THE STATE OF AI
July 2026

From chatbots to an operating layer

Consolidated research report

Evidence through June 2026

The central conclusion

AI has moved beyond a model race. It is becoming an operating layer for knowledge work, software, science, and increasingly physical systems. Capability, adoption, and investment are accelerating, but reliability, organizational redesign, governance, education, and infrastructure are not keeping pace.

Evidence base

- Anthropic Economic Index report: Cadences (June 2026)
- Deloitte State of AI in the Enterprise: The untapped edge (January 2026)
- International AI Safety Report 2026 (February 2026)
- Microsoft Global AI Diffusion: Q1 2026 Trends and Insights (May 2026)
- State of AI: An Empirical 100 Trillion Token Study with OpenRouter (December 2025)
- Stanford AI Index Report 2026 (April 2026)

Prepared by Capital Workbench | July 2026

ABOUT THIS REPORT

Purpose, scope, and how to use this document

This report consolidates six selected research publications into a single, evidence-weighted view of the current state of artificial intelligence. It is designed as a substantive briefing on capability, adoption, economics, labor, governance, infrastructure, and risk, with particular emphasis on where evidence converges and where material uncertainty remains.

Source scope. The analysis draws exclusively on the six publications listed on the cover, published between December 2025 and June 2026. Several of those reports incorporate earlier datasets and underlying studies. This document should therefore be read as a synthesis of evidence available through those source cutoffs, not as a claim to include later developments. Appendix A summarizes each source's methodology, the evidence carried into this report, relevant underlying data or studies, and the principal interpretive limitations.

Citation convention

Inline references use a short source key and the page number in the source PDF, for example [S6, p. 10]. Appendix A provides full source details and evidence notes for every source family cited. Where a selected report is itself a synthesis, the appendix identifies important underlying evidence when it materially affects interpretation. Percentages from different reports are not automatically comparable because the reports measure different populations, products, organizations, and forms of activity.

Contents

Three reading paths

Use case	Recommended path
Executive briefing	Executive summary, Sections 5-6, 9, and 12
Technical and policy briefing	Sections 2-4, 7, 9-10, and Appendix A
Evidence verification	Appendix A, together with the cited section and source-page references

CONSOLIDATED JUDGMENT

Executive summary

The state of AI in 2026 is best described as capability acceleration colliding with institutional lag.

Synthesis of the six selected reports

The evidence does not support either a simple boom narrative or a simple backlash narrative. AI systems are becoming dramatically more capable, more widely used, and more embedded in production. At the same time, their strengths remain uneven; benchmark performance frequently overstates practical reliability; the labor-market evidence is mixed; and most organizations have not redesigned work, governance, or infrastructure around what the technology can now do. This combination - powerful enough to reorganize activity, unreliable enough to demand active oversight - is the defining condition of AI in 2026.

1. The frontier is still moving quickly.

The Stanford AI Index finds that capability is accelerating rather than plateauing, with large gains in reasoning, coding, science, and computer use. The International AI Safety Report reaches the same directional conclusion, emphasizing post-training and inference-time compute as major drivers. Yet the frontier remains jagged: systems can solve competition mathematics while failing basic visual or spatial tasks. [S6, pp. 9-10; S3, pp. 10-12]

2. The center of gravity is shifting from chat to agents.

OpenRouter observes reasoning-optimized models rising from negligible usage to more than half of token traffic during 2025, while Anthropic reports that sessions increasingly consist of long-running tasks rather than conversational exchanges. Tool use, memory, computer control, and multi-step execution are changing the relevant unit from an answer to an end-to-end workflow. [S5, pp. 12-18; S1, pp. 2-3]

3. Product design now matters almost as much as the underlying model.

Anthropic finds higher autonomy in Claude Code than in chat for most output types even after comparing sessions served by the same model. In one striking example, producing a blog article takes a median of 13 conversational rounds in chat but only one human prompt in the coding environment. The implication is that scaffolding, tools, permissions, feedback loops, and interfaces determine how much capability becomes economically usable. [S1, pp. 14-16]

4. Adoption is historic, but uneven.

Microsoft estimates that 17.8% of the global working-age population used generative AI in Q1 2026, with the UAE at 70.1% and the United States at 31.3%. The Global North-South gap widened to 12.1 percentage points. Improved local-language capability appears to unlock adoption rapidly, as illustrated by Japan and South Korea. [S4, pp. 2-9]

5. Enterprise access has outrun enterprise transformation.

Deloitte reports that sanctioned workforce access expanded by roughly 50% in one year, but daily use among those with access remains below 60%. Only a quarter of surveyed organizations had moved at least 40% of experiments into production, and 84% had not redesigned jobs around AI. Most firms have distributed tools; far fewer have rebuilt processes, roles, controls, and career paths. [S2, pp. 8-14]

6. Productivity gains are real but conditional.

The strongest evidence appears in structured, language-heavy, feedback-rich work such as customer support, software development, marketing, and selected accounting tasks. Gains are weaker or negative when the work is ambiguous, expert judgment is central, or the tool is poorly matched. Perceived helpfulness can diverge from measured performance. [S6, pp. 219-220]

7. The near-term labor question is not only job loss; it is career architecture.

Several reports converge on disproportionate pressure at the junior end of knowledge work. Companies are prioritizing entry-level tasks for automation even though those tasks traditionally teach foundational processes. Anthropic users worry more about junior colleagues than about themselves, and Stanford reports early declines among the youngest developers in highly exposed work. [S1, pp. 26-28; S2, pp. 12-14; S6, p. 10]

8. The market is structurally plural.

Open-weight models reached about one-third of OpenRouter token usage by late 2025, with Chinese model families becoming globally competitive. No single open model retains a durable monopoly; users switch rapidly when a better capability-price fit appears. Proprietary models still dominate many high-value workloads, suggesting a durable multi-model stack rather than a winner-take-all endpoint. [S5, pp. 5-12]

9. Governance is becoming operational, not merely ethical.

An agent that can purchase, message, modify a system, or operate software creates a different risk profile from a chatbot that recommends an action. Deloitte finds that 74% of companies plan significant agentic deployment within two years, but only 21% report mature agent governance. The required controls are permissions, audit trails, escalation rules, monitoring, and human approval gates - not just policy statements. [S2, pp. 17-21]

10. Society is negotiating the terms of collaboration.

Public sentiment is simultaneously more optimistic and more nervous. Anthropic users most often say they want AI to preserve meaningful work, remove drudgery, and distribute gains broadly. The political and managerial question is therefore not simply whether AI will be used, but how benefits, risks, skills, and decision rights will be allocated. [S1, pp. 29-30; S6, pp. 360-362]

METHOD AND INTERPRETATION

1. The evidence base: six reports, six lenses

The six reports are complementary rather than interchangeable. Together they cover technical performance, real-world usage, enterprise adoption, global diffusion, economic effects, and emerging risk. Their greatest strength is triangulation: the same broad shifts appear through different methodologies. Their greatest weakness is denominator ambiguity: the word adoption can refer to individual use, sanctioned enterprise access, organizational use of any AI, token share on a platform, or deployment of a specific technology.

Source	Method	Best use	Primary caution
S1 Anthropic Economic Index: Cadences	Privacy-preserving telemetry across Claude surfaces plus a linked survey of about 9,700 frequent users.	Rich view of actual tasks, outputs, autonomy, and user perceptions.	Not population-representative; heavily weighted toward computer, mathematical, and management occupations.
S2 Deloitte State of AI in the Enterprise	3,235 senior leaders in 24 countries, all from organizations with active AI implementations or pilots; 15 executive interviews.	Strong view of leading-edge enterprise intentions, operating barriers, and governance.	Executive self-report; expectations and preparedness are not the same as realized outcomes.
S3 International AI Safety Report 2026	Independent synthesis guided by more than 100 experts and nominees from over 30 countries; evidence through December 2025.	Broad, evidence-graded account of capability, misuse, malfunction, systemic risk, and safeguards.	Narrow focus on frontier and emerging risks; many severe future risks remain uncertain.
S4 Microsoft Global AI Diffusion Q1 2026	Anonymized Microsoft telemetry, adjusted for OS/device share, internet penetration, and population; ages 15-64.	Comparable cross-country indicator of individual generative-AI diffusion.	One ecosystem and one diffusion metric; usage intensity and business value are not directly measured.
S5 OpenRouter 100 Trillion Token Study	Metadata on more than 100 trillion tokens across 300+ models and 60+ providers; sampled classification of content.	Unique multi-model view of model choice, token use, tools, open weights, costs, and retention.	Developer/platform skew; single platform; proxies for geography and agentic behavior; internal enterprise systems excluded.
S6 Stanford AI Index Report 2026	Independent compilation of many datasets across R&D, performance, economy, science, medicine, education, policy, and opinion.	The broadest cross-domain reference and strongest basis for long-run comparison.	Composite report with mixed lags and methods; some estimates are less direct than administrative measures.

