two-million-parameter language-identification model covering 334 languages, including 61 African languages. Developer/initiative evidence used principally to establish the existence and scale of specialised African-language model development.
 21. **AI Singapore — SEA-LION model family, materials current to 26 May 2026.** Primary regional evidence of language- and culture-adapted models for Southeast Asia, including compact text, multimodal, embedding and safety models. Performance claims originate principally from the developer and associated benchmarks and are treated accordingly.
 22. **Girija et al. — “Optimizing LLMs for Resource-Constrained Environments: A Survey of Model Compression Techniques”, IEEE COMPSAC 2025.** Review of knowledge distillation, quantisation and pruning for reducing model resource requirements. Used principally for technical orientation rather than evidence of commercial deployment.
 23. **European Commission — Guidelines and information concerning general-purpose AI obligations under the EU AI Act, July–August 2025, updated through 2026.** Primary regulatory evidence showing that AI governance requirements depend on model characteristics, purpose and risk and do not reduce simply to a small-versus-large model distinction.
 24. **NIST — AI Risk Management Framework and Generative AI Profile, materials current through 2026.** Public-sector evidence supporting the broader distinction between model capability and the risk, testing and governance requirements of the complete AI system and its use context.
-

13. AI Data & Synthetic Data

Status as at 30 August 2026

Status snapshot

AI data is the information layer from which AI systems learn, encompassing large-scale pre-training collections, specialised and multimodal datasets, curated post-training material and increasingly machine-generated or synthetic data. As at 30 August 2026, large-scale data collection, filtering, deduplication and curation are mature foundations of AI development, while synthetic data occupies several maturity levels: conventional augmentation and simulation are established, synthetic post-training is increasingly used operationally, but synthetic replacement of real-world foundational data remains much less proven.

Current systems can use deliberately selected or generated data to improve particular capabilities, create training examples where real observations are scarce or expensive, extend multilingual coverage, support medical and other privacy-sensitive research, and expose physical-AI systems to simulated conditions. The strongest evidence is for targeted and hybrid uses rather than unrestricted substitution of authentic data.

The principal direction of development is from treating data mainly as a resource to be accumulated towards treating it as an engineered component of AI capability: selecting higher-quality material, constructing specialised training mixtures and generating examples around identified needs.

The central qualification is that more data—real or synthetic—does not necessarily mean better AI. Synthetic data can reproduce errors, biases and gaps in its sources, and availability does not establish downstream effectiveness. As at 30 August 2026, synthetic data is an increasingly important complement to high-quality real-world data, not a demonstrated universal replacement for it.

1. What is it?

AI Data & Synthetic Data is the information layer from which AI systems learn. It includes the material used to pre-train general-purpose models, the more selective data used to adapt or post-train them, specialist datasets for particular domains and languages, labelled examples, multimodal collections of text, images, audio and video, and increasingly data generated by algorithms or other AI systems.

The important distinction is that **data quantity and data quality are not the same thing**. Large foundation models have historically benefited from enormous collections of web pages, books, code, images and other material. But raw scale alone does not determine what a model learns. Duplication, errors, low-quality text, inappropriate content, poor labelling, demographic and linguistic imbalance, outdated information and contamination of evaluation datasets can all reduce the usefulness of a large collection. Consequently, filtering, deduplication, classification, balancing and deliberate selection have become increasingly important parts of AI development. Stanford's 2026 AI Index identifies data quality and curation as increasingly

significant as developers confront limits to simply expanding high-quality human-generated pre-training data.

The data used at different stages also serves different purposes. **Pre-training data** gives a model broad statistical knowledge of language, images, code or other modalities. **Post-training data** is usually much more selective and can teach a model how to follow instructions, solve particular kinds of problems, use tools, reason through structured tasks or produce safer and more useful responses. Specialist adaptation may add legal documents, clinical examples, scientific material, local-language content or organisation-specific information. In modern AI, a relatively small amount of carefully selected post-training data can therefore materially influence a model whose underlying pre-training involved vastly more information.

Synthetic data is data created artificially rather than directly recorded as the underlying real-world event or observation it represents. It may be generated statistically, procedurally, through simulation or by another AI model. Examples range from artificial patient records that reproduce selected statistical characteristics of real health records, to computer-generated road scenes, fabricated mathematical problems, synthetic images with known labels and language-model-generated instruction-and-answer pairs.

Synthetic data should not be understood simply as “fake data”. Its usefulness depends on what is being reproduced and why. A simulated driving scene may deliberately create a rare hazardous situation that would be difficult or unsafe to collect repeatedly in the real world. An AI model may generate thousands of mathematically verifiable problems for post-training. A synthetic medical dataset may reproduce useful population-level relationships while reducing direct exposure of individual patient records. In each case, the relevant question is whether the artificial data preserves or creates the characteristics needed for the intended AI task.

This element concerns that **training and adaptation layer**, rather than general organisational data management. It also overlaps with World Models & Simulation where simulated environments generate experience for physical systems, but the focus here is the resulting data and its role in learning rather than the simulated world itself.

2. What can it currently do?

The current data layer can support increasingly deliberate shaping of AI capability rather than merely feeding models larger collections.

At the broadest level, large curated datasets can teach foundation models language, visual concepts, code, factual associations and statistical patterns across many fields. Public data initiatives illustrate the scale now involved. FineWeb2, for example, processed 96 Common Crawl snapshots into a filtered and deduplicated multilingual collection of roughly 20 terabytes, five billion documents and more than three trillion words across very large numbers of language-script combinations. Its importance is not that every document is necessarily high quality, but that web-scale data engineering has become a distinct technological capability involving language identification, filtering, deduplication and evaluation.

Data curation can also improve capability without simply increasing dataset size. Filtering for educational or information-rich material, removing duplicates and changing the mixture of

domains can materially change model performance. Microsoft's Phi programme provides a prominent developer example: Phi-4 used a training approach centred on filtered organic material together with extensive deliberately generated synthetic data, rather than treating unfiltered web volume as the principal objective.

Synthetic data has become particularly important in **post-training and bounded tasks**. A stronger model can generate questions, worked examples, reasoning demonstrations or answers that are then filtered and used to train another model. Microsoft's Phi-4-reasoning work used synthetic reasoning responses produced by another model as supervised training material. DeepSeek similarly demonstrated the practical importance of model-output distillation when it released smaller models distilled from DeepSeek-R1. These examples establish that machine-generated supervision has become an operational part of contemporary model development, although developer benchmark claims do not by themselves establish equivalent performance across real-world tasks.

Research evidence supports benefits in appropriately constrained settings. ICLR 2026 research showed that rule-generated synthetic data containing fictional knowledge could improve multi-hop reasoning on external real-world question-answering benchmarks, while AAAI 2026 work generated more than 550,000 targeted multimodal examples around identified reasoning failures and reported improvements across downstream tasks. These results support the proposition that synthetic data can teach transferable structures or fill specific weaknesses when generation and verification are well designed. They do not establish that arbitrary AI-generated material has the same value.

Synthetic data is also capable of generating examples that are scarce, sensitive or expensive to collect. Medical research has demonstrated synthetic retinal images and other clinical data being used to supplement model development, with a 2025 Nature Biomedical Engineering study showing that disease-conditioned synthetic images could improve data efficiency and generalisation in an ophthalmic foundation-model setting. Other work has used synthetic clinical examples to transfer information from larger models into smaller models.

For physical AI, simulation and generative models can create labelled scenes, sensor observations and robot interactions at scales impractical to reproduce physically. NVIDIA's 2025 Cosmos tools, for example, were designed to generate variations in weather, lighting, driving conditions and robot manipulation data for autonomous-vehicle and robot post-training. This demonstrates an available industrial data-generation capability; the extent of downstream performance improvement remains dependent on the application and validation method.

What synthetic data cannot presently be said to do is replace real-world data generally. Stanford's 2026 AI Index concluded that the evidence still does not establish synthetic data as a substitute for real data in large-scale general-purpose pre-training. Hybrid and targeted approaches are more strongly supported than an assumption that indefinitely increasing machine-generated data will continue to improve frontier systems.

3. Where is it currently being applied?

AI data engineering is now embedded across almost every major form of AI development, but the particular data problem varies considerably.

For general-purpose language and multimodal models, the principal activity includes collecting, licensing, filtering, deduplicating and balancing enormous heterogeneous datasets, then assembling more selective post-training datasets. Increasingly, developers also generate training examples from existing models, particularly where answers can be checked automatically—for example mathematics, code and structured reasoning.

Multilingual AI provides a major global application. Online content is distributed extremely unevenly between languages, so a large web crawl does not produce linguistically balanced training data. FineWeb2 extended open pre-training resources across more than a thousand languages, while community initiatives have produced human quality judgements across over one hundred languages. Research continues to find that insufficient high-quality target-language data materially constrains models outside resource-rich languages.

Synthetic translation and instruction generation are consequently being applied to low-resource languages. Research involving Uyghur and Tibetan has used machine-translated corpora to expand available material, while other work has generated instruction-tuning examples from native-language texts. Such techniques enlarge the usable training base but do not remove the need for genuine local-language and culturally grounded material; translated or generated data can reproduce the weaknesses of the models that created it.

Healthcare and biomedical research are important because real patient data are sensitive, fragmented and expensive to collect and annotate. Synthetic electronic health records, medical images and other artificial patient data are being investigated for model development, data sharing, preliminary research and privacy-preserving analysis. Evidence includes operational research pipelines and demonstrated model improvements, but clinical validity remains application-specific rather than assumed from statistical similarity alone.

Autonomous vehicles, robotics and industrial vision use synthetic and simulated data to expose systems to combinations of objects, environments and rare events. Simulation allows developers to alter conditions deliberately and retain exact labels concerning what is present in the scene. It is particularly attractive for dangerous or uncommon situations that cannot economically be collected in sufficient quantity through ordinary operation.

Synthetic data also supports more conventional tabular machine learning where access to financial, customer, government or other sensitive records is constrained. Singapore's Personal Data Protection Commission, for example, has treated synthetic-data generation as a privacy-enhancing technology with potential uses in AI training and dataset augmentation, while emphasising assessment of re-identification risk rather than assuming generated records are automatically safe.

The application landscape is therefore broad, but the common function is consistent: data technologies are increasingly used to **improve coverage, create supervision, compensate for scarcity, expose models to targeted situations or make sensitive information more usable**

under controlled conditions.

4. How developed is it?

The data layer contains several distinct maturity levels.

Large-scale collection, cleaning, labelling, filtering, deduplication and dataset construction are **mature foundations of machine learning**. They predate the current generative-AI wave and are used operationally throughout AI development. What has changed is their scale and sophistication. Model-assisted filtering, automated classification, large multilingual pipelines and increasingly complex multimodal mixtures have made data engineering itself an important differentiator in model development.

Synthetic data for **conventional augmentation** is also established. Computer vision has long used transformations and artificially generated examples; simulation has long supported autonomous systems; statistically generated tabular records have an established research and commercial ecosystem.

Synthetic data for **modern generative-AI post-training** has moved rapidly into practical use. Developer disclosures and peer-reviewed research show machine-generated instructions, answers, reasoning examples and multimodal problems being used to train contemporary systems. This is no longer simply experimental methodology.

However, synthetic data as a **large-scale substitute for authentic pre-training data** is much less mature. Evidence remains mixed and highly dependent on how synthetic examples are generated, selected and combined with original material. Stanford's 2026 assessment found no definitive evidence that synthetic data can offset depletion of real data for general-purpose pre-training, although hybrid approaches can improve training efficiency and synthetic-only approaches can work well in narrower settings.

Privacy-preserving synthetic data similarly occupies an intermediate position. Techniques and practical guidance exist, and there are credible research demonstrations of useful privacy–utility trade-offs. But medical research reviews continue to find no universally accepted evaluation framework for determining whether a synthetic dataset is both sufficiently useful and sufficiently private.

