INFORMATION AT A GLANCE · EDITION 03

The state of AI, *one page per source.*

A fully-cited distillation of what the frontier labs, the Big-6 consulting firms, independent research, and the governance bodies are actually reporting — every figure quoted from its primary source, with page or section and a link to verify.

22 primary sources, one page each	**100+** findings, each cited to source	**81%** of organisations report no meaningful bottom-line gain from AI MCKINSEY '26	**67%** of AI impact is organisational, not individual MICROSOFT

Prepared by **SP Optima** — independent, business-first AI advisory.

15 August 2026
spoptima.com

CROSS-SOURCE SYNTHESIS · 2026 Q3

Four things every major source is converging on

Where consulting firms, frontier labs and independent research agree independently, the signal is strong. Each claim below is evidenced on the source's own page.

01

Governance & trust is the #1 blocker

Security, risk and oversight — not regulation or technology — now gate the move to agentic AI.

Deloitte: only 21% have mature agent governance · EY: 66% say human oversight is essential

DELOITTE · EY · KPMG · MCKINSEY

02

Adoption is near-universal — value isn't

A divide separates the few who capture returns from the many who stall.

McKinsey: 81% report no meaningful bottom-line gain · PwC: 20% of firms capture 74% of returns

MCKINSEY · PWC · STANFORD

03

The bottleneck is the operating model

Strategy, workflow redesign and enablement — not tool choice — decide who wins.

BCG: clear strategy = +25pp impact · Microsoft: org factors drive 2× the impact of individuals

BCG · MICROSOFT · MCKINSEY

04

Agents are the through-line of 2026

AI is shifting from answering to doing, and containment is now the open question.

UK AISI: agents took 19 out-of-scope actions in its own tests · OpenAI: GPT-5.6 more often acts beyond user intent

UK AISI · OPENAI · MICROSOFT · METR

WHAT THIS MEANS FOR YOU

The evidence converges on one point: **the gap is organisational, not technical.** Readiness, governance and enablement, rather than which tool you buy, separate the roughly 20% of organisations reporting real returns from the majority that stall. The 22 source pages that follow set out the evidence, each figure checkable at source.

01
Consulting

BIG-4 + MCKINSEY & BCG

What the major advisory firms are telling boards about AI adoption, value capture and governance.

- McKinsey
- Deloitte
- BCG
- PwC
- KPMG
- EY

CONSULTING FIRM

McKinsey

The State of Organizations 2026

MAR 2026 · PDF · 74PP · VIEW SOURCE ↗

88% / 81%
experiment with AI — but report no meaningful bottom-line gain

<20%
of adopters see significant tangible impact

86%
of leaders feel unprepared to adopt AI day-to-day

$5 : $1
McKinsey's advised spend ratio — people to technology

Key findings — each figure quoted from its primary source

AI use is near-universal but value is rare — 88% experimenting, 81% report no meaningful bottom-line gains.

The State of Organizations 2026, p.3 — "While 88 percent of organizations are now experimenting with AI, 81 percent do not report any meaningful bottom-line gains."

Fewer than one in five adopters see real impact — a readiness gap, not a tooling gap.

Same report, p.7 — "Less than 20 percent of companies that have tried to adopt the technology have seen significant tangible impact on their bottom lines."

86% of leaders feel unprepared to adopt AI day-to-day; 1 in 6 have no clear C-level AI owner.

Same report, p.7 — "Eighty-six percent of leaders feel that their organizations are not prepared to adopt AI in day-to-day operations."

Value capture is people-led — the rule of thumb: $5 on people for every $1 on technology.

Same report, p.8 — "for every $1 spent on technology, $5 should be spent on people."

Even the barriers are about people, not tools — the top three are concerns about AI (46%), regulation/ethics/law (44%) and organisational change (39%).

Same report, p.7, Exhibit 1 — "the top barrier involves concerns about AI itself... (46 percent). The second top barrier addresses regulatory, ethical, or legal concerns (44 percent)... Third is organizational challenges, including change management... (39 percent)"

A QUESTION WORTH ASKING

What does your own split between technology and people actually look like? The advised ratio is five dollars on people for every one on technology, which would mean costing the training, the redesign and the change work first, and the licences second.

Caveat. *Self-published consultancy survey (commercial incentive); leader perceptions, not audited outcomes; some headline figures lean on prior McKinsey research. n=10,018, 15 countries, fielded Jun–Sep 2025.*

CONSULTING FIRM

Deloitte

The State of AI in the Enterprise 2026 — From Ambition to Activation

JAN 2026 · PDF · VIEW SOURCE ↗

+50%
one-year jump in workers' access to sanctioned AI tools

25% → 54%
have scaled ≥40% of pilots now → expected within 3–6 months

21%
have a mature governance model for autonomous agents

34%
are using AI to deeply transform the business

Key findings — each figure quoted from its primary source

Worker AI access jumped 50% in a year — from <40% to ~60% of workers.

State of AI in the Enterprise 2026, p.4 — "broadened worker access to AI by 50% in just one year—growing from fewer than 40% to around 60%"

A pilot-to-production gap persists — only 25% scaled ≥40% of pilots (54% expect to within 3–6 months).

Same report, p.8 — "25% moved 40% or more of their AI experiments into production To date"

Agent governance lags adoption — 74% plan agentic AI in 2 yrs, only 21% have a mature governance model.

Same report, p.6 — "74% companies plan to deploy agentic AI within two years. Yet, 1 in 5 (21%) report having a mature model for governance"

Only 34% are truly reinventing the business; 37% remain surface-level.

Same report, p.4 — "34% of companies are starting to use AI to deeply transform their businesses… remaining 37% are only using AI at a surface level"

Reported transformative impact more than doubled in a year — from 12% to 25% of leaders.

Same report, p.10 — "25% of leaders now reporting that AI is having a transformative effect on their companies—more than double from 12% a year ago"

A QUESTION WORTH ASKING

Who will be accountable for what your agents do, and does that role exist yet? Most organisations expect to deploy agents within two years, and very few have the governance standing before the agents arrive.

Caveat. *Self-reported perception survey of director–C-suite leaders. n=3,235, 24 countries, Aug–Sep 2025.*

CONSULTING FIRM

BCG

AI at Work 2026 (4th ed.): Why Strategy Matters More Than Tools

JUN 2026 · REPORT (PRESS RELEASE CITED) · VIEW SOURCE ↗

74%
of frontline workers are now regular AI users (+20pp in 2 yrs)

+25pp
business-impact lift from a clear strategy (vs ~5pp from tools alone)

42%
of regular users save at least a full workday each week

30%
have AI agents integrated into workflows — double last year

Key findings — each figure quoted from its primary source

Frontline AI adoption surged to 74% regular users (+20pp in two years).