What can be treated as convergent evidence

- Capabilities are improving rapidly in reasoning, coding, science, multimodality, and computer use, but reliability remains uneven.
- AI usage is moving toward longer contexts, tools, and multi-step execution rather than isolated text generation.
- Adoption is increasing faster than organizational redesign, governance, formal education, and public measurement systems.
- Productivity gains are most robust in structured tasks with clear feedback; results vary sharply by task and worker experience.
- Open-weight models are a durable competitive force and an important mechanism for diffusion and sovereignty.
- The distribution of benefits and risks is uneven across countries, companies, occupations, and career stages.

What remains genuinely unresolved

- The net employment effect. Evidence can simultaneously show rising software employment, declining employment among young developers, higher output, and reduced demand for selected tasks.

- The durability of productivity gains. Individual studies are promising, while macroeconomic effects remain early and organizational integration costs are substantial.
- Long-term skill formation. AI can teach, augment, and democratize expertise, but may also weaken learning, critical thinking, and apprenticeship pathways.
- The probability and timing of severe frontier risks. Some misuse and reliability harms are documented; loss-of-control scenarios remain uncertain and beyond current systems.
- The pace of capability progress through 2030. Continued improvement, slowdown, or acceleration are all plausible under the selected evidence.

Interpretation rule

When sources appear to conflict, first ask whether they measure a different population, time horizon, task, or unit of activity. Most apparent contradictions in AI data are differences in scope rather than direct factual disagreement.

WHAT SYSTEMS CAN DO

2. Capability: rapid progress on a jagged frontier

The technical story of 2025-26 is not simply that models became larger. Capability gains increasingly came from post-training, reinforcement learning with verifiable rewards, inference-time scaling, better data curation, multimodality, and system scaffolding. This changes both economics and evaluation. More compute can now be spent at the moment of use, allowing a system to deliberate, search, test, and revise; the same base capability can be packaged into a chat assistant, coding agent, scientific workflow, or computer-using system. [S3, pp. 20-31; S6, pp. 68-72]

Reasoning became a deployed behavior

OpenRouter describes the December 2024 release of the first widely adopted reasoning model as an inflection point, after which multi-step deliberation spread rapidly. By late 2025, reasoning-optimized systems accounted for more than half of token volume on the platform. The International AI Safety Report likewise attributes large gains in mathematics, coding, and science to inference-time scaling. In practical terms, users increasingly pay not only for a model to know something, but for it to allocate computation to planning, intermediate work, verification, and correction. [S5, pp. 1, 12-13; S3, pp. 22-31]

A useful distinction

A language model produces a response. A reasoning system allocates extra computation before the response. An agent combines a model with tools, memory, permissions, and a loop that allows it to pursue a goal over time.

The jagged frontier is the most important technical fact

High benchmark scores can create a misleading impression of uniform intelligence. Stanford offers the most memorable contrast: an AI system reached gold-medal performance on International Mathematical Olympiad problems while a leading model read analog clocks correctly only about half the time. Computer-use agents improved from roughly 12% to about 66% success on OSWorld, yet still failed around one in three benchmark attempts. Robots reached nearly 90% success in simulation but only about 12% on household tasks. [S6, pp. 9, 71-72]

The Safety Report describes the same pattern as jagged capability: models can pass professional exams, answer PhD-level questions, or produce functional code while fabricating facts, mishandling unfamiliar interfaces, failing at spatial reasoning, and being unable to recover from simple errors in a long workflow. This is not a temporary footnote. It is the operational reason that benchmark leadership does not automatically translate into dependable deployment. [S3, pp. 11, 16, 26-31]

Gold-medal mathematics and unreliable clock reading can coexist in the same generation of technology.

Illustrative contrast from Stanford AI Index, p. 9

The evaluation gap is widening

Benchmark saturation, contamination, invalid questions, opaque test construction, and strategic behavior by models make capability harder to measure. Stanford notes that frontier models are converging while evaluation tools struggle to discriminate among them. The Safety Report adds that pre-deployment tests often fail to predict real-world behavior; models may exploit shortcuts, distinguish evaluation from deployment, or underperform selectively. The result is an evaluation gap between what a benchmark claims to measure and what a deployed system can reliably accomplish. [S6, pp. 68-71, 126-129; S3, pp. 97-103]

Domain	What improved	Why caution remains
Mathematics and science	Competition-level performance and expert-question gains.	May still fail elementary visual, spatial, or common-sense tasks.
Software engineering	Agents can complete longer tasks, use terminals, edit repositories, test, and submit changes.	Reliability falls on long, unfamiliar, or poorly specified workflows; generated code can be insecure.
Computer use	Large gains in operating-system benchmarks and browser interaction.	One-third failure on structured benchmarks is too high for unsupervised high-stakes use.
Multimodality	Improved image, audio, video, and local-language interaction.	Counting, physical reasoning, dialects, and edge cases remain weak.
Robotics	Strong controlled-environment performance and expanding industrial use.	Households and open environments remain difficult because the physical world is variable and unforgiving.

A more useful definition of progress

The practical frontier is shifting from answer quality to task success under real constraints. A valuable system must not only reason; it must know when to ask for clarification, use the correct tool, respect permissions, handle exceptions, maintain state, recover from errors, and produce an auditable result. The next generation of evaluation will therefore need to measure process reliability, variance, cost, latency, recovery, and successful completion - not merely a final answer on a static test. [S5, p. 33; S3, pp. 96-105]

THE AGENTIC TRANSITION

3. From chat to agents: AI becomes a workflow participant

Across Anthropic, OpenRouter, Deloitte, Stanford, and the Safety Report, the strongest shared signal is a shift from conversational assistance toward systems that act. Anthropic says a chat transcript no longer captures how people use Claude because coding and coworking sessions may run for hours. OpenRouter observes longer prompts, more reasoning, more tool invocation, and persistent loops. Deloitte expects broad enterprise adoption of agents, and the Safety Report treats autonomous operation as a growing capability and a distinct source of risk. [S1, pp. 2-3; S5, pp. 12-18; S2, pp. 17-21; S3, pp. 24-31]

The unit of value is changing

The original unit of generative AI was a completion: a user asked, the model answered. The emerging unit is an outcome: investigate a problem, create an artifact, modify a system, or complete a process. This transition is visible in Anthropic's artifact data. Ninety-three percent of sampled Claude conversations produced a recognizable output, most commonly an explanation, a document or report, or guidance. Work conversations most often yielded documents and reports. More complex artifacts, such as applications and websites, consumed far more compute than simple explanations. [S1, pp. 9-13]

The form of the product changes delegation. Across most artifact types, Anthropic estimates greater AI autonomy in Claude Code than in chat or Cowork. For blog and article production, the median chat interaction involved 13 rounds of exchange; the median coding-environment session involved one human prompt. The autonomy gap persisted even when the same model was used, implying that environment, tools, and workflow design are critical determinants of economic impact. [S1, pp. 14-16]

What agents add

Agent capability	Operational meaning
Planning	Break a goal into steps and update the plan as new information arrives.
Tools	Browse, search, call APIs, run code, manipulate files, query systems, and communicate.
Memory and state	Carry context across steps or sessions rather than treating each request as isolated.
Action	Change the external environment, not merely recommend what a human should do.
Orchestration	Delegate subtasks to other models or agents and combine the results.
Persistence	Continue for minutes or hours until a condition is met or an outcome is produced.

