THE STATE OF AI · QUARTERLY EVIDENCE REVIEW · EDITION 2

The State of AI: September 2026 Update

A time-series evidence review for operational leaders

Perth AI Consulting

Version 1.0 (fact-checked) · Evidence as at 2 September 2026 · Research window 14 June to 2 September 2026

Perth WA · September 2026

Follows The State of AI in Mid-2026 (v1.1, evidence as at 13 June 2026). Permalink: perthaiconsulting.com.au/resources/state-of-ai/mid-2026/

> Drafted with Anthropic's Claude Fable 5.1 coordinating seven research streams; independently fact-checked by a separate Claude Opus pass that traced every cited source (118 findings, 13 refuted, all corrected and logged in Appendix C); human-reviewed before publication. Every claim is dated, vendor claims are labelled, and carried-forward judgements are cited to the previous edition.

Series metadata

FIELD	VALUE
Series	The State of AI (quarterly evidence review)
Edition	September 2026
Edition number	2
Evidence as at	2 September 2026 (research completed 2 September 2026)
Version	1.0 (fact-checked)
Research window	14 June 2026 to 2 September 2026
Previous edition	The State of AI in Mid-2026 (v1.1, evidence as at 13 June 2026, corrections applied 25 July 2026). Permalink: https://perthaiconsulting.com.au/resources/state-of-ai/mid-2026/ · PDF: https://perthaiconsulting.com.au/resources/state-of-ai-mid-2026-literature-review-v1.1.pdf
Next edition	Planned for December 2026 (evidence window 3 September to early December 2026)
Structural change since previous edition	First update edition. Introduces the fixed frame, the movement register, the scorecard of the previous edition's predictions, and the standing watchlist (Appendix B carried forward). The ten capability sections, four judgements, and four postures are unchanged in name, number, and order.

How to cite this edition. Perth AI Consulting. (2026). The State of AI: September 2026 update. A time-series evidence review for operational leaders (Version 1.0). https://perthaiconsulting.com.au/resources/state-of-ai/september-2026/

How to cite the previous edition. Perth AI Consulting. (2026). The State of AI in mid-2026: A literature review for operational leaders (Version 1.1). https://perthaiconsulting.com.au/resources/state-of-ai/mid-2026/

Movement register: June 2026 to September 2026

The previous edition closed each of ten capability sections with four judgements: what is deployable today, where demonstrations outrun production, what is oversold, and what is underestimated. This register records how each of those forty judgements moved between 13 June 2026 and 2 September 2026. The classifications are defined in §1.4 and Appendix D; in brief, **improved** means the evidence moved in the reader's favour (more deployable, a narrower gap, an oversold claim gaining independent support, an underestimated capability moving further along), **declined** means the evidence moved against the reader (less reliable, a wider gap, a claim more oversold than June judged, a capability less far along), **unchanged** means the June judgement stands (with a note on whether new evidence confirmed it or no evidence of the

required standard appeared), and **reframed** means the judgement stands but the reason or boundary changed.