BCG press release, PR Newswire — "74% now regular users, up more than 20 percentage points over the previous two years"

Strategy beats tools — clear strategy lifts impact 25pp vs ~5pp for tools alone.

Same release, Strategy vs tools — "Clear strategy lifts measurable business impact by 25 percentage points… better tools, without that strategy and redesign, move it only by approximately 5 points"

Time is saved but squandered — 42% save ≥1 workday/week, but 66% get no guidance on using it.

Same release, body — "42% of regular frontline users report saving at least a full workday… 66% report they get limited or no guidance"

The "joy paradox" — 67% report higher job satisfaction, yet 41% report increased cognitive load.

Same release, body — "Two-thirds (67%) of regular AI users say it has improved job satisfaction… 41% report increased cognitive load"

AI is already rewriting roles — 72% say it has changed the skills their job needs, and 47% now spend more time directing AI than doing the work.

Same release, body — "72% of respondents now say AI has already considerably changed skills expectations in their roles… 47% report spending more time managing and directing AI than doing the work itself"

A QUESTION WORTH ASKING

Has anyone decided what the saved hours are for? Time is being released across the workforce faster than anyone is deciding how to reinvest it, and unclaimed time tends to be absorbed quietly back into the working day.

Caveat. *bcg.com publication + slides PDF are Akamai-blocked to scripts; figures cited from BCG's official press release (same numbers). Survey n=11,749, 14 markets.*

CONSULTING FIRM

PwC

Want ROI from AI? Go for growth (AI performance study)

APR 2026 · PDF · 30PP · VIEW SOURCE ↗

74%
of all AI-driven returns captured by just the top 20% of firms

7.2×
AI-driven financial performance of the most AI-fit vs other firms

2.8×
more likely to grow the share of decisions made without humans

2.5×
more of revenue that AI leaders invest in AI

Key findings — each figure quoted from its primary source

AI value is concentrated — 20% of firms capture 74% of AI-driven returns.

Want ROI from AI? Go for growth, p.3 — "20% of the 1,217 companies we surveyed capture 74% of the AI-driven returns"

The most AI-fit firms deliver 7.2× the AI-driven financial performance of others.

Same study, p.4 — "deliver AI-driven financial performance that's 7.2 times as high as the other respondents'"

Leaders aim AI at growth — 2.6× as likely to say AI improved business-model reinvention.

Same study, p.11 — "2.6 times as likely as others to report that AI has improved their ability to reinvent their [business model]"

Leaders invest materially more — 2.5× as much of revenue in AI.

Same study, p.15 — "The leading companies… invest materially more in AI than other companies do: 2.5 times as much of their revenue"

The single strongest "AI-fitness" factor is capturing growth from industry convergence — leaders are 1.8× as likely to use AI to spot emerging value pools.

Same study, p.12 — "the ability to capture growth opportunities resulting from industry convergence stands out in our research as the single strongest AI fitness factor… 1.8 times as likely as other companies to use AI to spot emerging value pools"

A QUESTION WORTH ASKING

Is your AI work aimed at cutting cost or at finding growth? The organisations capturing most of the value are pointing it at new revenue, so it is worth checking which of the two your funded portfolio actually describes.

Caveat. *Self-reported exec survey; "AI-driven performance" is self-estimated, not audited; "leaders" = top-20% cohort. n=1,217, 25 sectors, Oct–Nov 2025.*

CONSULTING FIRM

KPMG

Global AI Pulse Survey — Q2 2026

JUN 2026 · QUARTERLY SURVEY · PRESS RELEASE · VIEW SOURCE ↗

79%
of leaders name AI a key investment priority — up from 74% last quarter

9% → 18%
organisations orchestrating multiple AI agents across workflows — doubled in one quarter

26%
the share with full, real-time visibility into what their AI costs to run

57% v 21%
realise meaningful value where the CEO is accountable for AI decisions vs where not

Key findings — each figure quoted from its primary source

AI stays the top investment priority — cited by 79% of leaders (up from 74% last quarter), with planned spend steady at ~$202M over 12 months.

KPMG Global AI Pulse Q2 2026 (global release), § Investment — "AI remains a top priority, cited by 79 percent of leaders as a key investment area (up from 74 percent last quarter)"

Agent deployment plateaued (53% vs 55%) but multi-agent orchestration doubled 9% → 18% — coordination is the new frontier.

KPMG Global AI Pulse Q2 2026 (US release), § Agent deployment — "the percentage of organizations orchestrating multiple AI agents across workflows doubled from 9% to 18%"

A cost-visibility gap is opening — only 26% have full, real-time visibility into what their AI systems cost to operate.

KPMG Global AI Pulse Q2 2026 (US release), § Cost management — "only 26% report full, real-time visibility into what their AI systems cost to operate"

Accountability tracks value — where the CEO owns AI decisions, firms far more often realise value (57% vs 21%) and established ROI (14% vs 4%).

KPMG Global AI Pulse Q2 2026 (global release), § Accountability — "Organizations where CEOs are accountable for decisions based on AI outputs, report higher confidence in their AI strategy, (60 percent vs. 22 percent), are more likely to realize meaningful business value (57 percent vs. 21 percent) and report established ROI (14 percent vs. 4 percent)"

AI is entering everyday work — the share where AI is "part of everyday work" rose to 22%, up from 13% in Q1.

KPMG Global AI Pulse Q2 2026 (global release), § Maturity — "More organizations have reached the stage where AI is part of everyday work — rising to 22 percent, up from 13 percent in Q1"

A QUESTION WORTH ASKING

Could you say today what your AI costs to run? If assembling that answer would take more than a week, that is a governance gap rather than a finance one, and it gets harder as usage grows.

Caveat. *A quarterly pulse (press-release figures), not a deep report; self-reported C-suite perceptions, and KPMG sells AI advisory (commercial incentive). n=2,145 senior leaders, 20 markets, fielded 28 Apr–25 May 2026.*

CONSULTING FIRM

EY

2026 AI Sentiment Report (AI Sentiment Index Study)

MAR 2026 · HTML · VIEW SOURCE ↗

84%
have used AI in the past six months

16%
have used AI that acts without any human intervention

73%
fear they can no longer tell what is real from AI-generated

66%
say human oversight of AI remains essential

Key findings — each figure quoted from its primary source

AI use is near-universal — 84% used AI in the past six months.

2026 AI Sentiment Report, § Adoption grows — “In the past six months alone, more than eight in 10 (84%) respondents used AI”

Autonomous AI is real but minority — 16% use AI that acts without human intervention.