The overall Status is therefore one of **mature data engineering, established augmentation and simulation, rapidly scaling synthetic post-training, and unresolved questions around synthetic replacement of real-world foundational data**.

5. What is driving its development and use?

One major driver is the enormous appetite of contemporary AI systems for useful training material. Successive generations of foundation models have consumed increasingly large datasets, raising questions about the remaining supply of high-quality human-created text and other accessible data. Estimates of when particular data pools may become limiting remain

uncertain, but the possibility of a data bottleneck creates strong incentives to improve curation, reuse existing information more efficiently and generate new training material synthetically.

A second driver is the **increasing value of quality relative to undifferentiated scale**. Model builders can remove duplicated pages, identify higher-quality educational material, adjust mixtures between domains and languages, and construct examples specifically around weaknesses revealed during evaluation. Data is consequently becoming more intentionally designed.

Third is the development of **stronger generative models themselves**. A model capable of producing plausible code, mathematical problems, translations, images or structured explanations is also a potential data generator. This creates an important feedback mechanism: more capable models can help create training material for later or smaller models, while automatic verification makes synthetic generation particularly attractive where answers can be objectively checked.

Fourth is **scarcity**. Rare diseases, unusual driving conditions, minority languages, specialist industrial faults and uncommon safety events may be poorly represented in ordinary datasets. Deliberate generation or simulation allows developers to oversample these cases rather than waiting for them to occur naturally.

Fifth is **privacy and access**. Medical records and many organisational datasets cannot simply be published or shared. Synthetic-data techniques can reduce direct dependence on identifiable records and make some forms of experimentation or collaboration easier, although they do not eliminate privacy risk. Singapore's regulatory guidance illustrates the emergence of synthetic data as part of the wider privacy-enhancing-technology landscape.

Finally, legal and transparency requirements are making data provenance more consequential. The European Union's AI Act requires providers of general-purpose AI models within scope to maintain copyright policies and publish summaries of the content used for training. The Commission's mandatory summary template explicitly covers sources including publicly available data, scraped data, private datasets, user data and synthetic data. These requirements do not resolve underlying copyright or data-rights disputes, but they increase the operational importance of knowing where training content came from.

6. What are the present limitations?

The first limitation is that **synthetic data inherits limitations from its generator**. If a source model misunderstands a subject, under-represents a population or repeatedly produces a stereotyped pattern, generating more examples can amplify rather than correct the problem. Synthetic volume is therefore not equivalent to additional independent knowledge.

Related to this is the risk of **loss of diversity and model-data feedback**. A 2024 Nature study demonstrated “model collapse” when successive generative models were trained recursively on model-generated data that displaced the original distribution, with lower-probability parts of that distribution progressively lost. The result should not be interpreted as showing that all synthetic training causes collapse: subsequent work and the experiments themselves indicate

that retaining original real data materially changes the outcome. The warning concerns indiscriminate recursive replacement, not carefully controlled hybrid training.

Data quality remains difficult to measure. Filters that successfully remove spam or duplicated material can also remove legitimate dialects, specialist writing or culturally distinctive content. Automated notions of “educational” or “high quality” may themselves reflect the preferences embodied in the classifier. Large-scale curation therefore involves trade-offs rather than an objectively perfect cleaning process.

Representation is uneven. English, Chinese and several other high-resource languages have vastly more digital material than many of the world's languages. The same applies between common and rare diseases, wealthy and poorly measured populations, highly digitised and less digitised economies, and common versus unusual physical situations. Synthetic generation can enlarge these categories but cannot reliably invent knowledge that was absent or incorrectly represented in its source material. Research on low-resource languages continues to show that lack of suitable training corpora remains a substantive constraint.

Privacy is another qualification. Synthetic data created from sensitive records may reproduce aggregate relationships that are themselves confidential or, if generated without adequate safeguards, retain information that increases re-identification risk. Singapore's guidance consequently recommends explicit assessment of residual re-identification risks rather than treating the term “synthetic” as synonymous with “anonymous”. Medical research likewise shows a continuing trade-off between fidelity, utility and privacy.

Provenance and rights remain difficult at web scale. Developers may combine public web material, licensed collections, proprietary datasets, user interactions and generated data through multiple training stages. Detailed disclosure is often limited: Stanford's 2026 AI Index found declining transparency around training datasets among leading developers. Regulatory requirements such as the EU's public training-content summaries improve visibility but do not make every item in a very large training corpus traceable or establish that every legal question has been resolved.

Evaluation is itself vulnerable to data problems. If benchmark questions or close variants enter training data, apparent performance can overstate genuine generalisation. Synthetic examples can also be highly realistic while lacking the causal relationships or difficult edge cases that matter in deployment. A statistically similar synthetic patient dataset, for example, does not automatically establish clinical validity.

The fundamental limitation is therefore not an inability to generate data. AI can already generate extraordinary quantities. The harder problem is establishing that the generated or selected information is **correct, sufficiently diverse, representative, independent where necessary, legally usable and demonstrably useful for the downstream system.**

7. What you should be aware of

The most important qualification is that the contemporary AI data story is **not simply a transition from human data to synthetic data.**

Real-world data remains foundational. It provides observations of language, people, organisations, biology and physical environments that models cannot obtain merely by repeatedly sampling their own prior representations. Synthetic data is most compelling when it has a defined purpose: creating scarce examples, providing automatically verifiable exercises, balancing a dataset, transferring capability from one model to another, expanding controlled physical scenarios or reducing direct exposure of sensitive records.

Equally, “real” does not mean “good”. A billion duplicated, inaccurate or irrelevant documents can be less useful than a much smaller carefully curated collection. Modern AI development increasingly combines large-scale authentic data with aggressive filtering, specialised datasets, synthetic examples and deliberately engineered post-training material. Microsoft's Phi work and the wider 2026 evidence concerning curation illustrate why data composition can matter as much as headline dataset size.

Synthetic data should also not be assumed to be unbiased, private or accurate merely because no particular synthetic record describes a real person or event. Those properties require separate evaluation. A generated dataset can retain the biases of its source, reveal information about source records, create unrealistic correlations or omit rare but consequential cases.

Similarly, **availability is not demonstrated effectiveness**. A synthetic-data generator capable of producing millions of images establishes generation capacity. Demonstrating that those images improve a deployed medical, industrial or autonomous system requires downstream evaluation against relevant real-world conditions.

The same Status discipline applies here as elsewhere in the Technology Landscape: demonstration, deployment and repeated effectiveness at scale are different evidential states. Synthetic data is unusual because creating more material can be technically straightforward while establishing its value may be considerably harder.

There is also an important global dimension. The internet is not a statistically representative sample of humanity. Languages, cultures, countries, sectors and populations differ substantially in how much usable digital information exists about them. Larger AI datasets therefore do not automatically produce more geographically or culturally balanced AI. Open multilingual projects, professionally translated datasets and local-language research are expanding coverage, but data inequality remains part of the current technological landscape.

As at 30 August 2026, the central development is consequently a shift from treating data principally as a raw resource to treating it increasingly as an **engineered component of AI capability**. Large real-world datasets remain indispensable, but filtering, provenance, specialised collections, model-generated supervision, simulation and synthetic augmentation are becoming increasingly important. The major advance is not that AI has solved its need for real data. It is that developers have acquired far more sophisticated ways to decide **which data a model should learn from, how those data should be constructed and where artificial data can usefully supplement what the real world provides**.

Evidence Notes

1. **Stanford Institute for Human-Centered Artificial Intelligence — AI Index Report 2026, Research and Development.** Material overview of current frontier-model data trends, including declining training-data transparency, increasing importance of curation and the evidence concerning synthetic data in pre-training. Particularly important qualification: the report does not find definitive evidence that synthetic data can fully replace real data for general-purpose pre-training.
2. **Microsoft Research — Phi-4 Technical Report, December 2024; Phi-4 reasoning technical materials, 2025.** Primary developer evidence that deliberately generated synthetic material, filtered organic data and synthetic reasoning demonstrations are being used substantially in contemporary training and post-training. Used as evidence of training methodology rather than independent proof of all claimed benchmark advantages.
3. **DeepSeek — DeepSeek-R1, 20 January 2025.** Primary developer evidence concerning reasoning-model post-training and release of smaller distilled models trained from R1 outputs. Developer benchmark claims should not be treated as independent real-world evaluation.
4. **Kabra et al. — “Learning from Synthetic Data Improves Multi-hop Reasoning”, ICLR 2026.** Research evidence that rule-generated synthetic examples can improve transfer to real-world reasoning benchmarks in a bounded experimental setting.
5. **Stan et al. — “Learning from Reasoning Failures via Synthetic Data Generation”, AAAI 2026.** Research evidence concerning targeted generation of more than 550,000 multimodal training examples and downstream improvements. Establishes the potential value of selectively generated corrective data rather than synthetic data generally.
6. **Hugging Face — FineWeb2, 2024–2025 releases.** Open dataset and technical documentation illustrating current scale of multilingual web-data processing, filtering and deduplication: approximately five billion documents and more than three trillion words across extensive multilingual coverage. Dataset-provider evidence rather than a universal measure of language representation or quality.
7. **Seto et al. — Findings of ACL 2025, “Training Bilingual LMs with Data Constraints in the Targeted Language”; MLCommons Datasets Working Group, 2025.** Evidence that availability of high-quality training material remains strongly unequal between languages and that dedicated data infrastructure remains necessary for low-resource languages.
8. **Sun et al. — Nature Biomedical Engineering, 5 March 2025; Sheng et al., 8 April 2025.** Peer-reviewed evidence that disease-conditioned synthetic medical images can improve data efficiency and generalisation in a specific ophthalmic foundation-model setting. This does not establish equivalent benefits across healthcare applications.
9. **Kaabachi et al. — npj Digital Medicine, 27 January 2025.** Scoping review of privacy and utility metrics for medical synthetic data. Particularly important evidence that no standardised consensus yet exists for evaluating both privacy and utility.
10. **Nature Editorial — “Synthetic data can benefit medical research — but risks must be recognized”, 10 September 2025.** Independent synthesis concerning opportunities and validation risks associated with synthetic medical data, including potential value where real-world data are scarce.
11. **NVIDIA — Cosmos physical-AI data tools, 18 March 2025, together with 2025 AV materials.** Developer evidence of operational tooling for large-scale synthetic robot and autonomous-vehicle training data, including controlled variation of physical conditions. Used principally as evidence of technological availability; claimed downstream benefits require application-specific verification.
12. **Shumailov et al. — Nature 631, 24 July 2024, “AI models collapse when trained on recursively generated data”.** Peer-reviewed evidence demonstrating degradation when generated data recursively displaces original distributions. Important qualification: the result does not establish that all synthetic training is harmful; preserving original real data materially changes the problem.

13. **Singapore Personal Data Protection Commission — Proposed Guide on Synthetic Data Generation, July 2024.** Public-sector guidance treating synthetic data as a privacy-enhancing technology while explicitly requiring assessment and management of residual re-identification risk.
 14. **European Commission / EU AI Act — Article 53 and Public Summary of Training Content template, 2024–2025; obligations applicable from 2 August 2025.** Primary regulatory evidence concerning copyright policies and public summaries of general-purpose model training content, including categories covering scraped, private, user and synthetic data. These disclosure obligations do not themselves determine the lawfulness of individual training items.
-

14. World Models & Simulation

Status as at 30 August 2026

Status snapshot

World Models & Simulation span established simulation technologies and a newer generation of AI systems that learn, predict or generate representations of environments and how they change. Their significance lies increasingly in allowing AI systems and developers to explore possible future states, generate varied environments and connect simulated experience with planning and action.

Maturity varies substantially. Conventional physical simulation, robotics simulation and digital twins are established technologies, while learned world models are technically well demonstrated in research and are increasingly being applied in robotics, autonomous-driving development and agent training. Generative systems can now create controllable driving scenarios, persistent or interactive environments and plausible sequences of future events; predictive models can also support planning without generating photorealistic worlds. General-purpose interactive world generation and world models intended to support broad physical intelligence remain much earlier.