§	SECTION	DEPLOY TODAY	DEMO VERSUS PRODUCTION GAP	OVERSOLD	UNDERESTIMATED
3.1	Large language and reasoning models	Unchanged (confirmed): three closed-lab flagship-class releases (GPT-5.6, Opus 5, Fable 5.1) and two open-weight releases (Kimi K3, DeepSeek V4-Flash) shipped; none changed what is dependable with review.	**Declined**: OpenAI's own 8 July finding that about 30 per cent of SWE-bench Pro tasks are broken widens the benchmark-to-production gap on the field's contamination-resistant coding benchmark.	Unchanged (confirmed): no new independent hallucination measurement; frontier labs now publicly dispute each other's benchmark margins.	**Improved**: DeepSeek V4-Flash (MIT licence, one-million-token context, US$0.14/0.28 per million tokens; vendor-disclosed specifications and pricing) and Kimi K3 open weights extend the open-weight case, against Meta's exit and a tighter Qwen licence.
3.2	Agentic systems and tool use	Unchanged (confirmed): ChatGPT Work (9 July) ships with human approval gates; Anthropic's Cowork telemetry (fieldwork late May, pre-window) shows agents used overwhelmingly for bounded office work.	Unchanged (confirmed): no rerun of TheAgentCompany; METR's 21 July expenditure-horizon study found autonomous agents had "minimal effect" on AI research progress.	**Reframed**: the Ninth Circuit vacated the Comet injunction on 4 August, narrowing one legal theory against agent makers; the reliability and security judgement is unchanged.	**Improved**: the MCP specification's 2026-07-28 revision became the current protocol version, with a feature lifecycle policy (twelve months' notice before removal) and Tasks, Skills over MCP, and MCP Apps extensions.
3.3	Vision and multimodal models	Unchanged (no new evidence): June 2026 judgement carried forward.	Unchanged (no new evidence): no rerun of the OmniAI real-document benchmark; the roughly 75 per cent field-accuracy figure stands.	Unchanged (confirmed): again no independent evaluation of any vision-against-standards product was found.	Unchanged (no new evidence): June 2026 judgement carried forward.

§	SECTION	DEPLOY TODAY	DEMO VERSUS PRODUCTION GAP	OVERSOLD	UNDERESTIMATED
3.4	Video and image generation	Unchanged (confirmed): Seedance 2.5 (30-second single pass) and Grok Imagine Video 1.5 extend capability without changing the deployable boundary.	Unchanged (no new evidence): no independent stress test in the window; duration claims are vendor claims.	**Declined**: the Sora API sunset (24 September 2026) holds, Google shut down Veo 2.0 and Veo 3.0 (30 June) and Imagen 4 (17 August), and the Higgsfield pipeline removed Sora; product durability is worse than June judged.	Unchanged (no new evidence): one new per-second price point (Grok Imagine, vendor) is not a trend measurement; June judgement carried forward.
3.5	Audio, voice, and music generation	Unchanged (no new evidence): June 2026 judgement carried forward.	Unchanged (no new evidence): still no independent cross-platform voice-agent evaluation; the scarcity persists.	Unchanged (confirmed): Bland's Series C and ElevenLabs' reported US$500 million ARR are commercial, not evaluative, evidence.	Unchanged (no new evidence): no primary voice-cloning loss data appeared; ACCC's June release carries no breakout.
3.6	Knowledge management and synthesis	Unchanged (confirmed): Anthropic's 25 August memory merge and Google's June NotebookLM upgrade (pre-window) extend the deployable grounded pattern.	Unchanged (no new evidence): no new dated report on enterprise search governance.	**Declined**: a peer-reviewed taxonomy (4 July) catalogues 33 RAG failure modes, 12 without empirical study, including all eight agentic modes; a purpose-built hallucination detector fails to transfer across task types. Grounded hallucination is less bounded than June implied.	Unchanged (no new evidence): June 2026 judgement carried forward.

§	SECTION	DEPLOY TODAY	DEMO VERSUS PRODUCTION GAP	OVERSOLD	UNDERESTIMATED
3.7	Code generation and software creation	Unchanged (confirmed): Claude Code, Cursor, GPT-5.6, and Copilot continued to ship inside review workflows.	Unchanged (no new evidence): no new vibe-coding security scan; the Stack Overflow 2026 survey had not published results.	**Declined**: benchmark-derived capability claims are now contested by the labs themselves (the SWE-bench Pro finding); benchmark margins are less usable than in June.	Unchanged (no new evidence): price falls are directionally consistent; no study measured the effect.
3.8	Business automation platforms with AI built in	Unchanged (no new evidence): June 2026 judgement carried forward.	Unchanged (no new evidence): no independent resolution-rate measurement in the window.	Unchanged (confirmed): Salesforce reported Agentforce ARR above US$1.5 billion (26 August) with no customer count, continuing the pattern June described.	Unchanged (confirmed): outcome pricing stands as the buyer's tool; no new move either way.
3.9	Crypto and AI convergence	Unchanged (confirmed): still a watch category; the one machine-payment rail with hard numbers (x402) fell 93 per cent in daily volume year to date.	**Reframed**: Visa Intelligent Commerce went live with 30-plus European issuers (2 July, vendor-announced) but published no volume; no successor to Instant Checkout shipped.	**Declined**: x402 daily settlement fell 93 per cent year to date, to roughly US$28,000 to 42,000 a day in August from a late-2025 peak that repeatedly approached US$800,000; "the agent economy is here" is more oversold than June judged.	**Improved**: the rails kept standardising (Visa in production; US Treasury GENIUS Act rulemaking 18 August) and NEAR's confidential routing reached roughly US$30 million TVL.