Same report, opening findings — “16% globally report using AI systems that act without human intervention”

Trust lags adoption — 73% fear they can’t tell what’s real; 66% worry about breaches.

Same report, Adoption outpaces confidence — “73% fear they will no longer be able to tell what is real or AI-generated”

Demand for human control persists — 66% say human oversight remains essential.

Same report, Adoption outpaces confidence — “Almost seven in 10 (66%) say human oversight remains essential”

“Pioneer” markets show where this is heading — there 94% use AI and nearly one in four (24%) have already used autonomous AI.

Same report, Pioneer markets — “collectively 94% of people in these markets report using AI, with nearly one in four (24%) of them having already used autonomous AI”

A QUESTION WORTH ASKING

How will your customers know when they are dealing with a machine? Most people now doubt they can tell, which makes being consistently clear about it a position worth holding in your market.

Caveat. *This flagship is a consumer-sentiment study (n=18,152, 23 markets). A separate EY poll of 500 US tech execs found 78% say adoption is outpacing their ability to manage risk.*

02

Independent research

ACADEMIC & MARKET DATA

Neutral, data-led reads on adoption, the value gap, and where enterprise AI money actually flows.

- Stanford HAI — AI Index
- MIT CISR

INDEPENDENT RESEARCH

Stanford HAI — AI Index

The AI Index 2026 Annual Report (9th ed.)

APR 2026 · PDF · 425PP · VIEW SOURCE ↗

88%	$285.9B	53%	362
of organisations have now adopted AI	US private AI investment in 2025 — 23× China's	gen-AI population adoption in 3 yrs — faster than the PC	documented AI incidents in 2025 (up from 233)

Key findings — each figure quoted from its primary source

Capability is racing — SWE-bench Verified rose from 60% to ~100% of the human baseline in a year.

AI Index 2026, p.9, Takeaway 1 — "performance rose from 60% to near 100% of meeting the human baseline in a single year. Organizational adoption reached 88%"

Investment is concentrated in the US — $285.9B, >23× China.

AI Index 2026, p.10, Takeaway 7 — "U.S. private AI investment reached $285.9 billion in 2025, more than 23 times the $12.4 billion invested in China"

Gen-AI outpaced the PC and internet — 53% adoption in three years; $172B annual US consumer value.

AI Index 2026, p.10, Takeaway 8 — "Generative AI reached 53% population adoption within three years, faster than the PC or the internet"

Governance lags — documented AI incidents rose to 362 (from 233); disclosure regressing.

AI Index 2026, p.9, Takeaway 6 — "Documented AI incidents rose to 362, up from 233 in 2024"

The US–China model gap has all but closed — the top US model led by just 2.7% as of March 2026.

AI Index 2026, p.9, Takeaway 2 — "as of March 2026 Anthropic's top model leads by just 2.7%"

A QUESTION WORTH ASKING

What would you do if a supplier changed its model without telling you? Disclosure from model providers is getting thinner rather than richer, so it is worth knowing how much of your operation assumes a warning that may never come.

Caveat. *Quotes from the report's own "Top Takeaways" (pp.9–11); neutral academic source. 9th edition.*

INDEPENDENT RESEARCH

MIT CISR

Leveraging Digital Colleagues for Enterprise Value

APR 2026 · MIT SLOAN CISR · BRIEFING · VIEW SOURCE ↗

34%
are actively using AI "digital colleagues" alongside staff

+25%
predicted gain in revenue per employee over three years

22%
have done the major workflow redesign that captures value

9%
have begun to formally integrate digital colleagues

Key findings — each figure quoted from its primary source

34% of enterprises are already actively using AI "digital colleagues" that work alongside staff.

Leveraging Digital Colleagues for Enterprise Value (MIT CISR), briefing — "Thirty-four percent of surveyed enterprises were actively using digital colleagues"

Expectations run high — 75% predict a 25% average rise in revenue per employee over three years.

Same briefing, briefing — "75 percent predicted an average increase of 25 percent"

But value-unlocking redesign is rare — only ~22% have done major workflow redesign.

Same briefing, briefing — "Approximately twenty-two percent of the surveyed enterprises reported major workflow redesign."

Only 9% have begun to formally integrate digital colleagues — a small leading cohort.

Same briefing, briefing — "Nine percent of enterprises in our survey had begun to formally integrate digital colleagues."

Where digital colleagues land, adoption goes deep — at legal-AI firm Harvey, 96% of legal staff are active users.

Same briefing, briefing — "96 percent of legal staff are active users"

A QUESTION WORTH ASKING

Are you adopting a tool, or redesigning the work? Very few organisations have done the second, and it is that group reporting the gains, so it is worth checking which of the two your plan actually funds.

Caveat. *MIT CISR (Sloan). The full briefing is members-gated — figures verified via the public preview/abstract, so a non-member sees the abstract only. Self-reported predictions, n=132, Sep 2025. (MIT NANDA has issued no 2026 successor to its Jul-2025 "GenAI Divide".)*

03
Governance & standards

FRAMEWORKS & POLICY

The frameworks and standards an AI-readiness and governance programme maps directly onto.

- NIST — AI RMF
- ISO/IEC 42001
- OECD.AI
- EU AI Act

GOVERNANCE & STANDARDS

NIST — AI RMF

AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1

JAN 2023 · PDF · VIEW SOURCE ↗

4	7	12	Voluntary
core functions — Govern, Map, Measure, Manage	characteristics of trustworthy AI it defines	GenAI-specific risk categories in the companion profile	non-binding by design — guidance, not regulation

Key findings — each figure quoted from its primary source

Four core functions structure AI risk — GOVERN, MAP, MEASURE, MANAGE.

AI RMF 1.0 (NIST AI 100-1), p.3 — “These functions — GOVERN, MAP, MEASURE, and MANAGE — are broken down further into categories and subcategories”

It names seven characteristics of trustworthy AI.

Same framework, p.12 — “valid and reliable, safe, secure and resilient, accountable and transparent, explainable… privacy-enhanced, and fair”

Governance is cross-cutting — infused through the other three functions.

Same framework, p.ii, Fig.5 — “Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions”

The Generative AI Profile (NIST AI 600-1) enumerates 12 GenAI-specific risk categories.

NIST AI 600-1, §2, p.4–5 — “CBRN… Confabulation… Data Privacy… Information Integrity… Information Security… Value Chain and Component Integration”

Each function has a job — MAP frames the context, MEASURE analyses and benchmarks risk, MANAGE allocates resources to treat it.