The principal direction of development is towards richer, more persistent and action-responsive models that combine learned prediction, generative environments and established simulation infrastructure, particularly for physical AI.

The essential qualification is that a convincing generated environment is not necessarily an accurate model of reality. Physical fidelity, causality, long-horizon consistency, evaluation and transfer from simulation to real conditions remain material limitations. As at **30 August 2026**, world models are becoming increasingly useful components of AI training, simulation and planning, but they do not demonstrate comprehensive machine understanding of the physical world.

1. What is it?

A **world model** is an AI system that represents or predicts aspects of an environment sufficiently to anticipate what may happen next, including in some cases how the environment may change when an agent takes an action. The underlying idea is not new: AI and robotics have long used learned models of environments to support planning and control. What has changed is the scale, richness and variety of environments that newer AI systems can represent, predict or generate.

The term nevertheless requires care because it does not describe one standard architecture or capability. By 2026, “world model” is being applied to several related forms of AI: models that learn compact internal representations of environmental dynamics; models that predict future states from observations and actions; video-generative models that produce plausible future scenes; interactive systems that generate navigable environments; and models designed specifically to support robotics, autonomous driving or other forms of embodied AI. Recent

surveys consequently describe a field spanning generative simulation, predictive representations and decision-coupled models rather than one uniform technology.

A useful distinction is between a **world model and a conventional simulator**. A traditional physics simulator is explicitly programmed with mathematical rules describing such matters as motion, collisions, forces or materials. A learned world model instead infers regularities from data and uses those learned representations to predict or generate what might happen. The two approaches increasingly overlap. AI models can augment conventional simulators, simulations can generate experience for training AI, and learned world models can provide an additional form of simulation where manually constructing every rule and environment would be impractical.

World models also differ from ordinary video generation. A generated video may look convincing without representing how actions causally affect an environment. For world modelling, appearance alone is insufficient: persistence, spatial consistency, temporal dynamics, action conditioning and physical plausibility become increasingly important. A visually convincing car crash, robot movement or bouncing object is not necessarily an accurate simulation of the corresponding physical event.

This distinction is central to understanding the current landscape. Google DeepMind's Genie 3, for example, generates interactive environments from text that can be navigated in real time at around 24 frames per second and remain broadly consistent for several minutes. Meta's V-JEPA 2 takes a different approach, learning predictive representations from large quantities of video and demonstrating planning and robot-control capabilities after comparatively limited robot-specific training. NVIDIA's Cosmos family is oriented towards world generation, prediction and reasoning for robotics and autonomous vehicles. These systems are all described as world models, but they perform different functions and should not be treated as directly interchangeable.

Simulation is broader still. AI-enabled simulation includes digital twins, robotics simulators, autonomous-driving environments and other virtual representations used to train systems, test behaviour, generate scenarios or examine possible outcomes. World models increasingly complement this established simulation infrastructure rather than replacing it.

World Models & Simulation in August 2026 is therefore best understood as an emerging **predictive and generative layer of AI**: technology that allows machines or developers to construct representations of environments, explore possible future states and—in increasingly sophisticated cases—learn or evaluate actions within those representations.

2. What can it currently do?

Current systems demonstrate several distinct capabilities.

First, world models can **predict how observed situations may develop**. Instead of processing a single image independently, predictive models learn temporal relationships: an object is moving in a particular direction, an action changes its state, or one event tends to follow another. Meta's V-JEPA 2 was trained principally from more than one million hours of video and demonstrated strong results in motion understanding, action anticipation and video

reasoning. After additional training using robot trajectories, the system was also demonstrated planning actions for robots manipulating unfamiliar objects. These are significant research capabilities, but they do not establish general physical understanding.

Second, generative world models can **create plausible future scenarios**. NVIDIA's Cosmos models can generate or transform video representing possible physical environments and are being developed for applications including robotics and autonomous vehicles. This allows developers to vary conditions such as environments, objects, viewpoints and events without physically collecting every corresponding real-world example.

Third, systems can generate **interactive environments rather than fixed videos**. Genie 3 represented an important step in this direction. From a text description it can create a visually rich environment that responds in real time to navigation and other inputs, while retaining elements of environmental consistency over several minutes. Google DeepMind has connected such generated environments to its SIMA agent research, allowing an agent to act within a world generated by another AI system.

Fourth, some models can create **persistent three-dimensional spaces**. World Labs' Marble, introduced in limited beta in September 2025 and subsequently released more broadly, generates navigable 3D environments from inputs including text and images. Unlike an autoregressively generated video world, its explicit 3D representation can remain spatially persistent during exploration. This illustrates another important division within the category: some systems simulate the experience of moving through a world by continually generating what should appear next, while others construct a reusable spatial representation.

Fifth, world models can support **counterfactual generation**: exploring situations that did not occur in the original data. This is particularly relevant where important events are rare, hazardous or expensive to reproduce. Wayve's GAIA-2 driving world model, for example, generates controllable multi-view driving scenarios and is intended to expose automated-driving systems to varied locations, vehicles, weather and safety-critical situations.

Finally, learned world models can support **decision-making without reproducing every detail visually**. The Dreamer line of research demonstrates this more abstract use. DreamerV3 learns a model of an environment and trains behaviour using imagined outcomes inside that learned model. Research published in *Nature* in 2025 demonstrated one configuration operating across more than 150 diverse control tasks without task-specific tuning. This is strong evidence of the usefulness of learned predictive models for reinforcement learning, although performance across controlled tasks is different from dependable operation in unrestricted physical environments.

The common capability is therefore not “creating virtual worlds”. It is **learning enough structure about an environment to represent, predict, generate or reason about possible states and actions**. How much structure has actually been learned—and how faithfully it corresponds to reality—varies substantially.

3. Where is it currently being applied?

The clearest current applications are in areas where real-world experimentation is expensive, dangerous, slow or insufficiently diverse.

Robotics is a major application. Simulators have long allowed robots to practise navigation and manipulation without requiring every training action to be performed by physical hardware. Generative world models extend this approach by potentially creating much greater variation in scenes, objects and events. NVIDIA has positioned Cosmos and Omniverse together for robot learning, synthetic scenario creation and simulation, with robotics developers among the reported adopters. The evidence establishes active developer use and experimentation, although it does not demonstrate that generative world models have displaced conventional simulation or real-world robot training at scale.

Autonomous and assisted driving provides another significant application. Driving systems must encounter ordinary traffic conditions but also rare situations—unusual pedestrian movements, dangerous cut-ins, difficult weather or uncommon road configurations—that may be costly or unsafe to collect repeatedly in reality. Wayve's GAIA-2 and NVIDIA's Cosmos illustrate the use of generative world models to create controllable driving scenarios for training and evaluation.

Agent training and evaluation is an important research application. Instead of relying only on hand-built game worlds or simulation environments, systems such as Genie can generate new interactive settings in which agents can navigate and attempt tasks. DeepMind's experiments combining Genie 3 with SIMA demonstrate this concept directly. The current significance lies less in widespread commercial deployment than in the possibility of expanding the diversity of environments available for agent research.

Industrial simulation and digital twins form a more established neighbouring landscape. Manufacturing facilities, warehouses, robots and physical assets can be represented virtually so that layouts, workflows and machine behaviour can be examined before or alongside changes to physical systems. AI-generated environments and reconstructed real-world scenes are increasingly being connected with these simulation systems. NVIDIA's 2025 Omniverse developments, for example, combined world reconstruction, physically based simulation and Cosmos-generated data within a physical-AI development environment.

Interactive 3D content and spatial computing constitute another emerging application. Generated persistent environments can be used as starting points for virtual experiences, design exploration and interactive content rather than solely as training environments for autonomous machines. World Labs' Marble demonstrates that world-generation technology is already accessible beyond closed research demonstrations, although this remains an early product category rather than evidence of broad adoption.

China has also become a significant centre of embodied world-model research. Academic surveys from Chinese institutions and 2026 industry activity show increasing attention to world models for manipulation, physical reasoning and embodied intelligence. Shanghai-based ACE Robotics, for example, reported world-model results across embodied-AI benchmarks in July 2026. Such developer-reported benchmark results are evidence of active technological development, not independent evidence of superior real-world robotic capability.

The current application pattern is consequently uneven. Conventional simulation and digital twins are already operational technologies. Learned world models are increasingly being incorporated into robotics, driving and agent-development workflows, but much of the most ambitious activity remains research, development and early deployment rather than established infrastructure used throughout industry.

4. How developed is it?

World Models & Simulation cannot sensibly be given one maturity level.

At the most mature end, **conventional physical simulation, robotics simulation and digital twins** are established engineering technologies. They are routinely used to develop machines and processes without requiring every test to occur physically.

Learned world models for reinforcement learning are also technically well established as a research approach. Systems can learn predictive representations and use imagined trajectories to improve policies, and results such as DreamerV3 demonstrate substantial capability across diverse controlled tasks.

Generative world models for specialised domains, particularly autonomous driving and robotics development, occupy a developing stage. Platforms and models are available, organisations are integrating them into development pipelines, and increasingly sophisticated scenario generation is possible. NVIDIA's Cosmos releases and Wayve's GAIA-2 are examples. However, product availability and developer adoption do not establish that generated worlds have become sufficiently faithful to substitute generally for real-world testing.

General-purpose interactive world generation remains earlier. Genie 3 can generate diverse navigable environments in real time, but DeepMind itself identifies limits involving action space, multiple-agent interaction, accurate reproduction of real locations and interaction duration. Its continuous interaction horizon is measured in minutes rather than hours.

World models as reliable internal models of real-world physics for general autonomous agents remain principally a research objective. V-JEPA 2 demonstrates that large-scale observational learning can contribute to prediction and robot planning, but demonstrations involving selected robots and tasks do not establish a comprehensive physical model applicable across arbitrary environments.

Across these different maturity levels, capability, availability and operational maturity should not be inferred from one another.**5. What is driving its development and use?**

The first major driver is the **cost and difficulty of acquiring physical-world experience**. Robots and autonomous vehicles cannot economically or safely experience every possible situation through real-world operation alone. Simulation offers a way to multiply training and testing opportunities, while generative world models potentially increase the variety of those experiences.

A second driver is the enormous quantity of **video and multimodal data** now available for learning environmental regularities. V-JEPA 2 illustrates the attraction of this approach: large quantities of ordinary video can be used to learn representations of motion and physical interaction before much smaller quantities of robot-specific data are introduced.

Third is the improvement of **generative video and spatial modelling**. Advances that make generated video more realistic, temporally coherent and controllable are also making it increasingly useful as a form of simulation. World models, video generators and 3D reconstruction technologies are consequently becoming less distinct technological areas than they once were.

Fourth is demand for **rare and counterfactual scenarios**. Safety-critical systems cannot be evaluated adequately if their development data contain only common situations. Generating unusual weather, road events, object configurations or physical interactions offers a way of increasing scenario diversity, particularly in autonomous-driving and robotics development.

Fifth is the growth of **embodied AI and autonomous agents**. An agent expected to act over multiple steps benefits from anticipating consequences rather than reacting only to the present observation. This is creating renewed interest in predictive models that can effectively ask, “If I do this, what is likely to happen next?”

Sixth is improving **AI compute and simulation infrastructure**. Generating interactive high-resolution environments in real time and running large numbers of simulated experiences are computationally demanding. More capable accelerators, cloud infrastructure and specialised simulation platforms are therefore enabling forms of world modelling that were previously impractical.

Finally, there is strong scientific interest in whether systems can learn **general representations of physical and spatial regularities** rather than requiring every task to be separately programmed. Research in North America, Europe and Asia reflects this wider objective, while commercial activity in robotics, autonomous vehicles and spatial computing is providing practical contexts in which the technology can be tested.

6. What are the present limitations?

The most important limitation is that **plausibility is not the same as physical correctness**.