§	SECTION	DEPLOY TODAY	DEMO VERSUS PRODUCTION GAP	OVERSOLD	UNDERESTIMATED
3.10	Vertical AI applications	Unchanged (confirmed): Xero reported 100 million-plus auto-reconciled transactions (vendor); Australian regulators kept publishing rules.	Unchanged (no new evidence): no new controlled study of time saved in any vertical.	Unchanged (confirmed): the hallucination-cases database passed 2,000 entries (Australia 110); "hallucination-free" remains contradicted by the incident record.	**Improved**: the Fair Work Commission issued AI-disclosure rules effective 20 October, the TGA listed software as a medical device among its twelve compliance priorities for 2026-27, and the OAIC updated its facial-recognition guidance after the Bunnings decision; the regulatory floor is more specific than in June.

Summary count: 4 improved · 5 declined · 2 reframed · 29 unchanged (14 confirmed by new evidence, 15 with no new evidence of the required standard). Eleven of forty cells moved; nine of those moved directionally.

Reading the register. Eleven of forty cells moved, nine of them directionally. That is the headline of this edition: in a quarter that shipped three closed-lab flagship-class models, a new current MCP specification, a formally adopted EU AI Act amendment, and a fourth major-lab product shutdown, the evidence changed what a prudent business should do in almost none of the ten categories. The movements cluster in two places. The improved cells are infrastructure and regulation (open weights, MCP, payment rails, Australian regulators), which is where the field is compounding. The declined cells are measurement and durability (benchmark integrity, grounded hallucination, product shutdowns, machine-payment volume), which is where June's caution was, if anything, too mild.

What the Mid-2026 edition got wrong, and what has aged

A time series is only worth quoting if it scores itself. The previous edition made a set of dated predictions in its §2.4 and closed with a specific claim about which of its sections would age fastest. Both are scored here, item by item.