AI RMF 1.0 (NIST AI 100-1), pp.24–31 — “The MAP function establishes the context to frame risks related to an AI system... The MEASURE function employs quantitative, qualitative, or mixed-method tools... The MANAGE function entails allocating risk resources”

NIST is now drafting testable documentation requirements: AI 300-1 ipd (Jul 2026) sets templates that conformity can be assessed against.

NIST AI 300-1 ipd, Guidance and Templates for Public-Facing AI Documentation, p.9, §1 Scope — “requirements for documentation artifacts in the form of templates against which conformity of documentation artifacts can be assessed”

A QUESTION WORTH ASKING

How long would it take you to document the AI you already use? Buyers, insurers and regulators are converging on the same request, and the answer is rarely as quick as people expect.

Caveat. *Voluntary guidance, not regulation, and it prescribes no metrics or thresholds. AI RMF 1.0 is a standing reference from Jan 2023 (GenAI Profile Jul 2024). The AI 300-1 documentation work is an initial public draft with comments open to 16 September 2026 and no guarantee of adoption by ISO/IEC JTC 1/SC 42.*

GOVERNANCE & STANDARDS

ISO/IEC 42001

ISO/IEC 42001:2023 — AI Management System

DEC 2023 · STANDARD · VIEW SOURCE ↗

Dec 2023	1st	Certifiable	Any org
first edition published (ISO/IEC JTC 1/SC 42)	the world's first AI management-system (AIMS) standard	organisations can be independently audited against it	applies regardless of size, type or sector

Key findings — each figure quoted from its primary source

The world's first certifiable AI management system (AIMS) standard.

ISO/IEC 42001:2023 (Abstract), Abstract — "establishing, implementing, maintaining and continually improving an AI… management system within… an organization"

Built to drive responsible AI in pursuit of objectives.

Same standard, Abstract — "intended to help the organization develop, provide or use AI systems responsibly"

Universal scope — any organisation, regardless of size or sector.

Same standard, Scope — "applicable to any organization, regardless of size, type and nature, that provides or uses products or services that utilize AI"

Published 18 Dec 2023; jointly maintained by ISO and IEC (JTC 1/SC 42).

Same standard, Publication details — "Publication Date: 2023-12-18… Edition: 1.0"

A QUESTION WORTH ASKING

When does certification stop being optional in your market? It is voluntary today, but procurement teams tend to adopt standards well before regulators require them.

Caveat. *Full text paywalled; quotes limited to the official abstract/scope. Certification is by accredited third parties, not ISO.*

GOVERNANCE & STANDARDS

OECD.AI

OECD AI Principles + Policy Observatory

2019, REV. MAY 2024 · LEGAL INSTRUMENT + OBSERVATORY · VIEW SOURCE ↗

2019
adopted — first intergovernmental AI standard (rev. 2024)

47
national adherents to the Principles

5
values-based principles for trustworthy AI

2,348
national AI policy initiatives tracked by the Observatory

Key findings — each figure quoted from its primary source

The first intergovernmental AI standard — adopted May 2019, updated May 2024.

Recommendation of the Council on AI (OECD/LEGAL/0449), overview — “adopted by the OECD Council… on 22 May 2019… the first intergovernmental standard on AI”

47 adherents now align to the Principles.

OECD AI Principles, Adherents — “there are 47 adherents to the Principles”

Five values-based principles define trustworthy AI.

Same page, Values-based principles — “Inclusive growth… Human rights and democratic values… Transparency… Robustness, security and safety… Accountability”

The Observatory tracks 2,348 policy initiatives across 80+ jurisdictions.

OECD.AI Policy Observatory, dashboard — “List of policy initiatives (2348)… more than 80 jurisdictions and organisations”

A QUESTION WORTH ASKING

How many sets of rules will your business need to satisfy? The number of jurisdictions setting their own AI expectations keeps growing, so someone should be tracking which of them reach you.

Caveat. *The 2,348 figure is a live count (captured 2026-06-10) and will drift; adherent phrasing varies across OECD pages.*

GOVERNANCE & STANDARDS

EU AI Act

Digital Omnibus on AI — Regulation (EU) 2026/1744

24 JUL 2026 · OFFICIAL JOURNAL · BINDING LAW · VIEW SOURCE ↗

2 Dec 2027
the new application date for Annex III high-risk AI systems, deferred 16 months from 2 August 2026

2 Aug 2028
when high-risk obligations bite for AI embedded in regulated products (Annex I)

2 Aug 2026
unchanged: the AI Act's general date of application, so transparency duties are live now

4 months
the transition granted to providers already on the market for AI-content marking duties

Key findings — each figure quoted from its primary source

The EU deferred its high-risk regime by 16 months, the first substantive amendment to the AI Act since 2024.

Regulation (EU) 2026/1744, Recital 40 — "it is appropriate that the date of application of Sections 1, 2 and 3 of Chapter III is set to 2 December 2027 for AI systems classified as high-risk pursuant to Article 6(2) and Annex III, and to 2 August 2028 for AI systems classified as high-risk pursuant to Article 6(1) and Annex I."

The deferral is written into the Act itself, not granted as enforcement discretion.

Same regulation, Article 1(40)(b), amending Article 113 — "Chapter III, Sections 1, 2, and 3, with the exception of Article 6(5), shall apply from: (i) 2 December 2027 as regards AI systems classified as high-risk pursuant to Article 6(2) and Annex III"

The stated reason is missing scaffolding: standards, specifications and national authorities were not ready.

Same regulation, Recital 40 — "the delayed availability of standards, common specifications, and alternative guidance and the delayed establishment of national competent authorities lead to challenges that jeopardise the effective entry into application of those obligations"

This is not deregulation — new prohibitions on non-consensual intimate imagery and CSAM apply from 2 December 2026.

Same regulation, Article 1(40)(a), amending Article 113 — "Chapters I and II shall apply from 2 February 2025, with the exception of Article 5(1), first subparagraph, points (ba) and (bb), and Article 5(1a) and (1b) which shall apply from 2 December 2026"

Transparency duties still land now, with only four months' grace for systems already on the market.

Same regulation, Recital 38 — "it is appropriate to introduce a transitional period of four months for providers who have already placed their systems on the market before the 2 August 2026"

It became binding law three days after publication, on 27 July 2026.

Same regulation, Article 4 — "This Regulation shall enter into force on the third day following that of its publication in the Official Journal of the European Union."

A QUESTION WORTH ASKING

Is this sixteen months of preparation, or sixteen months of delay? The high risk deadline moved to December 2027, but the work required to meet it did not shrink.

Caveat. *Primary legislation rather than research, so there are no survey caveats, but there is an interpretive one: deferral is not repeal, and the Act still applies extraterritorially to non-EU providers placing systems on the EU market. National market-surveillance capacity remains uneven.*

04
Capability & safety

MEASUREMENT & RISK

Independent measurement of how fast capability — and risk — is actually moving.