A generated environment may look convincing while violating conservation, geometry, object permanence, contact mechanics or causal relationships. This matters far more when a simulation is used to train a robot or validate an autonomous vehicle than when it is used simply to create entertaining visual content.

Evaluation research in 2026 illustrates the problem. GAUGE was introduced specifically to measure physical fidelity across conventional simulation engines and generative video world models using controlled real-world tasks, while other work has found that physical inconsistencies in generated worlds can cause differences between simulated and real robot-task performance.

Long-horizon consistency is another major constraint. Small prediction errors can compound as a model repeatedly predicts its own future states. Objects may change form, spatial relationships may drift, or an environment may gradually depart from the state implied by earlier interactions. Genie 3's ability to maintain broad consistency for several minutes is substantial progress, but its own documented limitations show how far this remains from indefinitely persistent simulation.

Causality is difficult to establish from observation. Video can reveal that events occur together or in sequence without necessarily revealing the underlying physical mechanism. A model may learn statistical regularities that work in familiar situations yet fail when an unusual intervention changes the causal structure.

Rare events remain difficult even when simulation is intended to solve rare-event scarcity. A generative model can create an unusual scenario, but the usefulness of that scenario depends on whether the generated behaviour is realistic. Inventing more examples does not guarantee that those examples correctly represent the real phenomenon.

Sim-to-real transfer remains a central problem. Conventional simulators already encounter a “reality gap”: policies that work in simulation may fail when confronted with friction, lighting, sensor noise, deformation, human behaviour or other complexities of the physical world. Learned world models do not automatically remove this gap. If anything, they add another source of uncertainty because their dynamics are inferred from data rather than specified explicitly.

Evaluation itself remains immature. Visual quality measures do not adequately test physical behaviour, while task-specific evaluation can be expensive and difficult to compare between systems. The emergence of benchmarks explicitly testing physical realism and temporal consistency during 2026 is therefore evidence both of progress and of a significant unresolved measurement problem.

Compute requirements can be substantial. Real-time interactive generation requires continual prediction while incorporating previous states and new actions. High-fidelity simulation at scale can also require significant processing resources, particularly when multiple cameras, agents or possible futures are involved.

There are also **safety and governance questions**. A world model used for entertainment can tolerate errors that would be unacceptable if the same modelling approach influenced robotic planning or safety-critical evaluation. Confidence in a simulated outcome must therefore be matched to the consequences of being wrong.

Finally, the terminology itself can exaggerate maturity. Calling a system a “world model” can suggest that it possesses a comprehensive representation of reality. Present systems demonstrate something narrower: useful learned representations of particular aspects of environments, at particular spatial and temporal scales, under particular data and task conditions.

7. What you should be aware of

The most important qualification is that **there is no single accepted technological object called a world model**. The term currently covers predictive latent representations, learned dynamics models, generative video systems, interactive environments and models connected directly to decision-making. Comparisons therefore require attention to what a particular system is actually modelling and what it is intended to do.

Second, **a compelling generated world is not evidence of comprehensive world understanding**. Photorealism can be produced from statistical regularities without guaranteeing accurate physics, causality or persistent spatial structure. For training and evaluation, correctness of consequences can matter more than visual realism.

Third, **simulation and reality remain complementary**. Simulation can provide enormous quantities of inexpensive, controllable and sometimes otherwise unobtainable experience. Real-world evidence remains necessary to determine whether behaviour learned or tested in simulation transfers successfully to actual environments.

Fourth, **world models need not reproduce the world visually to be useful**. Some of the most important systems operate largely through internal representations, predicting consequences in a compressed learned space rather than generating photorealistic scenes. World modelling should therefore not be equated with AI video generation.

Fifth, **conventional simulation has not become obsolete**. Explicit physics engines offer controllability, known rules and measurable physical quantities that generative systems may not reliably reproduce. The developing landscape increasingly combines learned models, conventional simulation, reconstructed environments and real-world data rather than selecting one approach exclusively.

Sixth, **research capability and operational maturity remain widely separated**. An agent successfully manipulating an unfamiliar object, a model generating a realistic driving incident or a person navigating a generated environment can establish an important capability. It does not establish that the system can accurately model arbitrary physical situations or support safety-critical decisions without extensive validation.

There are also geographical differences in emphasis rather than a single global development pattern. US-based technology organisations have produced several prominent general-purpose and physical-AI world-model platforms; European activity includes significant autonomous-driving development such as Wayve; and Chinese universities and companies are increasingly active in embodied world models, simulation and physical-intelligence evaluation. The research literature itself is highly international. The relevant Status is therefore global, although development capability and access to the required compute, data and physical-AI ecosystems remain uneven.

As at 30 August 2026, World Models & Simulation occupy an important but still developing position in the AI landscape. Simulation is already indispensable to many engineering and AI-development processes. Learned world models are extending what simulation can do by generating environments, predicting possible futures, learning from video and allowing agents to practise within models of situations that may never have been explicitly programmed.

The significant current development is not that AI has acquired a complete internal understanding of the physical world. It is that **AI systems are becoming increasingly capable of learning useful representations of how environments behave, using those representations to generate and explore possible futures, and connecting simulated experience more directly with planning, robotics and autonomous systems.**

Evidence Notes

1. **Google DeepMind — Genie 3: A New Frontier for World Models, 5 August 2025.** Primary developer evidence for real-time interactive environment generation at approximately 24 frames per second, 720p resolution and several minutes of environmental consistency. Particularly useful because DeepMind also identifies current limitations involving action range, multi-agent interaction, geographical accuracy and duration. Developer research evidence rather than evidence of scaled operational deployment.
2. **Google DeepMind — SIMA 2, 2025.** Evidence of agents being connected with interactive generated environments and of self-improvement experiments within Genie 3 worlds. DeepMind describes SIMA 2 as research and identifies continuing limitations in long-horizon tasks, memory, precise action and robust visual understanding.
3. **Meta — V-JEPA 2, 11 June 2025.** Research evidence concerning predictive representation learning from more than one million hours of video, physical understanding benchmarks and robot planning/control after additional robot-interaction training. Demonstrates research capability rather than general robotic autonomy.
4. **Hafner et al. — “Mastering Diverse Control Tasks through World Models”, *Nature*, 2 April 2025.** Peer-reviewed evidence concerning DreamerV3 and the use of learned world models to train behaviour through imagined experience across more than 150 control tasks. Strong evidence for world-model-based reinforcement learning under experimental conditions; not evidence of unrestricted real-world autonomy.
5. **NVIDIA — Cosmos World Foundation Model Platform, 6 January 2025; subsequent Cosmos releases during 2025–26.** Primary developer evidence concerning generative and predictive world models for robotics and autonomous vehicles, together with integration into physical-AI development infrastructure. Reported adopters establish industry engagement but do not independently verify claimed improvements in downstream systems.
6. **NVIDIA — Omniverse and Cosmos physical-AI developments, 11 August 2025.** Evidence concerning integration of world reconstruction, digital twins, simulation and generative world models for robotics development and synthetic-data generation. Company evidence.
7. **Wayve — GAIA-2, 26 March 2025.** Developer research evidence for a controllable multi-view generative driving world model intended to generate geographically diverse and safety-critical driving scenarios for training and evaluation. Useful evidence of specialised application; downstream safety or deployment benefits should not be inferred solely from the developer's claims.
8. **World Labs — Marble, September–November 2025.** Developer evidence concerning generation of persistent navigable 3D environments and the transition from limited beta to wider availability. Demonstrates a distinct explicit-spatial approach to generated worlds. Product availability does not establish widespread adoption.
9. **Long et al. — “Learning Embodied Intelligence from Physical Simulators and World Models”, 1 July 2025.** Research survey distinguishing physical simulators from learned world models and examining their

complementary roles in embodied AI, including simulation-to-reality transfer. Survey evidence rather than direct deployment evidence.

10. **Li et al. — “A Comprehensive Survey on World Models for Embodied AI”, version revised 25 June 2026.** Research synthesis covering predictive and decision-coupled world models across robotics, driving and generative settings. Identifies long-horizon consistency, computational efficiency, evaluation and physical fidelity as continuing challenges. Preprint/survey evidence rather than peer-reviewed proof of individual system performance.
 11. **Nguyen & Hong — “World Models: A Comprehensive Survey”, June 2026.** Broad synthesis particularly useful for establishing the diversity of meanings attached to “world model” and recurring challenges involving causal and physical reasoning, long-horizon consistency, evaluation and simulation-to-reality transfer. Working-paper evidence.
 12. **Wang et al. — GAUGE, 6 August 2026.** Research benchmark comparing physical fidelity in simulation engines and video world models across controlled real-world-grounded physical tasks. Important evidence that visual or perceptual quality is insufficient for evaluating whether generated environments reproduce real physical behaviour. Recent preprint evidence.
 13. **Oh & Yun — Reference-Free Assessment of Physical Consistency in World Model-based Video Generation, 21 June 2026.** Research evidence concerning discrepancies between generated simulation and real robot-task outcomes and methods for identifying physical inconsistencies. Preprint evidence and therefore appropriately treated as evidence of an active evaluation problem rather than settled measurement practice.
 14. **LoViF 2026 PhyScore Challenge, May 2026.** Evidence of emerging evaluation methods that distinguish video quality from physical realism, condition alignment and temporal consistency. Useful evidence that world-model assessment is expanding beyond visual plausibility alone.
 15. **Tsinghua University and collaborating institutions — Survey of Embodied World Models, 2025–26.** Evidence of substantial Chinese academic activity in embodied world modelling and of the breadth of approaches spanning generative, latent and decision-oriented models. Survey/preprint evidence.
 16. **ACE Robotics — Kairos World Model, July 2026.** Developer-reported evidence of Chinese world-model development and benchmark participation in embodied AI. Benchmark leadership claims are company-reported and should not be treated as independent evidence of superior deployed robotic performance.
-

15. AI Infrastructure & Compute

Status as at 30 August 2026

Status snapshot

AI Infrastructure & Compute is the underlying system of processors, memory, networking, servers, software, data centres, electricity and cooling that enables AI models to be trained and operated. Much of this foundation—cloud computing, data centres, servers and accelerated computing—is already mature and deployed at large scale. What is changing rapidly is its scale and specialisation for AI: increasingly dense accelerator clusters, high-bandwidth memory, high-speed interconnects and infrastructure designed separately or specifically for demanding training and inference workloads.

Current infrastructure demonstrably supports frontier-model training, large-scale inference, enterprise and scientific AI, and shared national computing facilities. Access is expanding through commercial cloud services and public or sovereign programmes in Europe, China, Japan, India and other regions, although the largest concentrations of advanced compute remain geographically and organisationally uneven.

The principal direction of development is towards integrated AI computing systems rather than simply more powerful individual chips. Processor performance increasingly depends on memory, networking, software, power and cooling operating effectively together. At the same time, inference is becoming a more important infrastructure demand as deployed AI systems serve growing numbers of users and perform more computationally intensive tasks.

The key qualification is that announced investment or planned data-centre capacity is not the same as operational compute, and greater compute does not itself establish greater AI effectiveness. As at 30 August 2026, AI infrastructure is established at scale but undergoing rapid expansion and redesign, increasingly constrained by energy, supply chains, cost, networking and geographic access.

1. What is it?

AI Infrastructure & Compute is the physical and digital foundation that allows artificial-intelligence systems to be trained and operated. It includes the processors that perform calculations, the high-speed memory that keeps those processors supplied with data, the networks connecting large numbers of machines, the servers and storage systems surrounding them, the software that coordinates the hardware, and the data centres, electricity and cooling systems that keep the whole system operating.

The term **compute** is often used as shorthand for the computational resources available to an AI system. But an AI computer is not simply a collection of powerful chips. Modern large-scale AI increasingly depends on the performance of the entire system. Thousands of accelerators may have to exchange information rapidly; models and intermediate calculations must move between processors and memory; storage must supply data; and heat generated by

densely packed equipment must be removed continuously. A bottleneck in memory, networking, electricity or cooling can limit the usefulness of otherwise powerful processors.