§2.4 scorecard: announced versus speculative

JUNE 2026 STATEMENT	STATUS AT 2 SEPTEMBER 2026	EVIDENCE
EU provisional agreement (May; the baseline dated it 6 and 13 May, the Commission's page gives 7 May) to postpone AI Act high-risk obligations to December 2027 (Annex III) and August 2028 (Annex I); "provisional, not yet formally adopted"; transparency rules still commence August 2026.	**Shipped as announced.** The AI Omnibus Regulation was formally adopted and entered into force on 27 July 2026 with those dates; Article 50 transparency obligations commenced on schedule on 2 August 2026, with a transitional period to 2 December 2026 for machine-readable marking on existing generative systems.	European Commission (2026a, 2026b); Cooley (2026); Morgan Lewis (2026)
Australia: National AI Plan (December 2025) shelved the proposed mandatory guardrails in favour of existing law, sector regulators, and a new Australian AI Safety Institute; Privacy Act automated-decision-making transparency obligations commence 10 December 2026.	**Still open on the Institute; on track on the Privacy Act.** The 10 December 2026 date is unchanged. The OAIC's consultation on transparency guidance for automated decision-making closed on 15 June 2026; no draft or final guidance had been published by 2 September. The Institute's operational status could not be verified in the window (Appendix B, new item N1).	OAIC (2026a, 2026b)
United States: December 2025 executive order and March 2026 White House framework push toward federal preemption of state AI laws; Congress had enacted nothing.	**Did not ship.** A bipartisan discussion draft (the Great American Artificial Intelligence Act of 2026, 3 June 2026) proposes a three-year preemption; nothing was enacted by 2 September. Colorado had already repealed and replaced its AI Act with a narrower Automated Decision-Making Technology Act (signed 14 May 2026, effective 1 January 2027), a state-side retreat the June edition did not record.	Nextgov (2026); Skadden (2026)
Economics stress-tested in public: Anthropic confidentially filed for an IPO (reported US$965 billion valuation); an OpenAI listing widely expected; public listings would expose frontier-lab economics to audited scrutiny.	**Shipped differently; scrutiny not yet realised.** No public S-1 and no listing occurred in the window. Press reported Anthropic's run-rate at about US$65 billion at end-July (Bloomberg-sourced; not confirmed by the company) and an IPO "possibly as soon as this fall" at a reported US$2 trillion. Microsoft's FY2026 accounts, meanwhile, recorded net gains of US$4.96 billion on its OpenAI investments for the year, which complicates the "losses inferred from Microsoft filings" framing June used (see Appendix B item 28).	TechCrunch (2026c); Axios (2026b); Microsoft (2026)
Agent infrastructure consolidates: MCP, A2A, and the payment protocols (AP2, ACP, x402) under neutral foundations; compute supply keeps growing.	**Shipped as announced.** The MCP specification's 2026-07-28 revision became the current protocol version under the Agentic AI Foundation, with a formal feature lifecycle and an extensions framework. Visa's agent-payment programme went live in Europe on 2 July (reported 6 July). Hyperscaler capital expenditure guidance rose again in the July earnings round (§4).	Model Context Protocol (2026); Agentic AI Foundation (2026); Industry Spread (2026)

JUNE 2026 STATEMENT	STATUS AT 2 SEPTEMBER 2026	EVIDENCE
Discount: "specific names and dates for next-generation models circulate almost entirely on low-credibility blogs; none of the frontier labs had formally announced its next flagship as of 13 June 2026." Disciplined base case: "continued incremental frontier releases… no scheduled discontinuity."	**Base case held; one named model was real.** OpenAI shipped GPT-5.6 (Sol, Terra, Luna) on 9 July 2026, which June had correctly listed as anticipated rather than fabricated (Appendix B item 2). Anthropic shipped Opus 5 (24 July) and Fable 5.1 and Mythos 5.1 (1 September) as mid-cycle releases, not a new tier. Google shipped three Flash-tier Gemini models on 21 July and did not ship a 3.5 Pro flagship, which Bloomberg attributed to internal delays. No discontinuity.	OpenAI (2026a); Axios (2026a); Anthropic (2026); TechCrunch (2026b)
Discount: "claims of recursive self-improvement on an 18-month horizon are speculation."	**Held, now with independent evidence.** METR's expenditure-horizon study (21 July 2026) found that autonomous agent optimisation "has so far had minimal effect on AI R&D progress" on the NanoGPT speedrun, with newer models matching human returns only within budgets of roughly US$0 to 3,300.	METR (2026)

The fastest-ageing prediction

June's Conclusion said the sections "most likely to age fastest, agent reliability and cost, are the ones worth re-checking first." Half right. **Cost aged as predicted**: DeepSeek V4-Flash shipped on 31 July at US$0.14 per million input tokens with a one-million-token context under an MIT licence (vendor-disclosed); Anthropic cut cache-read pricing 75 per cent on 1 September; OpenAI split GPT-5.6 into three price tiers from US$1 to US$5 per million input tokens. **Agent reliability did not age in the way predicted.** The independent numbers June relied on (TheAgentCompany's 30 per cent, METR's task-horizon doubling) were not superseded, because no comparable independent evaluation of current agents was published in the window; what changed was the legal and product landscape around agents (a vacated injunction, a shut-down browser, a folded research prototype). The lesson for the series is that "fastest-ageing" has two modes, evidence ageing and landscape ageing, and the register should record which one moved.