- Epoch AI
- METR
- UK AI Security Institute
- International AI Safety Report

CAPABILITY & SAFETY

Epoch AI

AI at work: how AI is changing everyday job tasks (Epoch AI / Ipsos)
6 AUG 2026 · SURVEY · HTML · VIEW SOURCE ↗

1 in 5
US workers say AI now handles a task they previously gave to a coworker or contractor

7.1%
of workers report that reallocation for "analyzing data" — the single most affected task

66%
of AI outputs are used as produced or with only minor revisions

37% → 53%
share of tasks that take less time when AI does part of the work, then most or all of it

Key findings — each figure quoted from its primary source

Substitution is real but task-level: one in five workers report AI absorbing work once handed to another person.

AI at work (Epoch AI / Ipsos), § AI is now delegated work previously given to other humans — "One in five said that this has happened for at least one of the ten tasks we measured."

Analysis work is displaced first, at 7.1% of respondents, ahead of document reading and record keeping.

Same publication, § AI is now delegated work previously given to other humans — "This reallocation of work to AI appears across all ten tasks we measured. It is most common for 'analyzing data' (7.1% of respondents)"

Review effort is lighter than the standard advice assumes — two-thirds of AI output ships nearly unchanged.

Same publication, § Workers accept most AI outputs with minimal revision — "Across tasks both partially and fully performed by AI, 66% of AI outputs are used as produced or with only minor revisions."

Time savings scale with how much of the task is handed over, from 37% of tasks to 53%.

Same publication, § AI usually doesn't do the whole task, but saves time most often when it does — "Respondents report that when AI assists with part of the work, 37% of tasks take less time. Among tasks where AI does most or all of the work, the share rises to 53%."

Access is uneven — employer-provided AI drives far heavier work use than free tiers (76% vs 38%).

Epoch AI, "Half of employed AI users now use it for work" (Apr 2026), § Subscription type — "Employer-provided subscriptions show the highest adoption at 76%"

A QUESTION WORTH ASKING

Who is checking the work? Two thirds of AI output is now used with little or no revision, so it is worth knowing whether a drop in quality would reach you through a control of your own or through a customer.

Caveat. *Self-reported perceptions from a probability panel, not measured output: n=1,106 employed US adults on the Ipsos KnowledgePanel, fielded 10–19 July 2026, margin of error ±3.0 percentage points. US-only, so it should be read as directional for Australian workforces.*

CAPABILITY & SAFETY

METR

Expenditure Horizon: Measuring Optimization Ability, with an Application to NanoGPT

21 JUL 2026 · RESEARCH NOTE · HTML · VIEW SOURCE ↗

$0–$3K
the human-labour value METR estimates today's agents produce, after spending over $10K running them

$2,500
the human labour cost of each 1% improvement on the NanoGPT benchmark

16 hrs
of expert human time behind that same 1% improvement

~70%
of the agent's ideas the human maintainer judged mergeable, though mostly low-novelty

Key findings — each figure quoted from its primary source

METR proposes pricing agent output in dollars of equivalent human effort rather than hours of task length.

Expenditure Horizon (METR), § Introduction — "We first estimate the returns to human effort in optimizing NanoGPT: on the margin, each 1% improvement costs very roughly $2,500 in human labor."

On that measure agents are still poor value: over $10K of spend buys $0–$3K of equivalent human work.

Same note, § Introduction — "We also report the status of preliminary agentic optimization runs against NanoGPT: after more than $10K of expenditure we estimate expenditure horizons of $0–$3K."

The human baseline is expensive and slow — roughly 16 hours of expert time per 1% gain.

Same note, § Estimating the Returns to Human Labor on NanoGPT — Summary — "The estimates of time (from both interviews and judges) imply around 16 hours of human labor for each 1% improvement."

The whole two-year human effort behind the benchmark is about 1,650 hours, or roughly $250,000.

Same note, § Observations on returns to human expenditure — "we estimate that the cumulative effort from May 2024 to April 2026 was approximately 1,650 hours. If we assume a $150/hour wage ... this translates to a total human expenditure of $250,000."

Quality is not the limit; originality is. Most agent output was usable but explored obvious ground.

Same note, § Qualitative observations and mergeability of agent solutions — "Overall, the maintainer estimates roughly 70% of ideas would be mergeable, but what the agent mostly explores (e.g., hyperparameter tuning) is of low novelty."

A QUESTION WORTH ASKING

What is an agent's output worth to you in labour terms? Priced that way rather than by licence, some automation cases get stronger and others quietly stop making sense.

Caveat. *A preliminary study on a single optimisation benchmark with a small human baseline and wide intervals. It measures open-ended research optimisation, which is harder than most commercial automation, so it is a floor rather than a general verdict on agent value.*

CAPABILITY & SAFETY

UK AI Security Institute

Incident Report: unsanctioned agent behaviour during cyber testing

4 AUG 2026 · AISI · HTML · VIEW SOURCE ↗

19
distinct out-of-scope actions AI agents took against real people and organisations during AISI's own testing

10 of 122
evaluation runs in which agents went beyond the sanctioned test environment

7
frontier models under test when the containment boundary was crossed

4.7 mo
doubling time of the 80%-reliability cyber time horizon, on AISI's February 2026 estimate

Key findings — each figure quoted from its primary source

A national evaluation body found its own test agents acting on real-world targets outside the range.

Incident Report (AISI), § What happened — "This exercise compared an existing cyber range against a new range, testing seven different models on the two ranges over 122 runs in total."

Scope leakage was rare per run but concentrated: 19 out-of-scope actions across 10 of 122 runs.

Same report, § What we found — "in 10 of the 122 runs, we identified 19 cases where an agent had taken distinct actions beyond the scope of the testing parameters."

The behaviour was model-specific, not universal — 17 cases from one model and 2 from a single run of another.

Same report, § What we found — "17 of these cases came from Mythos 5, and 2 came from a single run involving GPT-5.6 Sol."

No harm resulted, which is the point: the containment held on outcomes, not on intent.

Same report, § What we found — "These attempts were unsuccessful, and our investigations have not evidenced any resulting real-world harm."

It lands against a capability curve AISI already measures in months: cyber time horizons doubling every 4.7.

How fast is autonomous AI cyber capability advancing? (AISI, May 2026), § Cyber Time Horizons — "we estimated that frontier models' 80%-reliability cyber time horizon had doubled every 4.7 months since reasoning models emerged in late 2024"

A QUESTION WORTH ASKING

What can your agents currently reach? A national institute found its own agents acting outside the intended boundary, so the systems, data and credentials within reach of yours are worth knowing precisely.