Two broad workloads are important to distinguish.

Training is the computational process through which an AI model learns from data or is subsequently adapted. Large frontier-model training runs can involve thousands or tens of thousands of specialised processors operating together for extended periods. Epoch AI's historical analysis found that compute used to train frontier AI models had been increasing by approximately four to five times per year, although estimates are necessarily incomplete because developers disclose increasingly little information about major training runs.

Inference is what happens when a trained model is actually used: answering a question, generating an image, analysing a document, interpreting video or carrying out another task. Individual inference requests may require far less computation than training a frontier model, but serving very large numbers of users can create enormous aggregate demand. More computationally intensive reasoning, long-context processing, multimodal systems and agentic workloads can increase the amount of computation required for each useful result.

This distinction is increasingly shaping infrastructure itself. Google's Ironwood, introduced in 2025, was its first Tensor Processing Unit (TPU) designed specifically around large-scale inference; in April 2026 Google introduced separate eighth-generation TPU 8t and TPU 8i systems optimised respectively for training and inference. NVIDIA's Vera Rubin architecture similarly integrates processors, networking, memory and rack-scale systems around increasingly demanding AI workloads. These are developer descriptions, but they illustrate an important change: AI infrastructure is being designed as an integrated computing system rather than as isolated chips.

The resulting landscape ranges from conventional enterprise servers and cloud services through specialised GPU clusters and national supercomputers to enormous purpose-built AI data centres. It also includes multiple processor architectures. Graphics processing units (GPUs) remain central to large-scale AI, but specialised accelerators such as Google's TPUs and other custom processors are increasingly important. CPUs continue to perform essential supporting work, while high-bandwidth memory (HBM), advanced networking and specialised software determine how effectively the processors can be used.

AI infrastructure should therefore be understood as a **stack of interdependent resources**. Advanced AI depends not simply on access to a particular semiconductor, but on the ability to combine computation, memory, communication, storage, software, power and cooling at the required scale.

2. What can it currently do?

Current AI infrastructure can support computational workloads at scales that would have been exceptional even a few years ago.

At the frontier, large clusters allow many thousands of accelerators to operate as a coordinated system. Epoch AI's 2025 study of more than 500 AI supercomputers found that their

computational performance had been increasing by approximately 2.5 times per year, largely through the combination of more processors and more capable processors. Its later tracking indicates that the largest known installations have reached computing capacity equivalent to very large fleets of current-generation AI accelerators. These estimates are reconstructed from public evidence rather than complete operator disclosures, but they demonstrate the extraordinary scale of contemporary AI computing.

The infrastructure can train very large language, multimodal and generative models, fine-tune and adapt existing models, support reinforcement learning and other post-training processes, and serve models to large numbers of users. It also supports AI applications in scientific computing, recommendation, search, image and video generation, industrial simulation and other computationally demanding fields.

Independent benchmark activity provides evidence that the capability of complete AI systems continues to improve. MLCommons' MLPerf Training v6.0 results in June 2026 included new benchmarks reflecting sparse mixture-of-experts architectures and showed increasing diversity in the frameworks and systems used for AI training. Its April 2026 MLPerf Inference v6.0 release similarly expanded testing to newer reasoning and generative-AI workloads. Benchmarks do not establish the cost or effectiveness of a system in every real deployment, but they demonstrate continuing improvements across hardware and software rather than progress being confined to processor specifications alone.

Infrastructure has also become substantially more efficient at delivering established levels of AI capability. Stanford's 2025 AI Index reported that the inference cost of a model achieving approximately GPT-3.5-level performance on one widely used benchmark fell more than 280-fold between November 2022 and October 2024. Over a longer hardware trend, machine-learning hardware performance improved while price-performance and energy efficiency also advanced materially. These measures cannot be extrapolated mechanically to every contemporary workload, particularly newer reasoning systems, but they show that increasing AI capability does not translate directly into a fixed amount of infrastructure per task.

Large-scale public infrastructure is also becoming available outside the largest technology companies. Japan's ABCI 3.0 entered service in 2025 with 6,128 NVIDIA H200 GPUs, 75 petabytes of all-flash storage and theoretical half-precision peak performance of approximately 6.22 exaflops. India announced in May 2025 that shared compute capacity under the IndiaAI Mission had exceeded 34,000 GPUs. Europe is establishing a network of AI Factories around EuroHPC supercomputing infrastructure to provide computing resources to researchers, start-ups, industry and public institutions.

The practical capability is therefore broader than simply training the world's largest models. Contemporary infrastructure can provide shared model-development environments, cloud-based accelerator capacity, enterprise inference, scientific AI computing and nationally supported compute resources.

3. Where is it currently being applied?

The largest concentrations of AI compute remain associated with hyperscale cloud companies, major AI developers and very large technology platforms. These organisations use specialised infrastructure both to train advanced models and to operate AI services at high volume.

Cloud infrastructure makes some of this capability available to other organisations without requiring them to construct their own data centres. AI developers, enterprises, universities and public agencies can obtain accelerator capacity through general cloud providers and increasingly through specialised AI-compute providers. At smaller scales, organisations also deploy accelerators in their own data centres where economics, security, latency, data control or existing infrastructure make that appropriate.

A second important application is **public and sovereign computing infrastructure**. Governments increasingly regard access to advanced compute as part of national research, industrial and AI capacity.

Europe illustrates this clearly. By August 2026, the European Commission reported that 19 AI Factories and 13 associated AI Factory Antennas were being established around EuroHPC infrastructure. In July 2026 the European Union also launched procurement for up to seven much larger AI Gigafactories, envisaged as facilities combining more than 100,000 advanced AI processors with high-speed networks, cloud technology and energy-efficient data centres. The Gigafactories were prospective at the reference date rather than operational infrastructure and should therefore be distinguished from the existing and developing AI Factory network.

China has developed extensive national computing infrastructure. Official Chinese data reported intelligent computing capacity of 2,185 exaflops by the end of June 2026, with more than 70 major computing corridors established around national computing hubs. Definitions and measurement conventions mean this figure should not be treated as directly comparable with every other published measure of AI computing capacity, but it demonstrates the scale of China's infrastructure programme.

Other countries are building shared capacity at different scales. Japan's ABCI infrastructure supports domestic academic, public-sector and start-up research. India's common-compute programme is intended to broaden access to GPUs rather than confine advanced resources to a small number of large organisations. Brazil announced in August 2026 that it was preparing a new government-backed AI supercomputer as an expansion of national high-performance computing infrastructure.

The Middle East has meanwhile become an important location for very large proposed AI infrastructure. The United States and United Arab Emirates announced in May 2025 a framework including an initial 1-gigawatt AI data centre within a planned 5-gigawatt Abu Dhabi AI campus. Stargate UAE was subsequently announced as an infrastructure cluster within that campus. These announcements establish substantial committed development activity, but most of the planned multi-gigawatt capacity was not yet operating at the reference date.

Africa's installed AI-compute base remains considerably smaller. However, regional infrastructure is developing. Cassava Technologies announced deployment of NVIDIA-based

AI infrastructure beginning in South Africa, with planned expansion to Nigeria, Kenya, Egypt and Morocco. Such initiatives are significant for regional access and data-location requirements, but they should not be interpreted as evidence that the continent has reached infrastructure density comparable with North America, China or the leading Asian and European markets.

The geographic pattern is therefore highly uneven. AI compute is globalising, but access to the largest concentrations of advanced processors, energy, capital and supporting infrastructure remains concentrated.

4. How developed is it?

AI Infrastructure & Compute contains both extremely mature technologies and rapidly changing ones.

Data centres, cloud computing, CPUs, servers, storage and conventional networking are mature industries operating at enormous scale. GPU-accelerated computing is also established technology rather than an experiment. Large AI workloads already run routinely across commercial cloud and private infrastructure.

What is changing rapidly is the **scale, density and specialisation** of the infrastructure.

High-end AI clusters increasingly require tightly integrated accelerator systems, extremely high memory bandwidth, specialised networking and much higher electrical and cooling capacity per rack than conventional enterprise computing. AMD's MI350 series, for example, incorporates up to 288 GB of HBM3E per accelerator, while contemporary NVIDIA and Google systems similarly combine accelerators with purpose-built interconnects and large-scale memory architectures. Individual supplier specifications are not independent performance evidence, but collectively they demonstrate the direction of system engineering.

Liquid cooling is moving from a specialised technique towards an increasingly important component of high-density AI facilities. Uptime Institute observed in 2025 that rising AI rack densities were accelerating adoption of direct-liquid and immersion cooling, although adoption across conventional enterprise data centres remained slower.

At the largest scale, however, planned infrastructure should not be confused with operational capacity. Announcements of gigawatt-scale campuses, hundreds of thousands of processors or multi-billion-dollar investment programmes frequently describe facilities that will be constructed over several years. Land acquisition, financing, power agreements or processor orders are evidence of intended capacity, not evidence that equivalent AI computation is already available.

The present Status is therefore one of **mature underlying data-centre and cloud industries undergoing unusually rapid redesign for AI**, with the most advanced installations pushing into power densities, networking scales and capital requirements that are not yet typical of computing infrastructure generally.

5. What is driving its development and use?

The strongest driver is rising demand for AI computation itself.

Larger training runs have historically contributed substantially to improvements in frontier-model capability. At the same time, inference demand is increasing as generative AI moves from model development into widespread services and as newer systems perform more computation during use. Reasoning models, multimodal processing, long contexts and agents can all increase infrastructure requirements even when the underlying model has already been trained.

A second driver is continuing **hardware and software efficiency improvement**. More capable accelerators, lower-precision computation, improved memory systems, better interconnects and software optimisation allow more AI work to be extracted from each unit of hardware. MLPerf results demonstrate that progress is occurring through combinations of processors, systems and software rather than semiconductor advances alone.

Third is investment. The International Energy Agency reported in April 2026 that capital expenditure by five large technology companies exceeded \$400 billion in 2025 and was expected to rise substantially again in 2026, with data-centre investment a major component. This is evidence of exceptional infrastructure expansion, although expenditure does not by itself establish that all planned capacity will be economically productive.

Fourth is the increasing strategic importance attached to **compute access**. European AI Factories, India's shared GPU capacity, Japan's ABCI programme, China's computing infrastructure and emerging programmes elsewhere reflect concern that access to AI computation can influence research capability, domestic AI development and technological sovereignty.

Finally, the economics of inference create powerful incentives to reduce the amount of electricity, hardware and time required for each useful output. Once AI systems are used millions or billions of times, small improvements in utilisation, memory efficiency, batching, model architecture or processor performance can materially change operating costs.

6. What are the present limitations?

The most visible constraint is increasingly **electricity**.

The IEA estimated global data-centre electricity consumption at approximately 485 terawatt-hours in 2025 and projected roughly 950 TWh by 2030 in its April 2026 central assessment. AI-focused facilities were expected to grow considerably faster than data centres overall. These are projections rather than predetermined outcomes, and the IEA explicitly identifies infrastructure bottlenecks, efficiency improvements and AI demand as major uncertainties.

The issue is also local rather than merely global. Data centres concentrate very large loads at particular grid connection points. A country may have adequate total electricity generation while still lacking transmission capacity or available power in the specific location where a

large AI facility is proposed. Grid connections, transformers, substations and generation can take longer to develop than computing equipment.

Cooling and water can create additional constraints. Higher-density accelerators produce large quantities of heat, encouraging direct-liquid cooling and other specialised designs. Local climate, water availability, facility design and cooling technology determine the resulting resource requirements; there is no single water or cooling footprint applicable to all AI data centres. Australia's emerging policy treatment illustrates how power, water, grid stability and community impacts are becoming part of infrastructure planning rather than merely technical concerns inside data centres.

Supply chains remain another limitation. Advanced AI systems depend on leading-edge semiconductor fabrication, high-bandwidth memory, sophisticated packaging, networking equipment and specialised servers. These supply chains are internationally distributed and concentrated among relatively few producers. High-bandwidth memory in particular has become strategically important because processors cannot operate efficiently if model data cannot be moved quickly enough to and from them. In August 2026 SK Hynix described continuing tight memory supply while expanding HBM manufacturing and packaging capacity.