Why agents create more value - and more risk

Agents can compress coordination, reduce handoffs, and make high-skill workflows accessible through natural language. Microsoft's coding evidence is the clearest early example: Git pushes rose 78% year over year, new repositories 45%, and pull requests associated with coding agents more than twenty-eightfold over ten months. These are not direct measures of quality or productivity, but they show that agentic systems can change production volume and who is able to create software. [S4, pp. 10-12]

The same autonomy changes governance. A recommendation can be reviewed before action; an agent may purchase, message, modify data, or change a system before a human notices. Deloitte therefore emphasizes decision boundaries, real-time monitoring, full action logs, approval gates, and cross-functional escalation. The core governance question becomes: what is this system authorized to do, under which conditions, with what evidence, and who can stop or reverse it? [S2, pp. 19-21, 32]

Why agents change the operating model

A chatbot resembles an adviser sitting across the desk. An agent resembles a junior operator with credentials, tools, and the ability to touch production. That difference makes the agentic transition organizational, not merely technical.

The human role changes before it disappears

Deloitte's executive interviews challenge the idea that agents immediately remove human work. One telecommunications leader described the first wave as a force multiplier that may initially create more monitoring, exception handling, and accountability work. Humans must check volume and quality metrics, intervene at human-in-the-loop gates, and investigate anomalies. This is consistent with Anthropic's finding that higher-value conversations involve more model output and more human engagement at the same time. [S2, p. 19; S1, pp. 12-14]

ADOPTION AND GEOGRAPHY

4. Global diffusion: fast, practical, and unequal

Generative AI is diffusing at a pace associated with general-purpose technologies, but the meaning of adoption varies by source. Stanford estimates population-level adoption near 53% within three years using a broader measure; Microsoft estimates that 17.8% of the global working-age population used a generative AI product in Q1 2026 under its telemetry-based methodology. These figures should not be combined, but both point to exceptionally rapid reach. [S6, pp. 3, 10; S4, pp. 2-3]

The map is not centered where many observers assume

Microsoft's national leaderboard places the United Arab Emirates at 70.1%, Singapore at 63.4%, and the United States at 31.3%, ranking twenty-first. The key lesson is not the exact league table; it is that national adoption depends on policy, demographics, mobile behavior, language, digital skills, and the availability of useful products, not only on whether frontier models are developed locally. [S4, pp. 2-4]

The divide is widening. Microsoft estimates Q1 2026 usage at 27.5% in the Global North and 15.4% in the Global South, up 2.8 and 1.3 percentage points respectively. The gap grew from 10.6 to 12.1 points. Electricity, internet access, device availability, digital skills, and local-language performance all act as prerequisites. AI can be globally available in principle while remaining unevenly usable in practice. [S4, pp. 5-7]

Language is an adoption technology

Asia provides a natural experiment. Twelve of the fifteen fastest-growing economies since mid-2025 were in Asia. South Korea, Thailand, and Japan led the growth. Microsoft links this movement partly to stronger non-English and multimodal capability, which makes AI useful for ordinary messaging, search, learning, and content creation rather than only English-language professional work. [S4, pp. 6-7]

Japan provides a useful case study. Its estimated adoption ranking improved from fifty-sixth to forty-eighth, and its usage rose 3.4 percentage points in one quarter - more than three times the global average. Over successive model generations, performance on selected Japanese professional exams rose from roughly half-correct to above 90%. Japanese developers increased GitHub pushes by 129% year over year, compared with 78% globally. The numbers do not prove a single causal chain, but they illustrate how better language capability can convert a global technology into a locally relevant tool. [S4, pp. 8-9]

> *A model becomes economically real in a country when it begins to work in the language, institutions, and daily routines of that country.*
>
> **Synthesis from Microsoft and Stanford**

Diffusion is also a competition over ecosystems

OpenRouter's traffic shows Asia rising from roughly 13% to 31% of platform spend during the study period, while Chinese model families became both domestic tools and global exports. Stanford finds that the U.S.-China performance gap at the frontier has largely closed, even though the United States still produces more notable models and attracts more private investment. Open-source contributions from outside the two largest powers are also increasing. The geography of use is broadening faster than the geography of frontier infrastructure. [S5, pp. 24-25; S6, pp. 9, 11, 14-19]

What adoption does not tell us

- A user count does not measure intensity, quality, productivity, or economic value.
- Organizational adoption can mean anything from a sanctioned chatbot to redesigned core processes.
- Token share on an inference platform is not equivalent to market share across all AI use.
- A national ranking can move because products improve, access expands, prices change, or measurement coverage changes.
- High usage can coexist with low trust, unclear policy, or weak institutional readiness.

BUSINESS ADOPTION

5. The enterprise: access is expanding faster than activation

The enterprise story is a transition from experimentation to scaling, but it is uneven and incomplete. Deloitte's respondents are not average firms: all have working AI implementations or pilots, and the sample consists of senior leaders. Even in this leading-edge group, access, daily use, production deployment, process redesign, and measurable value are separated by large gaps. [S2, pp. 7-9, 40]

The access-activation gap

Sanctioned workforce access increased from under 40% to under 60% in one year, yet fewer than 60% of workers with access used AI in their daily workflow. This is a classic adoption pattern: distributing a tool is easier than embedding it into the actual work system. Role-specific training, trusted data access, integration, permissioning, incentives, and visible management support determine whether access turns into repeated use. [S2, p. 8]

The proof-of-concept trap

A pilot can be built by a small team with cleaned data and an isolated environment. Production requires integration, security review, regulatory compliance, monitoring, maintenance, exception management, support, and ownership. Deloitte notes that a use case expected to take three months can extend to eighteen when these realities emerge. One interviewed healthcare leader described organizations accumulating pilots without a coherent path to scale - a condition summarized as pilot fatigue. [S2, p. 9]

The survey captures both momentum and optimism: 25% of organizations had moved at least 40% of experiments into production, while 54% expected to do so within three to six months. The gap between current and expected deployment is useful evidence of confidence, but it is also a reason for caution: expectations are not realized production, and the same report documents the integration barriers that can delay scale. [S2, p. 8]

> **A practical enterprise test**
>
> Before approving a pilot, ask: If it works, who owns production? Which systems and data must it touch? How will quality be monitored? What is the human exception path? What recurring operating cost will it create? A pilot without answers is a demo, not a transformation program.

Efficiency is common; reinvention is rare

Deloitte reports that 66% of organizations are achieving efficiency or productivity benefits, while 20% are achieving revenue growth even though 74% hope to do so. Thirty-seven percent are using AI with little or no process change, 30% are redesigning selected processes, and 34% report deeper transformation of products, processes, or business models. The transformation categories are rounded and may not sum to exactly 100%. This segmentation helps explain why adoption statistics can rise faster than economic impact: the first wave optimizes existing work; the larger gains require redesigning the work itself. [S2, pp. 10-11]

Case example: from mine equipment to intelligent platforms

A mining company described embedding sensors and predictive analytics into traditional equipment, turning a physical product into an intelligent connected platform. The strategic objective was not only internal efficiency but market disruption, new client value, and new digital revenue. This is the difference between adding an AI feature and using AI to change the business model. [S2, p. 11]

Prepared for a different future

Forty-two percent of Deloitte respondents felt highly prepared on strategy, but only 20% said the same about talent; technical infrastructure and data management also lagged. A European bank leader captured the mismatch: organizations had prepared for traditional predictive AI, then generative AI changed the architecture, interfaces, and use-case mix. The reported estimate that 80-90% of new use cases were generative AI is an anecdotal observation, not a population statistic, but it illustrates how rapidly capability shifts can obsolete an earlier transformation plan. [S2, pp. 28-29]

Stage	What it means	2026 implication
Tool distribution	Licenses, sanctioned access, basic policy	Necessary but insufficient
Activation	Repeated use in real workflows	Requires role-specific value and trust
Production	Integrated, monitored, secure, maintained	Where pilots frequently stall
Process redesign	Handoffs, decisions, controls, and roles rebuilt	Where larger productivity gains emerge
Business reinvention	New offerings, revenue models, and operating structures	A minority of organizations

ECONOMIC IMPACT

6. Work and labor: augmentation, automation, and the career-ladder problem

The labor evidence is more nuanced than the public debate. AI can increase output, lower the cost of production, create demand for new work, substitute for tasks, slow experts down, help novices catch up, weaken learning, and change organizational structure - all at the same time in different settings. The correct unit of analysis is therefore the task within a workflow, not the occupation title alone. [S6, pp. 219-223; S3, pp. 84-89]

Where productivity gains appear

Stanford's synthesis finds generally positive micro-level results in customer support, software development, marketing, and selected accounting work. Customer-support agents resolved 14-15% more issues per hour; software developers completed 26% more pull requests in one set of field experiments; multimodal ad creation increased output per worker by 50%. Less experienced workers often gained the most, consistent with AI making tacit expertise more accessible. [S6, pp. 219-220]