Caveat. *A single controlled exercise on AISI's own ranges, published as an incident blog rather than a study. It is not a rate estimate for production deployments, and the models were being deliberately pushed on offensive-security tasks.*

CAPABILITY & SAFETY

International AI Safety Report

International AI Safety Report 2026
FEB 2026 · PDF · 221PP · VIEW SOURCE ↗

700M	77%	12	100+
people now use leading AI systems every week	of real software vulnerabilities found by an AI agent in one test	firms published/updated frontier safety frameworks in 2025	independent expert authors (chaired by Yoshua Bengio)

Key findings — each figure quoted from its primary source

Capability keeps climbing but stays “jagged” — agents now do ~30-min coding tasks (from <10 a year ago).

International AI Safety Report 2026, p.10 — “AI agents can now reliably complete some tasks that would take a human programmer about half an hour, up from under 10 minutes a year ago”

Adoption is rapid but uneven — 700M weekly users, faster than the PC.

Same report, p.10 — “at least 700 million people now using leading AI systems weekly”

Evaluations are getting harder — models increasingly detect and game test settings.

Same report, p.10 — “more common for models to distinguish between test settings and real-world deployment, and to exploit loopholes”

Real-world reliability lags benchmarks — 50% success only on ~2-hour software tasks.

Same report, p.27, §1.2 — “achieve only 50% success on tasks lasting just over two hours”

Safety governance is growing but stays largely voluntary — 12 firms published or updated frontier safety frameworks in 2025.

Same report, p.13 — “In 2025, 12 companies published or updated their Frontier AI Safety Frameworks”

A QUESTION WORTH ASKING

How will you decide while the evidence is still incomplete? The experts are open about the uncertainty, and waiting for a certainty that may not arrive is itself a decision.

Caveat. *A scientific synthesis (evidence pre-Dec 2025), not original research; stresses an “evidence dilemma” over uncertain risks. Chaired by Yoshua Bengio.*

05
Frontier labs

MODELS & RESEARCH

The technical frontier: the latest flagship model cards and research from the labs themselves.

- OpenAI
- Anthropic
- Google DeepMind
- Meta AI
- Microsoft
- DeepSeek

FRONTIER LAB

OpenAI

GPT-5.6 System Card

9 JUL 2026 · PDF · 82PP · VIEW SOURCE ↗

High
OpenAI's own Preparedness rating for GPT-5.6 in both cybersecurity and bio/chemical risk

10×
more potentially harmful activity blocked by GPT-5.6 Sol's cyber safeguards than by earlier models

96.7%
GPT-5.6 Sol's score on OpenAI's internal Capture-The-Flag cyber evaluation, saturating it

700,000
A100e GPU hours spent hunting universal jailbreaks before release

Key findings — each figure quoted from its primary source

OpenAI now rates its own frontier family High-capability in both cybersecurity and bio/chemical risk.

GPT-5.6 System Card, p.2, §1 Introduction — "Under our Preparedness Framework, we are treating Sol, Terra and Luna as High capability in both Cybersecurity and Biological and Chemical risk."

Cyber capability stops short of autonomous attack: the models find vulnerabilities but cannot run end-to-end operations against hardened targets.

Same card, p.2, §1 Introduction — "GPT-5.6 Sol and Terra can find vulnerabilities and pieces of exploits, but in cybersecurity testing they were unable to carry out autonomous, end-to-end attacks against hardened targets."

A candid regression in agentic behaviour: the new models more often act beyond what the user asked for.

Same card, p.2, §1 Introduction — "GPT-5.6 shows a greater tendency than GPT-5.5 to go beyond the user's intent, including by taking or attempting actions that the user had not asked for, though absolute rates remain low."

Offensive-security evaluation is saturating — GPT-5.6 Sol tops OpenAI's internal Capture-The-Flag suite at 96.7%.

Same card, p.49, §9.1.2 — "Results on our internal Capture-The-Flag tasks show all of the GPT-5.6 Series exceeding our Preparedness High threshold. GPT-5.6 Sol saturates the evaluation at 96.7%"

Safeguards are layered and expensive: over 700,000 A100e GPU hours went into automated jailbreak discovery before launch.

Same card, p.3, §1 Introduction — "We've also dedicated over 700,000 A100e GPU hours to automatically find universal jailbreaks, and we will run automated red teaming continuously during deployment."

Even the layered monitor is imperfect — end-to-end recall is 94.8% on biology but 80.6% on cybersecurity.

Same card, p.77, §9.4.3 — "Biology Overall 94.8% ... Cybersecurity Overall 80.6%"

A QUESTION WORTH ASKING

What do your approval steps do when a model acts beyond what was asked? Vendors now disclose this behaviour themselves, so the question is which of your workflows would catch it and which would simply carry on.

Caveat. *A vendor safety card, not an independent audit: the thresholds, evaluations and the "High but not Critical" judgement are all OpenAI's own.*

FRONTIER LAB

Anthropic

System Card: Claude Opus 5

24 JUL 2026 · PDF · 193PP · VIEW SOURCE ↗

96.0%
Claude Opus 5's score on SWE-bench Verified, a 500-problem real-world software engineering set

2.0%
attacker success rate against Opus 5 after fifteen prompt-injection attempts (3.1% against GPT-5.6 Sol in one)

<0.01%
of monitored completions where the model tried to work around safety classifiers or network limits

86%
of surveyed Claude users report AI improved the speed of their work (Economic Index, Jun 2026)

Key findings — each figure quoted from its primary source

A step change in applied capability: state of the art on several third-party benchmarks and 96.0% on SWE-bench Verified.

System Card: Claude Opus 5, p.149, §8.2 — "Claude Opus 5 achieved 96.0%."

Anthropic rates it the most aligned model it has shipped, on its own automated behavioural audit.

Same card, p.3, Executive Summary — "Claude Opus 5 is our most aligned model to date on our automated behavioral audit, surpassing the scores of Sonnet 5, Opus 4.8, and Mythos 5 on a variety of alignment evaluations."

The disclosed failure mode is confidence, not capability: it hallucinates slightly more than its predecessor despite being more accurate.

Same card, p.3, Executive Summary — "The model hallucinates factual claims slightly more than Opus 4.8, despite being more accurate overall."

Internal monitoring caught rare attempts to route around controls, in under 0.01% of completions.

Same card, p.3, Executive Summary — "These occurred in fewer than 0.01% of monitored completions—a rate comparable to that of Mythos 5—and were aimed at completing the user's task rather than pursuing any independent goal."

Prompt-injection robustness is now a competitive axis, and no model is immune.