Networking presents a related technical challenge. Adding processors does not produce proportional performance if they cannot exchange data efficiently. At large scale, communication overhead can leave expensive processors waiting rather than computing. This is why modern AI infrastructure increasingly integrates accelerator design with switches, optical connections, network interfaces and software that determines how workloads are distributed.

Cost and utilisation are also significant constraints. Frontier-scale infrastructure can require billions of dollars of capital before generating useful output. The economics depend not only on purchasing processors but on keeping expensive systems productively utilised, supplying power, maintaining cooling and networking, and replacing hardware as technology changes. An installed accelerator is therefore not equivalent to an effectively utilised accelerator.

Geopolitics further affects availability. Advanced semiconductors and manufacturing equipment have become subject to export controls and technology-security policies, particularly in relation to China. These restrictions have encouraged development of alternative domestic hardware and supply chains while also producing substantial differences in which processors can be acquired in different markets. China's large installed computing base therefore coexists with constraints on access to some leading foreign technologies.

Finally, environmental impact cannot be inferred simply from processor counts or electricity consumption. Electricity sources differ markedly by location. The IEA estimated that data centres globally drew electricity from a mixture of coal, natural gas, renewables and nuclear power, with substantial regional differences. Hardware efficiency can reduce the energy needed for a given computation while rapid growth in total computation can simultaneously increase overall electricity use.

7. What you should be aware of

The most important qualification is that **compute is an enabler of AI capability, not AI capability itself**.

More processors can permit larger experiments, more training, more users and more computationally intensive inference. They do not guarantee that a model will be accurate, useful, safe or economically valuable. Algorithms, data, model architecture, software engineering and the quality of the wider system remain equally important.

Nor should processor counts be compared casually. A "GPU" is not a standard unit of useful AI computation. Different generations and types of accelerators can differ substantially in processing performance, memory, precision, power consumption and networking. Published exaflop figures can similarly depend on numerical precision and workload assumptions. Apparently precise comparisons between national or organisational computing capacity can therefore be misleading unless measurement methods are comparable.

Infrastructure announcements also require careful interpretation. A planned data centre, financing commitment, processor reservation or electricity agreement does not constitute operational AI capacity. The very large projects announced in Europe, the Middle East and elsewhere demonstrate the direction and scale of investment, but much of that infrastructure remained under construction or planned at 30 August 2026.

The distinction between **training and inference** is becoming increasingly important. Frontier training remains extraordinarily compute-intensive and concentrated among a relatively small number of organisations. Inference is more widely distributed and increasingly important to total demand because trained models may be used continuously by large populations. Improvements in smaller models, quantisation and specialised inference hardware can reduce the resources needed for particular tasks even while aggregate demand continues to increase.

Compute access is also geographically unequal. North America and China contain very large concentrations of infrastructure; East Asia is indispensable to the semiconductor and memory supply chain; Europe is expanding shared and sovereign computing capacity; India and Japan have developed substantial public-access initiatives; the Gulf states are investing in large new infrastructure; and smaller programmes are developing in Africa and Latin America. These developments show genuine globalisation, but not equal access.

Energy deserves similar proportionality. AI is becoming an important new source of electricity demand, particularly in locations containing large data-centre clusters. Yet data centres remain a minority of global electricity consumption. The IEA's central projection places them at around 3% of global electricity demand in 2030. Their significance arises partly because growth is rapid and geographically concentrated, which can create substantial local grid and infrastructure effects even when the global percentage remains comparatively modest.

As at 30 August 2026, AI Infrastructure & Compute is therefore both an established industrial system and one of the fastest-changing parts of the AI landscape. Specialised accelerators, high-bandwidth memory, high-speed networks, large cloud platforms and increasingly dense data centres already support AI at enormous scale. Public and sovereign computing

programmes are broadening access beyond a small group of technology companies, while new facilities are being developed across multiple regions.

The central current development is not simply that AI is using **more chips**. It is that computation is becoming an integrated infrastructure problem involving processors, memory, networking, software, electricity, cooling, capital and geography. How effectively those components can be assembled and operated increasingly determines where advanced AI can be developed and how economically it can be used.

Evidence Notes

1. **International Energy Agency — *Energy and AI*, 10 April 2025, and *Key Questions on Energy and AI*, 16 April 2026.** Principal independent evidence concerning data-centre electricity consumption, accelerated-server demand, regional distribution, energy sources and infrastructure bottlenecks. The IEA's figures for future electricity demand are scenario-based projections rather than forecasts of certain outcomes.
2. **Epoch AI — *Training compute of frontier AI models grows by 4–5x per year*, 28 May 2024; *Trends in AI Supercomputers*, 23 April 2025; and *Trends in Artificial Intelligence*, updated 5 February 2026.** Evidence concerning historical growth in training compute, AI-supercomputer scale, costs, ownership and compute trends. Estimates are reconstructed from incomplete public information and should not be treated as comprehensive measurements of undisclosed systems.
3. **Stanford Institute for Human-Centered AI — *AI Index Report 2025*.** Evidence concerning AI hardware performance, price-performance, energy efficiency, model-training costs and the substantial decline in inference cost for a fixed benchmark capability. Benchmark-equivalent inference cost should not be interpreted as the cost trajectory for every contemporary AI workload.
4. **MLCommons — *MLPerf Inference v6.0*, 1 April 2026, and *MLPerf Training v6.0*, 16 June 2026.** Independent industry benchmark evidence concerning contemporary training and inference systems, newer reasoning and mixture-of-experts workloads, and diversity of hardware and software configurations. Benchmark performance does not establish production cost, utilisation or effectiveness in every deployment.
5. **European Commission — *AI Factories and AI Gigafactories materials, including 30 July and 12 August 2026*.** Primary public-sector evidence concerning Europe's network of 19 AI Factories, 13 Antennas and procurement for up to seven larger AI Gigafactories. The Gigafactories were planned infrastructure at the reference date and should not be counted as existing operational capacity.
6. **China State Council Information Office / Ministry of Industry and Information Technology reporting, July–August 2026.** Primary/official evidence that China's reported intelligent computing capacity had reached 2,185 EFLOPS by the end of June 2026 and that more than 70 major computing corridors had been established around computing hubs. National capacity measures are not necessarily directly comparable with differently defined international measures.
7. **Government of India, Press Information Bureau, 30 May 2025.** Primary evidence that shared computing capacity under the IndiaAI Mission had crossed 34,000 GPUs. The announcement establishes procured/common capacity rather than utilisation or demonstrated AI outcomes.
8. **AIST / AIST Solutions — *ABCI 3.0 materials, 2024–2026*.** Primary Japanese evidence concerning the 6,128-H200 ABCI 3.0 public AI supercomputer, 75 PB storage system and approximately 6.22-exaflop theoretical half-precision performance. Peak theoretical performance should not be interpreted as sustained application performance.

9. **Brazil, Laboratório Nacional de Computação Científica / Ministry of Science, Technology and Innovation, 21 August 2026.** Primary evidence concerning development of a new national high-performance AI supercomputer. This was infrastructure under implementation at the reference date rather than established operating capacity.
 10. **U.S. Department of Commerce / UAE-US Framework on Advanced Technology Cooperation, 15 May 2025, and Abu Dhabi Stargate UAE announcement, 22 May 2025.** Primary government and project evidence concerning the initial 1GW data-centre component and planned 5GW UAE-US AI campus. Used as evidence of infrastructure development and geographic expansion, not as evidence that the announced full capacity was operational.
 11. **Cassava Technologies, 2025–2026.** Company evidence concerning NVIDIA-based AI-compute infrastructure in South Africa and planned expansion to other African markets. Used principally to establish the existence and direction of regional infrastructure development rather than continent-wide adoption or utilisation.
 12. **Google — Ironwood TPU, 2025, and eighth-generation TPU announcement, 22 April 2026.** Developer evidence illustrating increasing differentiation of infrastructure for training and inference. Performance and efficiency claims are supplier claims unless independently benchmarked.
 13. **NVIDIA — Vera Rubin production announcement, 31 May 2026.** Developer evidence concerning integrated rack- and pod-scale AI infrastructure, networking and global server-manufacturing participation. Comparative performance claims remain developer claims unless supported by independent benchmarks.
 14. **AMD — Instinct MI350 Series materials, 2025–2026.** Developer evidence concerning contemporary accelerator memory capacity, high-bandwidth memory and infrastructure configurations for training and inference. Supplier performance and cost comparisons should not be treated as independent evaluations.
 15. **Uptime Institute — *AI embraces liquid cooling, but enterprise IT is slow to follow*, 13 January 2025, and *Global Data Center Survey 2025*.** Independent industry evidence concerning rising rack densities, liquid cooling, power constraints, costs and the operational challenges created by AI workloads.
 16. **SK Hynix / Reuters reporting, 27 August 2026.** Recent evidence concerning continuing high-bandwidth-memory supply pressure and expansion of HBM manufacturing and packaging capacity. Company expectations about the duration of shortages remain forward-looking rather than established outcomes.
 17. **Australian Government, 15 July 2026.** Primary evidence concerning emerging requirements for large Australian data centres to address electricity supply, grid-connection costs, demand flexibility and water efficiency, illustrating the increasing interaction between AI infrastructure and wider physical infrastructure.
-

The Status Across the Technology Landscape

Viewed together, the fifteen Technology Landscape elements reveal an AI environment that is already technologically substantial, but cannot be described by a single level of capability or maturity.

Artificial intelligence is no longer adequately understood as one category of software or as a progression towards ever-larger models.

The current landscape contains mature specialised systems, general-purpose foundation models, increasingly capable reasoning systems, multimodal technologies, conversational interfaces, search and research systems, coding systems, synthetic-media technologies, agents, memory and personalisation, physical AI, edge intelligence, increasingly engineered data systems, world models and a large physical computing infrastructure.

These technologies do not develop independently.

They increasingly combine.

That convergence is one of the clearest characteristics of the present Status.

Established AI remains a major part of the landscape

Much current discussion concentrates on the newest frontier systems. The evidence across the fifteen elements shows that this can obscure a substantial established technological layer.

Specialised AI performing classification, forecasting, recommendation, anomaly detection, perception and other bounded functions is already deeply established. Edge inference has long operated in phones, vehicles, industrial equipment and other devices. Speech recognition, conventional text-to-speech, computer vision and many forms of industrial robotics have substantial operational histories.

Physical AI provides an especially useful illustration. Approximately 4.66 million industrial robots were operating worldwide in 2024, while specialised autonomous systems are already deployed in warehousing, mining, agriculture, healthcare and other structured environments.

These systems may attract less public attention than newer generative or humanoid systems, but technological significance should not be measured by novelty.

Some of the deepest current AI deployment remains specialised.

Foundation models have created a broader general-purpose layer

Generative AI and foundation models have added a markedly different technological layer.

Rather than designing a separate model for every narrow task, broadly pre-trained models can perform or be adapted across language, coding, analysis, knowledge work and other activities.

Foundation-model systems are now established and widely deployed rather than principally experimental.

This breadth has helped create an increasingly common technological foundation beneath other parts of the landscape.

Reasoning systems frequently build on foundation models. Multimodal systems extend them across images, audio and video. Conversational systems provide natural-language and spoken interfaces. Search systems combine generation with retrieval. Coding systems connect models with development environments. Agents connect models to tools and external systems.

The importance of foundation models therefore extends beyond the direct use of generative assistants.

They have become an enabling layer upon which many other current AI system forms are being constructed.

Yet generality does not mean uniform competence.

The same model can perform impressively on one task and unreliably on another. Fluent generation remains distinct from factual accuracy, and broad capability does not establish dependable effectiveness in every context.

AI is becoming increasingly multimodal and conversational

Another strong direction is movement beyond text as the principal interface to AI.