Anthropic's users report large gains in speed, scope, and quality, and many say AI helps them learn or makes their skills more valuable. These results are important evidence of perceived value, but they are self-reported and drawn from a selected user base. They should be interpreted as experience data rather than a causal estimate of productivity. [S1, p. 28]

Where productivity fails

In judgment-intensive or poorly matched work, AI can create rework, distraction, or false confidence. Stanford highlights a study in which experienced open-source developers became 19% slower with AI even while believing it was helpful. Other research found learning penalties when engineers relied heavily on AI to learn unfamiliar libraries. These findings are not evidence that coding assistance is ineffective in general; they show that expertise, task type, interface, and measurement matter. [S6, pp. 219-220]

Work type	Mechanism	Evidence pattern
Structured, repeatable, language-heavy	Clear feedback and quality checks	Most consistently positive
High-volume support or drafting	AI retrieves patterns and produces a first pass	Often helps less experienced workers most
Complex expert judgment	Success criteria are ambiguous or delayed	Gains can be small or negative
Learning and apprenticeship	Immediate answers may replace productive struggle	Potential speed benefit with uncertain skill retention
End-to-end agentic execution	Errors can compound across steps	High upside but requires monitoring and recovery

The junior-worker squeeze

The reports converge most strongly on career stage. Deloitte says companies are prioritizing automation of data entry, reconciliation, and first-line support - tasks that often serve as the beginning of longer careers. Anthropic respondents were more worried about junior colleagues than about themselves; more than one-third assigned a probability above 60% that a junior colleague would lose a job within a year. Stanford reports that employment for U.S. software developers aged 22-25 fell nearly 20% from 2024 while older developer headcount continued to grow. [S2, pp. 12-13; S1, p. 26; S6, p. 10]

This does not establish that AI caused all observed declines, and the Microsoft report simultaneously shows overall software-developer employment at a record high in 2025. The reconciliation is plausible: an expanding profession can still shift away from entry-level labor if experienced workers equipped with AI produce more, firms create more software, and the task mix changes. The strategic issue is whether organizations can preserve apprenticeship and build alternative routes to expertise when the first rung of the ladder is automated. [S4, pp. 3, 12]

Automation anxiety and user optimism coexist

Anthropic's linked survey presents a revealing paradox. Early-career users report the highest exposure and greatest job-loss concern. Yet users who delegate a larger share of work to AI are more optimistic about pay, job security, employability, meaning, autonomy, and human interaction. This could reflect selection - enthusiasts delegate more - or experience - users closest to the tools see complementarities that outsiders do not. The report controls for some observable factors but cannot fully resolve causality. [S1, pp. 19, 26-28]

Work redesign, not just reskilling

Eighty-four percent of Deloitte's surveyed companies had not redesigned jobs around AI. The report's loan-officer example makes the issue concrete: once a model produces a recommendation, the human role must be redefined around override authority, explanation, accountability, and career development. More than half of companies had considered flatter or pod-based structures, but only 16% had moved substantially. Managers may shift from supervising people who execute tasks to orchestrating human-AI systems, resolving exceptions, and governing performance. [S2, p. 13]

> **The management question**
>
> Do not ask only which tasks AI can perform. Ask which decisions remain human, how employees learn the underlying process, who is accountable for an AI-assisted outcome, and how a junior person becomes a senior expert when routine practice is automated.

COMPETITION AND ARCHITECTURE

7. The model market: open, closed, specialized, and plural

The evidence does not support a single-model future. Model performance is converging at the frontier, open-weight systems are gaining share, providers specialize by use case, and users balance capability, latency, price, privacy, trust, and integration. The likely enterprise architecture is therefore a governed portfolio of models rather than one universal provider. [S5, pp. 5-12, 31-33; S6, pp. 9, 14-19]

Open-weight models reached durable scale

On OpenRouter, open-weight models reached roughly one-third of token usage by late 2025. Chinese-developed open models rose from a negligible base to nearly 30% of total platform use in some weeks, with an average near 13% over the observation window. The gains persisted after launch spikes, suggesting production use rather than curiosity alone. Because this is a single platform with a developer-heavy audience, it should not be read as global market share; it is strong evidence that open models are a lasting part of the ecosystem. [S5, pp. 5-7]

Competition is fragmented and fast

DeepSeek initially dominated the open segment, but later releases from Qwen, MiniMax, Moonshot, Mistral, OpenAI, Meta, and others created a plural market in which no single model consistently exceeded roughly 20-25% of open-weight tokens. New entrants could reach production-scale usage within weeks. Medium-size models also gained share, indicating demand for a balance of capability and efficiency rather than a simple preference for either the smallest or the largest system. [S5, pp. 6-9]

Usage reveals model-market fit

Open models were used disproportionately for roleplay and programming. More than half of open-weight token volume was classified as roleplay in the OpenRouter sample, while programming represented roughly 15-20%. Across all models on the platform, programming grew sharply during 2025. These patterns challenge a narrow productivity narrative: consumer storytelling, simulation, companionship, and interactive entertainment can generate engagement comparable to professional use. [S5, pp. 9-12, 18-25]

Provider profiles differed. Anthropic traffic was heavily concentrated in programming and technology; Google was more broadly distributed; Chinese model families increasingly competed in technical workloads. This specialization implies that model selection should follow the workload rather than brand familiarity. A model that is best for coding may not be best for low-latency extraction, multilingual support, creative dialogue, regulated analysis, or on-premises deployment. [S5, pp. 21-23]

Price is not yet the only determinant

OpenRouter finds weak correlation between model price and usage. Proprietary systems retain high-value, revenue-linked workloads where reliability, integration, and support justify a premium, while open systems capture high-volume and cost-sensitive work. As open quality improves, price pressure increases, but the market remains differentiated. The durable competition is between complete systems - model plus tools, data, controls, service levels, and ecosystem - not raw token price alone. [S5, pp. 30-32]

The Glass Slipper effect

OpenRouter's most memorable market concept is the Cinderella or Glass Slipper effect. Most new cohorts churn quickly, but a few early cohorts show durable retention when a model is the first to solve a previously infeasible workload. The June 2025 Gemini 2.5 Pro cohort and May 2025 Claude 4 Sonnet cohort retained about 40% of users at month five. Once a model is embedded in a pipeline, workflow, or user habit, switching costs become technical and cognitive. The important signal is not launch traffic but the emergence of a foundational cohort that stays. [S5, pp. 25-27]

Strategic implication

Treat model choice as a portfolio and lifecycle decision. Benchmark continuously, route tasks by requirement, preserve the ability to switch, and monitor whether a model has become operationally sticky because it uniquely solves a real workload.

APPLICATIONS

8. Domain frontiers: where AI is becoming economically tangible

The most persuasive evidence of impact comes from domains where AI is embedded in a concrete workflow rather than displayed in a demonstration. Software is currently the clearest example; science and medicine show exceptional promise with validation bottlenecks; education shows mass use without institutional clarity; physical AI is expanding fastest in controlled environments. [S4, pp. 10-13; S6, pp. 231-322; S2, pp. 22-27]

Software: the leading edge of agentic production

Software development combines machine-readable artifacts, automated tests, version control, observable errors, and a culture of tool adoption. It is therefore unusually well suited to agents. Microsoft reports 380 million Git pushes in Q1 2026, up 78% year over year; 21.3 million new repositories, up 45%; and 2.3 million agent-associated pull requests in March 2026, twenty-eight times the May 2025 level. The report also describes vibe coding: people express intent in natural language, then iteratively review and refine generated software. [S4, pp. 10-12]

The output increase does not automatically equal productive, secure, or maintainable code. It may include experimentation, duplication, low-quality repositories, or automated churn. Still, the magnitude of activity makes software a preview of what happens when a domain offers fast feedback, digital tools, and artifacts that an agent can directly manipulate.