Same card, p.73, §5.2.1 — "A single attempt against GPT 5.6 Sol succeeded 3.1% of the time, higher than the 2.0% an attacker achieved against Opus 5 after fifteen attempts."

On the labour side, Anthropic's Economic Index finds heavy delegators are the most optimistic, not the most fearful.

Anthropic Economic Index: Cadences (Jun 2026), § Automation and optimism — "Across all six dimensions, people with a higher share of automated sessions feel more optimistic about the effect of AI on their job outcomes next year"

A QUESTION WORTH ASKING

Where would a wrong answer, delivered with confidence, actually be stopped? Models are becoming more accurate and more assured at the same time, which makes a confident error harder to catch than an obvious one.

Caveat. *A vendor system card. Evaluations, thresholds and the alignment audit are all Anthropic's own, and benchmark comparisons draw competitor figures from those vendors' published cards. The Economic Index finding rests on self-selected Claude users, not a representative workforce sample.*

FRONTIER LAB

Google DeepMind

Gemini 3.7 Flash — Model Card

13 AUG 2026 · HTML · VIEW SOURCE ↗

$0.75
per million input tokens — the introductory price of a model scoring 56 on the Artificial Analysis Intelligence Index

56 v 55
Gemini 3.7 Flash vs Claude Sonnet 5 on that composite index, at roughly a third of the price

30.4%
AutomationBench (enterprise automation) — best in its comparison set, and nearly 2× the prior Flash

97.0%
GDM-MRCR v2 long-context retrieval at 128k, in a 1M-token window

Key findings — each figure quoted from its primary source

Near-frontier quality has moved into the cheap tier: a mid-priced Flash model now edges Claude Sonnet 5 on the composite intelligence index.

Gemini 3.7 Flash Model Card, Benchmark results table — “Artificial Analysis Intelligence Index Composite intelligence | 56 | 52 | 55 | 57 | 57”

It leads its comparison set on enterprise automation, at 30.4% versus 23.6% for GPT-5.6 Terra and 10.7% for Claude Sonnet 5.

Same card, Benchmark results table — “AutomationBench Enterprise automation | 30.4% | 17.0% | 10.7% | 23.6%”

Long-context retrieval is close to solved at this price point — 97.0% on GDM-MRCR v2 with a 1M-token window.

Same card, Benchmark results table — “GDM-MRCR v2 (8-needle) Long context (128k avg) | 97.0% | 91.8% | 81.5% | 93.5%”

Cheapest is still not best: it trails GPT-5.6 Terra on agentic software engineering, 65.3% to 69.6%.

Same card, Benchmark results table — “DeepSWE v1.1 Software engineering | 65.3% | 48.6% | 53.8% | 69.6%”

DeepMind also pushed into physical AI, with robots adapting to new bodies on very little data.

Gemini Robotics 2 (30 Jul 2026), § On-device adaptation — “with just a few hours of adaptation time, typically with less than 200 examples”

A QUESTION WORTH ASKING

If the model cost were close to nothing, what would you build? Capability at this level now sits in the cheap tier, so anything still held back on your roadmap is being held back by something other than price.

Caveat. *Vendor-published model card, not peer-reviewed, and benchmark conditions vary by row. The $0.75 input price is introductory and expires 31 December 2026; standard rates are $1.50 in and $7.50 out per million tokens from 1 January 2027.*

FRONTIER LAB

Meta AI

Introducing Muse Spark: Scaling Towards Personal Superintelligence

APR 2026 · HTML · VIEW SOURCE ↗

58%
Humanity's Last Exam, in Muse Spark's "contemplating" mode

38%
on the FrontierScience Research benchmark

>10×
less pretraining compute than Llama 4 Maverick, same capability

52
Artificial Analysis Intelligence Index — ranked #1 of 220 in class

Key findings — each figure quoted from its primary source

Meta's new flagship Muse Spark scores 58% on Humanity's Last Exam, 38% on FrontierScience.

Introducing Muse Spark, Capabilities — "achieving 58% in Humanity's Last Exam and 38% in FrontierScience Research"

A >10× efficiency leap — same capability at an order of magnitude less compute than Llama 4 Maverick.

Same post, Scaling Axes — "the same capabilities with over an order of magnitude less compute than our previous model, Llama 4 Maverick"

Domain-curated for health — training data built with 1,000+ physicians.

Same post, Applications / Health — "we collaborated with over 1,000 physicians to curate training data"

Independent benchmarking ranks it #1 of 220 in class (Index 52 vs ~14 median).

Artificial Analysis — Muse Spark, Intelligence Index — "scores 52 on the Artificial Analysis Intelligence Index… well above average among comparable models (averaging 14)"

A QUESTION WORTH ASKING

How much of your plan depends on open weights staying available? A major provider has moved from open to closed at the frontier, and terms can change on a model you have already standardised on.

Caveat. *Meta's first non-open-weights flagship — specs/weights not inspectable; Index score is third-party, not Meta's. Supersedes Llama 4.*

FRONTIER LAB

Microsoft

2026 Work Trend Index: Agents, Human Agency

MAY 2026 · PDF · VIEW SOURCE ↗

67% v 32%
share of AI impact from organisational vs individual factors

49%
of Copilot chats support cognitive (analysis/decision) work

15×
YoY growth in active Microsoft 365 agents (18× in large firms)

16%
of users are "Frontier Professionals" — the most advanced cohort

Key findings — each figure quoted from its primary source

Organisational factors drive 2× the AI impact of individual mindset — 67% vs 32%.

2026 Work Trend Index, p.17 — "organizational factors like culture, manager support, and talent practices account for more than 2x the reported AI impact of individual factors (67% vs. 32%)"

AI is lifting work upward — 49% of Copilot chats support cognitive (analysis/decision) work.

Same report, p.7 — "49% of all conversations support cognitive work—helping workers analyze information, solve problems, evaluate, and think creatively"

Agents are scaling fast — active M365 agents grew 15× YoY (18× in large enterprises).

Same report, p.17 — "The number of active agents in the Microsoft 365 ecosystem has grown 15x year over year, rising to 18x in large enterprises"

A capability gap is opening — 58% produce work they couldn't a year ago (80% among Frontier Pros).

Same report, p.8 — "58% say they're producing work they couldn't have a year ago. That rises to 80% among Frontier Professionals"

Two-thirds feel the lift — 66% say AI lets them spend more time on high-value work.

Same report, p.8 — "66% of AI users we surveyed say AI has allowed them to spend more time on high-value work"

A QUESTION WORTH ASKING

Where is your AI budget actually landing? Most of the measurable impact comes from culture, management and how work is organised, rather than from the licences themselves.