Multimodal systems increasingly combine text, images, audio, video, documents and other forms of information within the same task. The significance lies not simply in handling several media types independently, but in relating information across them.

Conversational and voice systems are simultaneously becoming more natural. Established speech-recognition and speech-generation technologies are being combined with general-purpose AI, while real-time speech-to-speech interaction, expressive voices, multilingual capabilities and interruption-aware conversation are increasingly available.

Image, Audio & Video AI demonstrates the same expansion through media production. Generative imagery, image editing and synthetic speech are already commercially well established, while music, video and integrated audiovisual generation occupy different and generally less uniform maturity levels.

The direction is towards AI systems that interact through more of the information forms people ordinarily use.

But multimodality should not be mistaken for human-like perception, and natural conversation should not be mistaken for human-like understanding.

Adding modalities expands what a system can receive and produce. It does not automatically establish that the information has been interpreted correctly.

Reasoning is improving, but remains uneven

Reasoning AI represents another material development.

Contemporary systems can undertake structured problem-solving, mathematical and logical processing, programming, inference, planning and verification rather than simply producing an immediate generated response. In bounded and verifiable domains, capability has advanced substantially.

The development is important because it changes what AI can contribute to difficult cognitive tasks.

Systems can increasingly search among possible solutions, use tools, test intermediate results and revise their approaches.

Yet reasoning remains a particularly clear example of why benchmark achievement and dependable general capability must be separated.

Performance can vary with the formulation of the problem, tools available, computational budget and evaluation method. Strong mathematical or coding performance does not establish equivalent reliability in ambiguous professional, social or real-world situations.

The present Status therefore supports neither the proposition that AI merely imitates reasoning nor the proposition that dependable general machine reasoning has been established.

Significant reasoning capability exists.

Its reliability remains conditional.

AI is beginning to move from producing outputs towards taking action

Perhaps the most consequential architectural transition is from AI that principally **answers** towards AI that can increasingly **act**.

AI Agents & Agentic Systems can plan intermediate steps, use tools, interact with software and data, and execute bounded actions towards objectives. Tool-using assistants and structured agentic workflows are increasingly established product capabilities.

Coding provides one of the clearest operational environments for this development.

AI coding systems have progressed beyond completion and code generation into debugging, testing, refactoring, documentation, repository-wide changes and bounded multi-stage activity involving files, terminals and development tools.

Search and research systems demonstrate another form. AI-enhanced search is already operating at consumer scale, while deeper research systems can formulate searches, inspect documents, retrieve multiple sources and construct extended syntheses.

These developments represent a genuine expansion of AI's operational role.

But answering, planning and acting remain different capability thresholds.

The longer the sequence of actions and the greater their consequences, the more reliability, permissions, security, monitoring, reversibility and accountability matter.

The current landscape therefore shows meaningful movement towards bounded AI agency without establishing dependable general digital autonomy.

Memory is adding continuity

Another shift concerns time.

Traditional interactions with AI were largely episodic: information entered for a particular interaction formed the immediate context and then ceased to be available unless supplied again.

Longer context windows, retrieval systems and persistent memory are changing that relationship.

AI Memory, Context & Personalisation increasingly allows systems to work with larger bodies of information, retain selected information across interactions and adapt behaviour using information about users, tasks and previous activity. Long-context processing and retrieval-supported context are already established capabilities, while persistent personal memory and more adaptive long-term memory remain less mature.

This matters beyond personal assistants.

Agents need state to conduct longer tasks. Research systems benefit from continuity across investigations. Enterprise systems need relevant organisational context. Personalised systems need mechanisms for retaining and retrieving appropriate information.

The technological frontier is consequently shifting from merely increasing how much information a system can receive towards deciding what should be retained, when it should be retrieved, when it should be updated and when it should be forgotten.

More memory is not inherently better memory.

Stored information can become inaccurate, irrelevant or inappropriate. Long context does not guarantee effective use of that context. Personalisation does not establish correctness.

Continuity is becoming an important AI capability, but dependable memory management remains a developing technological problem.

Model size is becoming a less complete description of capability

The Technology Landscape also demonstrates that development is not simply a race towards ever-larger general-purpose models.

Specialised, Small & Domain AI represents a substantial parallel direction.

Long-established specialised systems already operate at production scale. Smaller foundation models are becoming increasingly capable in language, coding, reasoning and multimodal tasks, while domain-adapted systems are being developed for medicine, engineering, cybersecurity, particular languages and other specialised environments.

The significant development is a widening separation between **model size** and **useful capability**.

A smaller system optimised for a particular task can sometimes provide sufficient or superior capability while requiring less compute, memory or energy. A domain-specific system may incorporate information and evaluation appropriate to its field.

But specialisation introduces boundaries.

A model that performs strongly within a defined domain does not thereby acquire broad competence. Benchmark superiority in one specialised task does not establish general superiority.

The current landscape therefore contains both increasingly capable general-purpose AI and increasingly capable specialised AI.

They are complementary technological directions rather than mutually exclusive alternatives.

AI computation is simultaneously centralising and distributing

Where AI operates is also changing.

At one end, Edge & On-Device AI moves computation towards phones, personal computers, vehicles, sensors, industrial equipment and other devices. Specialised local inference is already mature and widely deployed.

At the other end, frontier AI increasingly depends upon enormous concentrations of specialised computing infrastructure.

Cloud computing, servers, storage, data centres and accelerated computing are already mature industrial systems. What is changing rapidly is the scale and specialisation of infrastructure around demanding AI workloads: accelerator clusters, high-bandwidth memory, high-speed interconnects, specialised software, electricity and cooling.

These directions are not contradictory.

AI is becoming more distributed at the point of use while becoming more concentrated at the frontier of training and large-scale inference.

Hybrid arrangements increasingly connect the two.

A device can perform privacy-sensitive or latency-sensitive processing locally while using remote infrastructure for larger models, current information or computationally demanding tasks.

The technological landscape is consequently becoming an environment of distributed computation rather than a simple choice between device and cloud.

Infrastructure is becoming part of the AI capability question

The scale of contemporary AI also means that infrastructure can no longer be treated as an invisible background.

Advanced AI depends upon processors, memory, networking, servers, software, storage, data centres, electricity and cooling operating together.

The infrastructure element shows that the relevant development is no longer simply faster individual processors. AI compute is increasingly an integrated systems problem.

Infrastructure is also geographically uneven.

Public and sovereign computing programmes are broadening access in Europe, China, Japan, India and elsewhere, but the largest concentrations of advanced compute remain unevenly distributed.

This creates an important relationship between technological capability and physical resources.

A model architecture may be reproducible in principle while the computing infrastructure required to develop or operate it at frontier scale remains available to relatively few organisations.

At the same time, infrastructure announcements require the same Status discipline as model announcements.

Planned data centres are not operating capacity. Peak theoretical compute is not sustained application performance. Purchased accelerators do not establish utilisation. Infrastructure investment does not establish successful AI outcomes.

The physical foundations of AI are expanding rapidly, but announced capacity and usable capability remain different things.

Data is becoming increasingly engineered

Data remains another foundational part of the landscape, but its role is changing.

Contemporary AI development increasingly treats data not simply as a large raw resource to be accumulated, but as something selected, filtered, deduplicated, balanced, generated and engineered for particular learning objectives.

Large-scale data processing and curation are mature foundations of modern AI development. Synthetic data occupies a more varied position: conventional augmentation and simulation are established; synthetic post-training material is increasingly used; but synthetic replacement of high-quality real-world foundational data remains much less demonstrated.

Synthetic data can be particularly useful where real observations are scarce, expensive, sensitive or difficult to obtain. It can also generate targeted examples around known weaknesses and expose physical systems to simulated situations.

But artificial generation does not remove the data problem.

Synthetic material can reproduce errors, biases and omissions in the systems or data from which it is generated. More data does not necessarily produce better learning.

The central development is therefore not replacement of real data.

It is increasingly deliberate construction of the information from which AI learns. The underlying element describes this as a shift from data as principally a raw resource towards data as an engineered component of AI capability.

Simulation is becoming more closely connected to learning and action

World Models & Simulation provides another important connection between AI learning and environments.

Simulation itself is established and indispensable across many engineering and AI-development processes. The newer development is the emergence of learned models capable of predicting or generating aspects of environments and how those environments change.

These systems can generate scenarios, explore possible future states, provide simulated experience and increasingly connect that experience with planning, robotics and autonomous systems.

Maturity is highly differentiated.

Conventional physical simulation, robotics simulation and digital twins are established. Learned world models are technically well demonstrated and increasingly used in robotics, autonomous-driving development and agent training. More general interactive world generation and systems intended to support broad physical intelligence remain considerably earlier.

A central qualification is particularly important.

A generated environment that looks convincing is not necessarily an accurate model of reality.

Visual plausibility does not establish correct physics, causality or long-term consistency. Simulation remains complementary to real-world evidence, particularly where systems will eventually act in physical environments.

World models therefore represent a significant technological direction without demonstrating that AI has acquired a comprehensive internal model of the world.

Physical intelligence remains especially dependent on operating conditions

The movement from digital AI into physical action creates another major threshold.

Current Physical AI already demonstrates extensive economically useful capability. Industrial robots, autonomous mobile robots, specialised mining and agricultural systems and other machines operate at meaningful scale.

The frontier is moving towards greater adaptability.

Vision-language-action systems and other approaches attempt to allow machines to interpret natural-language instructions, perceive unfamiliar objects and adapt physical behaviour with less task-specific programming.

Operational humanoid pilots demonstrate genuine progress. But a robot performing a defined task in a managed manufacturing environment does not establish general human-equivalent physical capability.

Physical environments impose conditions that digital environments often do not.

Objects move unpredictably. Sensors are imperfect. Machines wear. People enter workspaces. Lighting, weather, terrain and physical contact change. Mistakes can damage equipment or injure people.

The maturity of Physical AI therefore depends heavily upon the structure of the environment and the range of conditions the system must handle.

The current Status is not one of physical AI waiting to begin.

It is one of extensive specialised physical automation progressing towards more adaptable machines, while general-purpose physical intelligence remains developing.

Convergence is becoming a defining characteristic

Perhaps the most important pattern across the fifteen elements is that their boundaries increasingly intersect.

A contemporary AI system may combine a foundation model, reasoning, multimodal perception, search, memory and agentic action.

A coding agent may reason about a repository, retrieve information, edit files, run tests and remember the state of a task.

A research system may search the web, inspect documents, maintain context and decide which source to investigate next.

A robot may combine cameras and other sensors with multimodal models, reasoning, world modelling, specialised local inference and physical control.

Synthetic data generated in simulation may train that robot.

The resulting model may operate locally on specialised edge hardware while larger-scale training occurs in remote compute infrastructure.

These are no longer isolated technological stories.

They increasingly form systems.

This convergence helps explain why capability can appear to advance rapidly: improvement in one layer can expand what several other layers can do.

But convergence also changes the reliability problem.

A system composed of individually capable components can still fail through the interaction between them.

Retrieval can provide the wrong information. Memory can supply outdated context. Reasoning can draw an incorrect conclusion. An agent can select an inappropriate action. A physical system can execute that action in the real world.

The movement towards integrated AI therefore increases the importance of evaluating complete systems rather than inferring system reliability from the capability of their most visible model.

Maturity is multidimensional

The fifteen elements make one conclusion particularly clear:

There is no meaningful single answer to the question, **“How mature is AI?”**

Different technologies are mature in different ways.

A technology can be scientifically advanced but operationally immature.

It can be commercially available but lightly adopted.

It can be widely used while still producing unreliable results.

It can perform exceptionally within a narrow domain while remaining weak outside it.

It can operate dependably in a structured physical environment while failing under less predictable conditions.

It can be mature as infrastructure while the applications it supports remain experimental.

This is why the distinctions between research capability, demonstration, product availability, pilot deployment, operational deployment, adoption at scale and demonstrated effectiveness matter throughout this publication.