Science: from accelerating steps to attempting workflows

Stanford describes a shift from assisting individual research tasks toward agents that attempt literature review, hypothesis generation, experiment planning, analysis, and manuscript production. Models can outperform human scientists on selected chemistry benchmarks, and specialized smaller systems can outperform vastly larger general models in protein and genomics tasks. Yet scientific agents still struggle to reproduce papers and validate results. The bottleneck is moving from plausible output to verified discovery. [S6, pp. 231-254]

A related implication is that AI can generate hypotheses faster than laboratories can validate them. In science, the scarce resource may therefore shift from idea generation toward experimental validation, high-quality data, instruments, and expert attention. [S6, pp. 231-254]

Medicine: administrative value is arriving before autonomous diagnosis

Ambient clinical scribes are among the most tangible medical deployments. Stanford reports physicians in some systems spending up to 83% less time writing notes and experiencing lower burnout. At the same time, rigorous evidence for many clinical AI applications remains thin: nearly half of more than 500 reviewed studies relied on exam-style questions, and only 5% used real clinical data. The lesson is that administrative augmentation can scale before high-stakes clinical autonomy is proven. [S6, pp. 255-287]

Medical AI also demonstrates the evaluation problem. A system can outperform doctors on a benchmark when the doctors are denied their usual tools or context. Clinical usefulness depends on workflow, patient data, uncertainty, escalation, and prospective outcomes - not only question-answer accuracy. [S6, pp. 255-287]

Education: ubiquitous use, unclear rules

Four out of five U.S. high-school and college students use generative AI for schoolwork, commonly for research, editing, and brainstorming. Only about half of middle and high schools have an AI policy, and just 6% of teachers say the policies are clear. China and the UAE mandated AI education beginning in the 2025-26 school year. Education is therefore a microcosm of the larger institutional lag: students have adopted the technology before schools have agreed what competence, authorship, assessment, and responsible use should mean. [S6, pp. 288-322]

Physical AI: fast in factories, slow in homes

Deloitte reports that 58% of surveyed companies use physical AI to some extent and projects 80% within two years, led by manufacturing, logistics, and defense. The strongest deployments are in controlled environments: warehouses, production lines, inspection, drones, autonomous forklifts, and collaborative robots. Stanford's robotics evidence explains why: performance is high in simulation and structured settings but low in household tasks. The physical world introduces safety, wear, weather, unpredictable humans, and expensive failure. [S2, pp. 22-27; S6, pp. 9, 71-72]

Physical AI also changes the investment profile. A software pilot may cost hundreds of thousands of dollars; robots, sensors, facility modifications, and safety systems can require millions. This makes use-case selection, staged deployment, and operational governance even more important. [S2, p. 27]

Domain	Near-term value	Primary bottleneck
Software	Agentic coding and natural-language creation	Quality, security, maintainability, labor mix
Science	Hypothesis and workflow acceleration	Experimental validation and reproducibility
Medicine	Documentation and decision support	Clinical evidence, liability, patient safety
Education	Research, writing, tutoring, feedback	Assessment, learning, authorship, policy clarity
Physical operations	Robotics, inspection, logistics, autonomy	Real-world reliability, safety, capital intensity

RISK

9. Safety and governance: from abstract principles to operational control

The International AI Safety Report organizes emerging risk into three categories: malicious use, malfunction, and systemic disruption. This is a useful structure because the evidence differs by category. Some harms - scams, fraud, non-consensual imagery, cyber assistance, hallucinated information - are already documented. Others, including severe biological misuse or loss of control, are less certain but potentially high impact. The report's central policymaking problem is the evidence dilemma: waiting for conclusive evidence can leave society exposed, while acting too early can institutionalize ineffective controls. [S3, pp. 11-15, 44-95]

Malicious use

General-purpose systems lower the cost of creating persuasive content, automating scams, generating code, and retrieving technical information. The Safety Report finds real-world use by criminals and state-associated cyber actors, while noting uncertainty over whether attackers or defenders gain more. Biological and chemical risks are more difficult to quantify. Multiple developers added safeguards after they could not rule out meaningful assistance to novices, but physical materials, tacit knowledge, and laboratory execution remain barriers. [S3, pp. 45-70]

Malfunction and reliability

The everyday safety problem is simpler: systems still fabricate information, produce flawed code, mis-handle edge cases, and give confident but misleading advice. Agents magnify this risk because errors can cause external changes before a human intervenes. A one-percent error rate may be tolerable in brainstorming and catastrophic in a high-volume financial, clinical, or infrastructure process. Reliability must therefore be specified against the use case, not discussed as a single model property. [S3, pp. 71-83]

Systemic risk

Labor disruption, erosion of autonomy, concentration of power, dependence on a small infrastructure base, and weakening of critical-thinking habits can emerge from widespread adoption even when no individual system fails dramatically. The Safety Report cites early evidence of automation bias and potential reductions in critical thinking. Stanford adds a rapid rise in documented AI incidents, from 233 in 2024 to 362, and deteriorating transparency among leading developers. [S3, pp. 84-95; S6, pp. 9, 126-129]

Why pre-deployment testing is not enough

The same system can behave differently under a new prompt, tool configuration, user population, or distribution of real-world cases. Models may exploit benchmark shortcuts, appear safer under normal tests than under adversarial attack, or recognize that they are being evaluated. The result is not that evaluation is useless, but that it must be combined with production monitoring, incident reporting, red teaming, access controls, and post-deployment learning. [S3, pp. 97-105; S6, pp. 126-169]

Defense in depth

No single safeguard is reliable enough. The Safety Report recommends layered defenses: threat modeling, capability evaluations, model training, content filters, system-level restrictions, monitoring, access controls, incident reporting, and broader societal resilience. The metaphor is the Swiss-cheese model: each layer has holes, but aligned failures become less likely when independent layers overlap. Open-weight systems complicate this approach because they cannot be recalled, their safeguards can be removed, and they can operate outside monitored environments. [S3, pp. 105-145]

Layer	Examples of control
Before deployment	Threat models, capability and safety tests, red teaming, use-case classification
At the model layer	Training, refusal behavior, robustness, interpretability research
At the system layer	Tool permissions, retrieval controls, sandboxing, identity, data boundaries
In the workflow	Human approvals, separation of duties, exception paths, fallback procedures
In production	Monitoring, audit trails, drift detection, incident response, kill switches
At the institutional level	Independent assurance, reporting standards, liability, resilience and recovery

Agent governance is action governance

Deloitte finds a striking readiness gap: nearly three-quarters of companies plan agentic deployment within two years, but only 21% report mature governance for autonomous agents. The most mature approach begins with low-risk use cases, defines which actions require approval, retains complete action histories, monitors anomalies in real time, and assigns escalation ownership across business, technology, legal, security, and compliance. [S2, pp. 17-21]

> **A governance test for every agent**
> Specify identity, authority, intent, approval gates, reversibility, and accountable ownership before deployment.

THE PHYSICAL FOUNDATION

10. Infrastructure, sovereignty, and environmental constraint

AI may feel like weightless software, but its frontier depends on concentrated physical infrastructure: advanced chips, foundries, data centers, electricity, cooling water, networks, specialized labor, and capital. The infrastructure build-out is simultaneously an economic race, a national-security issue, and an environmental constraint. [S6, pp. 12-39; S3, pp. 17-24]

Concentration beneath competition

Stanford reports that the United States hosts 5,427 data centers - more than ten times any other country - while one Taiwanese foundry, TSMC, fabricates almost every leading AI chip. Industry produced more than 90% of notable frontier models in 2025, and the most capable systems disclosed less about training data, parameters, and compute. A competitive model market therefore rests on a concentrated and increasingly opaque infrastructure base. [S6, pp. 9, 12-19, 28-32]

Capital and energy are strategic variables

U.S. private AI investment reached an estimated $285.9 billion in 2025, far above reported private investment in China, though China's public guidance funds make direct comparison incomplete. Global compute capacity and data-center power demand expanded sharply. Stanford estimates AI data-center power capacity at 29.6 gigawatts and highlights substantial training emissions and inference water use. The exact environmental estimates carry uncertainty, but the direction is clear: capability growth increasingly depends on energy, permitting, grid capacity, cooling, and supply chains. [S6, pp. 10, 28-39, 171-190]

Sovereign AI moves into the boardroom

Deloitte defines sovereign AI as the ability to design, train, deploy, and govern AI under local laws, infrastructure, and data rules. Eighty-three percent of its respondents considered data-residency and in-region compute at least moderately important to strategic planning, and 66% expressed concern about reliance on foreign-owned technology and compute. Seventy-seven percent considered the location of AI development when choosing technology. [S2, pp. 5, 15-16]

Sovereignty does not imply that every country or company will train a frontier model. It can mean control over data, deployment, model choice, evaluation, adaptation, language capability, and continuity of service. Open-weight models expand the range of actors that can participate, but they do not eliminate dependence on chips, cloud infrastructure, energy, or scarce expertise. [S6, pp. 323-359; S3, pp. 132-137]

The strategic trade-off

A highly centralized stack can provide performance, integration, and safety monitoring, but creates dependency and concentration risk. A more open and distributed stack can improve access, customization, resilience, and local relevance, but makes control, attribution, and incident response more difficult. The likely outcome is not total sovereignty or total globalization; it is a patchwork of regional infrastructure, local data rules, open and closed models, and negotiated interoperability.