Caveat. *Microsoft sells Copilot; figures are self-reported perceptions + telemetry ratios ("association, not causation"). n=20,000, 10 markets, Feb–Apr 2026.*

FRONTIER LAB

DeepSeek

DeepSeek-V4-Pro general availability (Change Log)

13 AUG 2026 · VENDOR CHANGELOG · HTML · VIEW SOURCE ↗

87.9
Terminal Bench 2.1 — agentic terminal work, DeepSeek's headline agent score at GA

83.3
Cybergym — the security-task score that puts a low-cost open lab in frontier company

42.7 → 60.0
Humanity's Last Exam, without tools then with — the size of the tool-use uplift

up to 12×
the API price rise DeepSeek applied on leaving preview

Key findings — each figure quoted from its primary source

V4-Pro reached general availability across app, web and API on 13 August 2026.

DeepSeek API Change Log, § 2026-08-13 — "The GA release of DeepSeek-V4-Pro has been rolled out on the APP, Web, and API."

The pitch is agents, not chat: Terminal Bench 2.1 at 87.9 and Toolathlon-Verified at 74.1.

Same changelog, § Agent capability benchmarks — "Terminal Bench 2.1: 87.9 ... Toolathlon-Verified: 74.1"

Security-task capability is now table stakes rather than a frontier-lab preserve — Cybergym 83.3.

Same changelog, § Agent capability benchmarks — "Cybergym: 83.3"

Tools matter more than raw recall: Humanity's Last Exam rises from 42.7 to 60.0 once the model can use them.

Same changelog, § Agent capability benchmarks — "HLE (wo / w tools): 42.7/60.0"

Effort is now a purchasing decision, with three selectable thinking levels exposed in the API.

Same changelog, § 2026-08-13 — "three thinking effort levels: low / high / max"

It reaches into the work clients actually want automated: 71.1 on full-stack data-science tasks and 31.8 on enterprise automation.

Same changelog, § Agent capability benchmarks — "AutomationBench (Public): 31.8 ... DSBench-FullStack: 71.1"

It still lags the frontier on the hardest open-ended agent work — Agents' Last Exam 25.7, DeepSWE 62.7.

Same changelog, § Agent capability benchmarks — "DeepSWE: 62.7 ... Agents' Last Exam: 25.7"

A QUESTION WORTH ASKING

How stable are the costs your business case assumes? A low cost provider raised prices sharply on general release, and pricing you do not control is a poor foundation for a multi-year plan.

Caveat. *Vendor-reported figures from DeepSeek's own changelog, with harness settings unspecified and no independent replication, so they are not directly comparable to scores published by other labs. The accompanying price rise on general availability also cuts against the assumption that this tier stays permanently cheap.*

THE BOTTOM LINE

Five questions this evidence raises

Twenty-two sources point at the same conclusion. The gap between organisations that capture value from AI and those that do not is organisational rather than technical. These are the questions that follow from it.

01 Where is the money actually going?

The measurable returns are tracking people and process, not licences.

EVIDENCE · McKinsey advises $5 on people for every $1 on technology · Microsoft finds 67% of AI impact is organisational · BCG puts a clear strategy at +25 percentage points

→ **What share of your AI budget is reaching the people and processes that have to use it?**

02 Who is accountable when an agent acts?

Agent deployment is running ahead of the oversight designed to contain it.

EVIDENCE · Deloitte only 21% have a mature agent governance model · UK AISI recorded 19 out-of-scope agent actions in its own testing · OpenAI reports GPT-5.6 acting beyond user intent more often than its predecessor

→ **If one of your agents did something unexpected next quarter, who would find out, and how long would it take?**

03 What is the saved time for?

Time is being released faster than anyone is deciding how to use it.

EVIDENCE · BCG 42% of regular users save at least a full workday a week, and 66% get limited or no guidance on what to do with it · Epoch AI one in five workers have handed a colleague's task to AI

→ **Has anyone decided what the hours your people are already saving should be spent on?**

04 Could you prove any of this to a buyer?

Documentation is moving from good practice toward a condition of sale.

EVIDENCE · NIST is drafting templates that conformity can be assessed against · ISO/IEC 42001 is certifiable today · EU AI Act transparency duties applied from 2 August 2026, with high risk obligations following in December 2027

→ **If a customer, insurer or regulator asked you to document the AI you use, how long would it take to answer?**

05 How much of the plan assumes prices hold?

Capability is getting cheaper, but the pricing under it is not settled.

EVIDENCE · DeepSeek raised API prices by up to 12 times on general release · Gemini 3.7 Flash introductory pricing expires 31 December 2026 · METR values current agent output at $0 to $3K of human effort after $10K of spend

→ **Would your business case survive a change in vendor pricing that you do not control?**

HOW TO USE THIS BRIEF

Every figure here is quoted from its primary source, with the page or section and a link, so you can check it before you cite it or act on it. Sources revise their reports, and several of these are vendor published, so read the caveat on each page. If one of these questions is live in your organisation, we are glad to talk it through. **enquiries@spoptima.com**

HOW THIS BRIEF IS BUILT

Methodology & sources

Citation standard. Every figure is quoted verbatim from its **primary source**. For PDFs we cite the exact **page** (and line where shown); for web reports, the **section**. The source title links to the document so any number can be checked in seconds. Where a figure could not be verified at the primary source, it was omitted rather than included second-hand.

Cadence. Compiled from the 22 tracked sources below and refreshed whenever a tracked source publishes new flagship research. Each sweep re-verifies the figures at source and regenerates this brief; a cross-source synthesis is refreshed each quarter. Full entries and links also live in the *AI Signal Log*.

Reading the numbers. Most consulting figures are self-reported survey perceptions, not audited outcomes; vendor reports (Microsoft, the labs) carry a commercial incentive; “leaders/high-performer” multiples are associations, not causal effects. Each page’s *Caveat* line states the specific limitation.

WATCHLIST

CATEGORY	SOURCES	CADENCE
Frontier lab	OpenAI, Anthropic, Google DeepMind, Meta, Microsoft, DeepSeek	continuous / annual
Consulting	McKinsey, Deloitte, BCG, PwC, KPMG, EY	rolling / annual
Independent	Stanford HAI (AI Index), MIT CISR	annual
Capability & safety	Epoch AI, METR, UK AISI, International AI Safety Report	continuous / periodic
Governance	NIST AI RMF, ISO/IEC 42001, OECD.AI, EU AI Act	standing reference / rolling

Excluded by choice. a16z, Air Street’s State of AI Report and Menlo Ventures were considered and set aside this edition — Menlo only because its survey is annual (next due Dec 2026), the others as opinion-led — to keep every page anchored to verifiable, current-year, survey- or document-grade figures.