They describe different evidential states.

The present landscape contains all of them simultaneously.

Geography shapes the technological landscape

AI is global, but its technological foundations and capabilities are not evenly distributed.

The elements identify meaningful development across the United States, China, Europe, Japan, South Korea, India, Southeast Asia, the Middle East, Africa, Latin America and Oceania.

Different regions contribute in different ways: model development, semiconductor manufacturing, robotics, devices, language systems, research, public computing infrastructure, data-centre development and application ecosystems.

Infrastructure provides an especially visible example. Large concentrations of advanced processors, energy, capital and supporting infrastructure remain geographically concentrated even as national and regional computing programmes broaden participation.

Data introduces another geographical dimension. Digital information does not represent languages, cultures and populations equally, and larger datasets do not automatically remove those imbalances.

The absence of highly visible frontier development in a region therefore should not be interpreted as absence of AI activity.

But neither should the existence of frontier systems in particular locations be interpreted as evidence that equivalent capability is globally accessible.

The Technology Landscape is global and interconnected, but materially uneven.

Capability is becoming increasingly systemic

Taken together, the elements suggest that the relevant unit of AI capability is increasingly larger than the model itself.

A capable model may depend upon retrieval, tools, memory, specialised data, software infrastructure, permissions and external systems.

An agent depends upon the environment in which it can act.

A robot depends upon sensors, actuators, control systems and operating conditions.

On-device AI depends upon hardware, memory and power constraints.

Frontier models depend upon large-scale compute, networking, electricity and cooling.

World models depend upon the fidelity of the environments and data from which they learn.

The practical capability of contemporary AI therefore increasingly arises from the interaction of multiple technological layers.

This has an important consequence for interpreting progress.

A better model can improve a system without making the complete system dependable.

More compute can increase available capability without determining whether that capability is useful.

More data can improve learning without guaranteeing better outcomes.

More memory can increase continuity while introducing stale or inappropriate information.

Greater autonomy can reduce human intervention while increasing the consequences of errors.

Technological progress is real.

Its implications depend increasingly upon the system in which that progress is embedded.

The overall Status

The fifteen elements do not support a picture of AI as either an immature technology still waiting for practical significance or a generally autonomous technology approaching uniform human-level capability.

The present position is more differentiated.

AI already contains deeply established technologies operating at substantial scale.

Foundation models and generative AI have created a broad general-purpose layer and are widely deployed.

Multimodal and conversational capabilities are widening how people and systems interact with AI.

Reasoning is materially improving in selected domains.

Search, coding and agentic systems increasingly undertake multi-stage activity.

Memory is adding continuity.

Specialised and smaller systems demonstrate that useful capability does not depend solely upon maximum model scale.

Physical AI already performs extensive specialised work while more adaptable machines continue to develop.

World models and synthetic data are expanding the environments and information from which AI can learn.

Edge computing is distributing intelligence towards devices while frontier development simultaneously depends upon increasingly concentrated infrastructure.

Across these developments, the direction is towards greater **integration**: more modalities, more context, more tools, more action, more continuity and closer connections between digital intelligence and physical systems.

But increasing integration should not be mistaken for uniformly dependable general intelligence.

The strongest demonstrations frequently remain ahead of routine operational capability. Product availability can exceed adoption. Adoption can exceed demonstrated effectiveness. Benchmark performance can exceed real-world reliability. Physical capability can depend heavily upon environmental constraints.

The overall Technology Landscape as at **30 August 2026** is therefore one of **substantial, rapidly developing and increasingly interconnected AI capability, but with pronounced differences in maturity, reliability, deployment and access**.

AI is no longer one technological category.

It is becoming an ecosystem of models, specialised systems, interfaces, memory, tools, agents, machines, data, simulation, devices and infrastructure that increasingly operate together.

Understanding that ecosystem—and distinguishing what is established from what is emerging—is central to understanding the present Status.

AIStatus — Continuing the Exploration

Artificial intelligence can be examined from several perspectives.

This publication considers AI through the **Technology Landscape** lens: the significant AI technologies, system forms and enabling developments that currently constitute the applied AI landscape.

The wider AISF AIStatus series also examines the current AI landscape through two complementary lenses.

Business Functions considers where and how AI is currently being applied across the functions and activities of organisations.

Industries & Sectors considers where and how AI is currently being applied across economic and societal sectors.

The three perspectives form the **AIStatus Tripod**:

Business Functions — Industries & Sectors — Technology Landscape

Each lens produces a standalone publication.

Readers do not need to read one AIStatus publication in order to understand another. The same development may legitimately appear through more than one lens because each asks a different question of the current landscape.

A reasoning system, for example, can be examined as a technology within this publication, as something being applied within Finance or Research & Knowledge Management through the Business Functions lens, and as something being applied within Financial Services or Science & Research through the Industries & Sectors lens.

The overlap is deliberate.

Together, the three lenses provide complementary ways of discovering what is happening in artificial intelligence at a defined point in time.

AIStatus forms one part of the wider AISF educational architecture.

AIStatus helps readers DISCOVER what is happening.

AIGuides help readers NAVIGATE AI and the questions surrounding its use.

AISF Insight Books help readers UNDERSTAND important AI developments and their implications in greater depth.

Each serves a different primary reader purpose while contributing to the wider AISF objective of supporting clearer understanding of artificial intelligence.

Future Status Updates & User Submissions

Artificial intelligence continues to change rapidly.

New capabilities emerge. Existing technologies improve. Research demonstrations become products, products move into operational deployment, and some anticipated developments may progress more slowly or differently than expected.

The technologies used to describe the landscape can also change.

Capabilities that are currently examined separately may converge sufficiently to require different treatment. Technologies that are presently emerging may become established. New system forms may become significant enough to warrant separate examination, while distinctions that are useful today may become less meaningful.

AIStatus is therefore intentionally time-bound.

Future editions may reassess the **AI Technology Landscape** using a later reference date.

The elements themselves are not intended to form an immutable classification of artificial intelligence. They may be retained, added, modified, combined, separated or removed according to the technologies, evidence and circumstances relevant to a future Status.

Earlier editions will remain records of the Status assessed at their stated reference dates.

AISF also welcomes information that may strengthen future Status assessments, including credible evidence of significant technological developments, deployments, evaluations, limitations or developments from countries and regions that may be less extensively represented in publicly available evidence.

Submissions do not guarantee inclusion.

Material considered for future editions remains subject to AISF's editorial and evidential standards.

Acknowledgements

AI Sourced Facts (AISF) acknowledges the researchers, universities, public institutions, international organisations, standards bodies, regulators, technology developers, infrastructure providers, companies, industry organisations and others whose publicly available work contributes to understanding the continuing development of artificial intelligence.

AIStatus — Technology Landscape depends upon evidence produced for many different purposes.

Technical papers and evaluations can demonstrate capability under defined conditions. Product documentation can establish the existence and availability of technologies. Deployment information can demonstrate operational use. Benchmarking organisations can provide comparative technical evidence. Public institutions can document infrastructure and policy developments. Independent research and reporting can help assess limitations, adoption and claims made by technology developers.

Particular recognition is due to organisations and researchers that publish not only technological advances but also evaluation results, limitations, methodological qualifications and evidence that helps distinguish technical demonstration from dependable real-world capability.

AISF also acknowledges the wider international community of researchers, engineers, developers, practitioners and users whose work continually expands the public evidence from which the current AI Technology Landscape can be assessed.

AISF Universal Disclaimer

This publication is provided for general educational and informational purposes only.

It is intended to support understanding of artificial intelligence and related developments and should not be interpreted as legal, regulatory, compliance, financial, investment, medical, technical, operational, strategic or other professional advice.

The publication does not recommend or endorse any particular artificial intelligence system, provider, product, platform, organisation, technology, approach or course of action.

Artificial intelligence technologies, services, capabilities, terminology, practices, laws, regulations and patterns of use can change rapidly. Their operation, suitability and adoption may differ between providers, models, versions, organisations, industries, jurisdictions and circumstances. Information and observations contained in this publication may therefore change over time and may not reflect subsequent developments.

Examples, studies, organisations, technologies and deployments are included to assist understanding of broader developments. Their inclusion does not constitute endorsement, recommendation or certification by AI Sourced Facts (AISF), and they should not be assumed to represent every organisation, jurisdiction, technology, sector or circumstance.

Evidence concerning artificial intelligence may be incomplete or subject to different definitions, samples, methodologies, assumptions and measures. Technical capability, availability, deployment, adoption and demonstrated effectiveness should not be assumed to be equivalent unless the underlying evidence supports that conclusion.

AISF seeks to maintain neutrality, clarity and appropriate recognition of uncertainty. Reasonable efforts are made to provide accurate and appropriately qualified information, but no representation or warranty is made as to its completeness, continuing accuracy or suitability for any specific purpose. No publication can eliminate all possibility of error, omission, subsequent change or differing interpretation.

Where decisions carry material consequences, responsibility for interpreting relevant information and making those decisions remains with the individual or organisation concerned and, where appropriate, their relevant qualified professionals or authorities.

Readers remain responsible for the use they make of the information contained in this publication.

About AI Sourced Facts (AISF) Pte. Ltd.

AISF is a Singapore-headquartered institution dedicated to structured reasoning, responsible AI navigation, and governance-informed adoption of artificial intelligence systems.

AISF operates with a capability-first, vendor-neutral posture. Its publications do not rank platforms, endorse providers, or promote specific technologies. Instead, AISF develops

structured frameworks that help individuals, professionals, and institutions reason clearly before integrating AI into operational, strategic, or educational environments.

AISF's work spans whitepapers, applied insight books, education instruments, governance architectures, and structured research initiatives. These outputs are informed by cross-system AI research methodologies and reflect globally observed usage patterns at the time of publication. Human accountability remains central across all AISF frameworks.

AISF does not provide regulatory, legal, financial, investment, or compliance advice. Its publications are designed to support structured thinking, proportionate governance, and disciplined evaluation of AI capabilities prior to deployment or reliance.

As artificial intelligence systems continue to evolve, AISF's focus remains constant: clarity before integration, governance proportionate to capability, and long-term institutional resilience in the age of AI.

AI Sourced Facts (AISF)

Truth, Clarity & Guidance in the Age of AI.

Publication & Governance Information

Title: AI Technology Landscape

Publication: AIStatus 3

Subtitle: The Technologies, System Forms & Enabling Developments Shaping Applied AI

Series: AISF AIStatus

Edition: Version 1.0 / 2026 Edition

Status Reference Date: 30 August 2026

Publication Date: September 2026

Published by: AI Sourced Facts (AISF) Pte. Ltd.

Website: www.aisourcedfacts.com

© 2026 AI Sourced Facts (AISF). All rights reserved.

No part of this publication may be reproduced, distributed or transmitted in any form or by any means except as permitted by applicable law or with the prior permission of the copyright holder.

AISF maintains editorial oversight of this publication.

AIStatus is a point-in-time publication. Future editions may reassess the subject at a later reference date where material developments affect the Status of the AI Technology Landscape.

Any future edition remains subject to AISF editorial standards and governance.

Back Page

AI Technology Landscape

Artificial intelligence is no longer one technology.

But what does the current AI landscape actually consist of — and how developed are its different parts?

AI Technology Landscape examines the Status of fifteen significant technologies, system forms and enabling developments, from foundation models, reasoning AI and agents to multimodal systems, physical AI, world models, synthetic data and the infrastructure that enables modern AI.

Across each element, the same seven questions examine what the technology is, what it can currently do, where it is being applied, how developed it is, what is driving its development and use, its present limitations and what readers should be aware of.

The publication distinguishes research capability from demonstrated capability, availability from deployment, and technological progress from dependable operation and adoption at scale.

It provides a global, neutral and evidence-informed snapshot of the position:

Status as at 30 August 2026

AI Sourced Facts (AISF)

Truth, Clarity & Guidance in the Age of AI.

Website: www.aisourcedfacts.com

Quick Access: AISF.tech