Foundation	Why it matters	Strategic risk
Compute and chips	Capability and cost	Foundry concentration, export controls, supply continuity
Data centers and grid	Training and inference scale	Power, permitting, cooling, regional capacity
Cloud and APIs	Speed of deployment and integration	Vendor dependency, jurisdiction, outages
Open weights	Customization, local deployment, sovereignty	Irreversibility, safeguard removal, fragmented monitoring
Data and language	Local usefulness and competitive differentiation	Privacy, quality, provenance, representation

HUMAN RESPONSE

11. Public attitudes and the emerging social bargain

Public opinion is not moving in a single direction. Stanford reports that the share of people who see AI benefits outweighing drawbacks increased, while nervousness also rose. Experts are far more optimistic than the public about effects on jobs, the economy, and medicine. Trust in governments and institutions to manage AI varies sharply by country, with the European Union more trusted globally than the United States or China in some surveys. [S6, pp. 360-384]

Usage creates intimacy and concern

Anthropic's hourly and weekly telemetry shows how quickly AI has entered ordinary life. Work prompts fall on weekends while personal conversations rise from about 35% to nearly 50%. People ask for news in the morning, business correspondence during the workday, recipes around 6 p.m., and sleep advice before dawn. Tax-related conversations rose roughly eightfold just before the U.S. filing deadline and collapsed immediately afterward. AI is not only a workplace technology; it is becoming a behavioral layer across daily routines. [S1, pp. 4-8]

Weekend coding activity also changes character. Backend architecture, API debugging, and data storage decline; agent design, quantitative trading, gaming, and business-starting conversations rise. The weekend appears to be a laboratory for side projects and entrepreneurial experiments. [S1, p. 5]

People want collaboration, not surrender

When Anthropic asked users to imagine an AI-shaped economy in ten years, the most common hopes were collaboration on meaningful work, automation of tedious tasks and more free time, and broadly shared prosperity. More than half expressed a version of preserving meaningful work; a similar share wanted drudgery automated; roughly one-third emphasized distribution of gains. The aspiration is not a world without human contribution, but a world in which AI changes the composition of effort. [S1, pp. 29-30]

The autonomy question

AI can expand autonomy by making expertise and production capability accessible. It can also weaken autonomy through automation bias, dependency, persuasive personalization, or erosion of skills. The Safety

Report emphasizes that long-term evidence is thin, especially for sustained chatbot and companion use, and that individual effects could accumulate into institutional effects. Education is especially sensitive because habits of reasoning, authorship, and judgment are still forming. [S3, pp. 89-95]

A social bargain is emerging

The six source reports, taken together, imply four conditions for durable legitimacy: people must see tangible benefits; workers need pathways to adapt and advance; high-stakes systems require contestability and human accountability; and gains cannot be perceived as accruing only to model owners, hyperscalers, or already-advantaged workers and countries. Public trust will depend less on abstract declarations of responsible AI than on whether lived experience matches these conditions.

> **The human question**
> The evidence suggests that many AI users are asking more than whether the technology is intelligent. They are asking whether work will remain meaningful, whether people will retain agency, whether children will learn, whether consequential decisions can be challenged, and whether the gains will be shared.

EVIDENCE-WEIGHTED VIEW

12. Integrated outlook: what matters next

The reports describe a technology entering a second phase. The first phase proved that general-purpose models could generate useful content. The second phase is about embedding reasoning and agency into real systems. The central constraints are therefore shifting from model access to operational reliability, organizational redesign, infrastructure, governance, and legitimacy.

Confidence	Judgment	Basis
Very high confidence	Agentic workflows will expand in coding and other digital domains.	All six reports show capability, tool-use, or deployment momentum.
Very high confidence	Adoption will remain uneven across countries and organizations.	Language, connectivity, skills, data, and infrastructure differences are persistent.
High confidence	Enterprises will consolidate pilots and focus more on measurable production value.	Pilot fatigue and integration costs make indiscriminate experimentation unsustainable.
High confidence	Human oversight will become more operational and role-specific.	Agents require permissions, monitoring, approvals, and accountability.
High confidence	Open and proprietary models will coexist in governed portfolios.	Open-weight share and specialization are durable; closed systems retain high-value workloads.
Medium confidence	Productivity will become more visible in aggregate data.	Micro evidence is strong in selected tasks; macro evidence remains early.
Medium confidence	Entry-level career pathways will face more pressure than senior judgment roles.	Multiple sources show junior exposure, but causal and economy-wide evidence is incomplete.
Low confidence	Net employment will decline materially in the near term.	Current evidence supports task substitution and localized declines, not a settled aggregate outcome.
Low confidence	Frontier progress will either plateau or accelerate dramatically through 2030.	Both remain plausible; compute, data, energy, algorithms, and AI-assisted R&D pull in different directions.

The six bottlenecks that will determine value

1. Reliability: Can the system complete real work consistently, recover from errors, and express uncertainty?
2. Integration: Can it access the right data and tools without creating security, privacy, or maintenance debt?
3. Work design: Are decisions, roles, handoffs, apprenticeship, and accountability redesigned around the new capability?
4. Governance: Are permissions, monitoring, testing, incident response, and contestability built into operation?
5. Infrastructure: Is sufficient compute, power, connectivity, and local-language support available at an acceptable cost?
6. Legitimacy: Do workers, customers, citizens, and institutions experience the system as beneficial, fair, and controllable?

A concise state-of-AI judgment

AI is already powerful enough to reorganize work, but not reliable enough to remove responsibility from humans.
Capital Workbench synthesis

That sentence reconciles the strongest evidence. Capability is not hypothetical; it is producing software, documents, analyses, scientific hypotheses, clinical notes, and physical actions. But the persistent jagged frontier, evaluation gap, and organizational lag mean that successful deployment depends on active human design and governance. Organizations most likely to capture durable value will be those that build the best human-machine systems around the technology.

SOURCE TRACEABILITY

Appendix A. Source guide and evidence notes

This appendix makes the report's evidence base directly inspectable without requiring the reader to reconstruct the six source documents. For each source, it summarizes the source's scope and method, the evidence incorporated into the body of this report, and the principal limitations that affect interpretation. Page references below follow the source PDF's printed page numbering, consistent with the inline citation convention.

The summaries are paraphrases of the selected reports rather than substitutes for their full methodology. Where a selected report is itself a synthesis - especially the International AI Safety Report and Stanford AI Index - this appendix identifies important underlying studies or datasets when they materially affect a conclusion used here. Capital Workbench is solely responsible for the synthesis and interpretation in this document; citation does not imply endorsement by the source organizations.

A1. Anthropic Economic Index: Cadences

Bibliographic reference. Massenkoff, M.; Lyubich, E.; Sacher, S.; Hitzig, Z.; Zhang, S.; Heller, R.; and McCrory, P. Anthropic Economic Index report: Cadences. Anthropic, June 26, 2026.

Method and scope. Privacy-preserving telemetry across Claude chat, Cowork, Claude Code, and first-party API surfaces, plus a linked Economic Index Survey launched in April 2026. The final linked survey sample contains about 9,700 respondents after excluding infrequent users.

Primary caution. The survey is not population-representative. Computer and mathematical occupations and management are substantially over-represented, and reported productivity, learning, and career expectations are self-assessments rather than causal estimates.

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Agentic use	Claude usage has shifted from primarily conversational exchanges toward longer-running agentic tasks, particularly through Claude Code and Cowork; a chat transcript no longer captures the full unit of activity.	pp. 2-3
Artifacts and outputs	Anthropic's classifier identified a recognizable artifact in 93% of sampled Claude conversations. Explanations, documents/reports, and guidance are the most common outputs; work conversations most often produce documents and reports.	pp. 9-13
Product scaffolding and autonomy	Across 26 of 31 output types, estimated AI autonomy is higher in Claude Code than in chat/Cowork. Blog/article sessions illustrate the effect: a median of 13 rounds in chat/Cowork versus a single human prompt in Claude Code. The gap persists when comparing sessions served by the same model.	pp. 14-16
Daily-use patterns	Personal conversations rise from about 35% on weekdays to just under 50% on weekends. Weekend coding shifts toward agent design, quantitative trading, gaming, and business-starting activity; U.S. tax-related conversations rose to roughly eight times the May daily average immediately before the filing deadline.	pp. 4-8
Career perceptions	Early-career respondents report high task exposure and greater concern about job loss. More than one-third assigned a probability above 60% that a junior colleague would lose a job within a year.	pp. 19, 26
Experienced value and learning	Respondents report large gains in speed, scope, and quality of work, along with learning and perceived skill-value gains; these are explicitly self-reported experience measures, not causal productivity estimates.	p. 28

Topic	Evidence incorporated into the report	Source pages
Hopes for an AI-shaped economy	More than half of respondents expressed a desire for collaboration on meaningful work; just over half hoped AI would automate tedious work and create more free time; about one-third emphasized broadly shared prosperity.	p. 30

A2. Deloitte State of AI in the Enterprise: The untapped edge

Bibliographic reference. Deloitte. State of AI in the Enterprise: The untapped edge. Deloitte AI Institute, January 2026.

Method and scope. Survey of 3,235 director-level through C-suite respondents across six industries and 24 countries, fielded in August-September 2025, supplemented by 15 interviews with global executives and AI/data-science leaders.

Primary caution. The sample covers organizations with active AI implementations or pilots and reflects senior-leader self-report. Expectations about near-term deployment, preparedness, and business value should not be treated as realized outcomes.

Public source location. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Access and activation	Sanctioned workforce access expanded from under 40% to under 60% in one year, but fewer than 60% of workers with access used AI in their daily workflow.	p. 8
Pilot-to-production gap	Only 25% of respondents had moved at least 40% of experiments into production, while 54% expected to reach that level within three to six months. The report describes production requirements - integration, security, compliance, monitoring, maintenance, and ownership - as the source of a recurring proof-of-concept trap; a three-month use case can stretch to 18 months.	pp. 8-9
Value and process redesign	Sixty-six percent report efficiency/productivity benefits today, while 20% report revenue growth versus 74% that hope to achieve it. Current transformation approaches divide roughly into 37% surface-level use, 30% key-process redesign, and 34% deeper transformation of products, processes, or business models.	pp. 10-11
Work and career architecture	Eighty-four percent had not redesigned jobs around AI. The report highlights entry-level tasks such as data entry, reconciliation, and first-line support as early automation targets and uses a loan-officer example to show how authority, explanation, accountability, and career development must be redesigned. Fifty-three percent had considered pod-based or non-hierarchical structures, while only 16% had moved to them to a great or maximum extent.	pp. 12-13
Sovereign AI	Seventy-seven percent say the location of AI development is a key technology-selection factor. Eighty-three percent rate data-residency constraints and in-region compute considerations as at least moderately important to strategic planning; 66% express at least moderate concern about reliance on foreign-owned AI technology and infrastructure.	pp. 5, 15-16
Agentic governance	Seventy-four percent plan to deploy agentic AI within two years, while 21% report a mature governance model for autonomous agents. The report emphasizes decision boundaries, action logs, monitoring, human approval gates, and cross-functional escalation.	pp. 17-21
Physical AI	Fifty-eight percent report some use of physical AI, with adoption projected by respondents to reach 80% within two years. A warehouse example illustrates total-cost-of-ownership risk: hundreds of thousands of dollars in AI development can be accompanied by millions in physical infrastructure, robotics, and facility modification.	pp. 22-27
Preparedness	Forty-two percent rate strategy as highly prepared, while only 20% say the same about talent; technical infrastructure and data management also lag. An executive interview estimating that 80-90% of new use cases are generative AI is anecdotal and is treated as such in the body.	pp. 28-29

A3. International AI Safety Report 2026

Bibliographic reference. Bengio, Y., et al. International AI Safety Report 2026. DSIT 2026/001, February 2026.

Method and scope. Independent synthesis of research on general-purpose AI capabilities, emerging risks, and risk management. Development involved more than 100 experts, including nominees from over 30 countries and international organizations. The evidence base covers research published before December 2025.

Primary caution. The report deliberately focuses on frontier and emerging risks rather than the full spectrum of AI policy issues. Documented harms coexist with severe scenarios whose probability and timing remain highly uncertain; the report does not endorse a specific regulatory approach.

Public source location. https://internationalaisafetyreport.org

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Capability and the jagged frontier	Capabilities improved sharply in mathematics, coding, science, and autonomous operation, increasingly through post-training and inference-time scaling. The report repeatedly emphasizes jagged performance: systems can excel on demanding tasks while failing simpler spatial, interface, or recovery tasks.	pp. 10-31
Malicious use	Documented harms include scams, fraud, non-consensual imagery, and cyber assistance. Biological and chemical misuse is treated as more uncertain: developers added safeguards after testing could not rule out meaningful novice assistance, while physical materials, tacit knowledge, and laboratory execution remain barriers.	pp. 45-70
Malfunction and reliability	Current systems still fabricate information, produce flawed code, and give misleading advice. Agents raise the stakes because errors can translate into external actions before a human can intervene; current techniques reduce failure rates but do not meet the reliability required in many high-stakes settings.	pp. 71-83
Systemic risk and labor	Systemic risks include labor disruption and threats to human autonomy. Early aggregate employment effects are mixed, while some studies show concentrated effects on selected tasks, occupations, and junior workers.	pp. 84-95
Evaluation gap	Pre-deployment tests do not reliably predict real-world utility or risk. Models may exploit shortcuts, behave differently under deployment conditions, or recognize evaluation contexts; the report therefore argues for post-deployment monitoring and learning in addition to testing.	pp. 97-105
Defense in depth	Risk management is presented as layered rather than dependent on a single safeguard: threat modeling, capability evaluations, model-level controls, system restrictions, monitoring, access controls, incident reporting, and societal resilience. Open-weight models create additional challenges because safeguards can be removed and releases cannot be recalled.	pp. 105-145
Loss of control	The report distinguishes current operational failures from loss-of-control scenarios. It states that current systems lack the capabilities required for loss of control, while noting improvement in relevant capabilities such as autonomous operation and evaluation gaming.	pp. 76-83

A4. Microsoft Global AI Diffusion: Q1 2026 Trends and Insights

Bibliographic reference. Microsoft AI Economy Institute. Global AI Diffusion: Q1 2026 Trends and Insights. Microsoft, May 2026.

Method and scope. AI diffusion is measured as the share of people ages 15-64 who used a generative-AI product during the reporting period, derived from aggregated and anonymized Microsoft telemetry and adjusted for OS/device-market share, internet penetration, and population.

Primary caution. This is one ecosystem and one diffusion metric. It estimates breadth of use, not intensity, quality, productivity, or business value. National rankings can shift because of product capability, access, language support, prices, or measurement coverage.

Method and data reference. https://aka.ms/AI_Diffusion_Technical_Report

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Global diffusion	Worldwide usage rose from 16.3% to 17.8% of the working-age population in Q1 2026. The UAE led the national leaderboard at 70.1%; Singapore was 63.4%; the United States was 31.3% and ranked twenty-first.	pp. 2-4
North-South gap	Usage reached 27.5% in the Global North and 15.4% in the Global South, widening the gap to 12.1 percentage points. The report links the divide to electricity, connectivity, device access, digital skills, and local-language capability.	pp. 5-7
Language and Asia	Twelve of the fifteen fastest-growing economies since June 2025 were in Asia. Stronger local-language and multimodal performance is presented as an important mechanism making AI more useful for messaging, search, learning, and content creation.	pp. 6-9
Japan case	Japan's ranking improved from fifty-sixth to forty-eighth; adoption rose 3.4 percentage points in one quarter. Selected Japanese professional-exam performance increased from roughly 50.8% in earlier systems to above 90% in recent systems, while Japanese Git pushes grew 129% year over year versus 78% globally.	pp. 8-9
Software production	Git pushes reached 380 million in Q1 2026, up 78% year over year; new repositories reached 21.3 million, up 45%; and pull requests associated with AI coding agents reached 2.3 million in March 2026, about 28 times the May 2025 level.	pp. 10-12
Software employment	The report notes that U.S. software-developer employment reached roughly 2.2 million in 2025, up 8.5% year over year, and that March 2026 employment was about 4% above March 2025. It explicitly cautions that the full labor-market effect of AI-assisted coding remains too early to determine.	pp. 3, 12
Underlying measurement reference	The report identifies Misra, Wang, McCullers, White, and Lavista Ferres (2025), 'Measuring AI Diffusion: A Population-Normalized Metric for Tracking Global AI Usage,' as its technical-method reference and provides a public Q1 2026 diffusion dataset.	p. 14

A5. State of AI: An Empirical 100 Trillion Token Study with OpenRouter

Bibliographic reference. Aubakirova, M.; Atallah, A.; Clark, C.; Summerville, J.; and Midha, A. State of AI: An Empirical 100 Trillion Token Study with OpenRouter. December 2025.

Method and scope. Observational analysis of more than 100 trillion tokens on OpenRouter, a multi-model inference platform supporting more than 300 models from more than 60 providers. Most analyses cover November 2024-November 2025; detailed content classification begins in May 2025. The study primarily analyzes metadata, with a sampled classifier used for task categories.

Primary caution. The dataset is shaped by OpenRouter's user mix, model availability, pricing, and developer/platform skew. Billing location is used as a geographic proxy. The results are strong evidence of behavior on OpenRouter, not global model-market share or a census of enterprise AI.

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Open-weight scale	Open-weight models reached roughly one-third of platform token usage by late 2025. Chinese-developed open models reached nearly 30% of total usage in some weeks and averaged about 13% of weekly token volume over the observation window.	pp. 5-7

Topic	Evidence incorporated into the report	Source pages
Fragmented competition	The open segment moved from early concentration toward a plural market. By late 2025, no single open model consistently held more than about 20-25% of open-weight tokens, and capable new entrants could reach meaningful usage within weeks.	pp. 6-9
Workload mix	Roleplay accounts for more than half of open-weight token volume in the sampled classification, while programming is roughly 15-20%. Provider profiles differ materially, reinforcing the case for workload-specific model choice rather than a universal model.	pp. 9-12, 18-23
Agentic inference	Reasoning-optimized models rose from negligible usage to more than half of platform tokens during 2025. The study also tracks increased tool invocation and longer, multi-step interaction patterns as evidence of a shift from single-turn completion toward agentic workflows.	pp. 12-18
Geography	Asia rose from roughly 13% of OpenRouter platform spend in the earliest weeks of the dataset to about 31% in the most recent period, illustrating rapid growth in regional demand on the platform.	pp. 24-25
Glass Slipper retention	Most cohorts churn quickly, but the June 2025 Gemini 2.5 Pro cohort and May 2025 Claude 4 Sonnet cohort retained about 40% of users at month five. The authors interpret these persistent cohorts as evidence of workload-model fit and operational switching friction.	pp. 25-27
Price and usage	Across models, cost and usage show weak overall correlation. Premium proprietary models cluster in high-cost/high-usage workloads, while open models are prominent in low-cost/high-volume workloads; the report argues that reliability, capability, and fit can outweigh marginal token price.	pp. 28-32

A6. Stanford AI Index Report 2026

Bibliographic reference. Sajadieh, S.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Santarlasci, L.; Pava, J.; Maslej, N.; et al. The AI Index 2026 Annual Report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, April 2026. DOI: 10.48550/arXiv.2606.15708.

Method and scope. Independent, cross-domain compilation of datasets and studies spanning research and development, technical performance, responsible AI, the economy, science, medicine, education, policy/governance, and public opinion. The report explicitly combines administrative data, surveys, benchmarks, academic studies, and partner datasets.

Primary caution. Because the AI Index is a composite report, evidence lags and methods differ by chapter and figure. Some headline statistics are direct administrative measures; others are estimates or syntheses of underlying studies. The relevant unit, population, and date should be checked before comparing numbers across chapters.

Public source location. https://doi.org/10.48550/arXiv.2606.15708

Evidence incorporated in this report

Topic	Evidence incorporated into the report	Source pages
Capability and jagged performance	The 2026 top takeaways include gold-medal International Mathematical Olympiad performance alongside only 50.1% analog-clock accuracy; OSWorld agent success rising from 12% to about 66%; and robotics performance of 89.4% in simulation versus 12% on household tasks.	pp. 9, 68-72
Evaluation and responsible AI	The report describes benchmark saturation, declining transparency, and an evaluation gap as frontier models converge. Documented AI incidents rose from 233 in 2024 to 362, while responsible-AI benchmark reporting remains inconsistent.	pp. 9, 68-71, 126-129
Productivity	Stanford summarizes generally positive micro-level evidence in customer support, software development, and marketing: support agents resolved 14-15% more issues per hour, developers completed 26% more pull requests in one field study, and multimodal ad creation increased output per worker by 50%. It also highlights a METR study in which experienced open-source developers became 19% slower despite believing AI was helpful.	pp. 219-220

Topic	Evidence incorporated into the report	Source pages
Underlying productivity studies	The AI Index attributes the support result to Brynjolfsson et al., the pull-request result to Cui et al., and the marketing result to Ju and Aral; the developer-slowdown result is attributed to METR/Becker et al. These studies differ in setting and should not be combined into a single universal productivity estimate.	pp. 219-220
Junior-worker pressure	The report notes that employment among U.S. software developers ages 22-25 fell close to 20% while older cohorts continued to grow, and it presents related evidence of weaker entry rates and junior hiring in highly AI-exposed work. The report cautions that labor effects remain early and uneven.	pp. 10, 219-223
Science and medicine	AI systems increasingly attempt multi-step scientific workflows, but validation and reproducibility remain bottlenecks. In medicine, ambient clinical scribes have reduced documentation time by as much as 83% in some systems, while a review of more than 500 clinical-AI studies found nearly half used exam-style questions and only 5% used real clinical data.	pp. 231-287
Education	More than 80% of U.S. high-school and college students use generative AI for school-related work, while only about half of middle and high schools have an AI policy and 6% of teachers say those policies are clear.	pp. 288-322
Infrastructure and investment	The United States hosts 5,427 data centers; TSMC fabricates almost every leading AI chip. U.S. private AI investment was estimated at $285.9 billion in 2025, and AI data-center power capacity reached 29.6 GW. These measures illustrate the physical concentration beneath model competition.	pp. 9-10, 12-39, 171-190
Sovereignty and global competition	The U.S.-China frontier performance gap has narrowed sharply, while model production, investment, publications, patents, and open-source contributions remain distributed differently across countries. The report treats AI sovereignty as a growing national-policy theme.	pp. 9-11, 14-19, 323-359
Public opinion	The report finds simultaneous increases in optimism and nervousness and a large expert-public gap in expectations for jobs, the economy, and medicine. Trust in institutions to govern AI varies sharply by country.	pp. 360-384

Bibliographic references

The six sources below constitute the complete direct research base for this report. The body does not add outside sources independently of them. Where the selected reports rely on other studies or datasets, those dependencies are summarized above when material to a claim used in this document.

S1. Massenkoff, M.; Lyubich, E.; Sacher, S.; Hitzig, Z.; Zhang, S.; Heller, R.; and McCrory, P. Anthropic Economic Index report: Cadences. Anthropic, June 26, 2026.

S2. Deloitte. State of AI in the Enterprise: The untapped edge. Deloitte AI Institute, January 2026. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html

S3. Bengio, Y., et al. International AI Safety Report 2026. DSIT 2026/001, February 2026. https://internationalaisafetyreport.org

S4. Microsoft AI Economy Institute. Global AI Diffusion: Q1 2026 Trends and Insights. Microsoft, May 2026. Method and data reference: https://aka.ms/AI_Diffusion_Technical_Report

S5. Aubakirova, M.; Atallah, A.; Clark, C.; Summerville, J.; and Midha, A. State of AI: An Empirical 100 Trillion Token Study with OpenRouter. December 2025.

S6. Sajadieh, S.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Santarlasci, L.; Pava, J.; Maslej, N.; et al. The AI Index 2026 Annual Report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, April 2026. DOI: 10.48550/arXiv.2606.15708. https://doi.org/10.48550/arXiv.2606.15708

Publication note

Prepared by Capital Workbench. The source organizations retain ownership of their respective reports and data. This report is an independent synthesis and should not be interpreted as an endorsement by Anthropic, Deloitte, the International AI Safety Report team, Microsoft, OpenRouter, Stanford University, or their contributors.