Corrections to the June edition's picture of the world

Three things the June edition stated as current were already, or have since become, wrong, and a fourth is a figure that could not be re-verified.

1. **Google's Project Mariner.** June §3.2 listed Mariner among products Google "ships." Press reports place the retirement of the standalone Mariner prototype around 4 to 7 May 2026, before June's research date, with the capability folded into Gemini's agent mode. This edition treats Mariner as discontinued; the primary Google announcement was not located and the correction is logged for the fact-check (Appendix B, new item N2).
2. **OpenAI's Atlas browser.** June §3.2 described Atlas as a live product (launched October 2025). OpenAI announced on 9 July 2026 that it would sunset Atlas, and shut it down on 9 August 2026, less than ten months after launch, redistributing the agentic browsing capability into a ChatGPT Chrome extension, a browser inside the ChatGPT desktop app, and

a remote cloud browser for agents (TechCrunch, 2026a; 36Kr, 2026). This is the fourth flagship shutdown from a best-resourced lab in about a year, after Sora, Operator, and Instant Checkout (June counted the first three inside nine months).

3. **Australian SMB adoption figures.** June cited the National AI Centre's 40-to-69 per cent "regular use" and 9-to-28 per cent "daily use" figures and, in its Appendix B item 29, recorded them as confirmed against the NAIC's December 2025 to February 2026 summary. The NAIC's published dataset for that same fieldwork (7 May 2026) reports 43 to 44 per cent of SMEs with "some level of AI adoption". One dataset is therefore being reported two incompatible ways, and until the instrument is examined neither figure should be quoted as a time series (Appendix B item 29).
4. **OSWorld figures.** June cited frontier results "around 70 per cent on standard OSWorld and roughly 78 to 80 per cent on OSWorld-Verified (XLANG, June 2026)". The only primary XLANG page that could be fetched in this window was date-stamped 28 July 2025 and shows materially lower scores, and Anthropic's 1 September release notes now report against a new "OSWorld 2.0 (strict)" variant at 41.7 per cent. The June figures could not be re-verified against a live primary source and are carried with that caveat (Appendix B item 9).

Baseline-period developments the June edition missed

These fall before 14 June 2026 and are not window evidence. They are listed so the series does not silently absorb them: Mastercard's Agent Pay for Machines (10 June 2026, a machine-to-machine payment product with 30-plus partners); the Visa and OpenAI agent-payment integration announcement (12 June 2026, no launch date); Ideogram 4.0 (3 June 2026, open weights, text rendering); Google's NotebookLM reasoning and research upgrade (8 June 2026); MYOB's AI BAS and Smart Reconciliation suite (2 March 2026; vendor announcement), which puts a second Australian accounting platform alongside Xero in §3.10; the Fair Work Commission's Wibmer v Fujifilm decision (9 June 2026) on AI-drafted workplace complaints; and the Australian Government's own whole-of-government Microsoft 365 Copilot trial evaluation, which the June edition did not cite alongside the UK evidence (its publication date could not be confirmed from the fetched page; see Appendix B, new item N3).

Watchlist: what happened to the thirty claims June could not verify

June's Appendix B listed thirty circulating claims that could not be verified to the paper's standard. Every item was revisited in this window. June's own second fact-check pass had already resolved many of them in place (twelve promoted into the body and removed, three refuted, and fourteen of the thirty retained with a resolved or partly resolved status and a sharpened reason), so the list this edition inherited was a mix of open questions and closed ones kept for the record. In this window: **two moved to verified** (item 2, GPT-5.6 exists and shipped; item 27, REA Group's FY26 results name specific AI features shipped in the year); **one gained a**

complication (item 28: Microsoft's FY2026 accounts record net gains, not losses, on OpenAI); **two closed items are retired as no longer relevant** (item 13, Sora clip length, resolved by June and now moot with Sora discontinued; item 3, Llama 4 Behemoth, resolved by June as not shipped and now moot with Meta's exit from open weights); and **twenty-five are unchanged in status**, most of them because the independent evidence they need does not exist rather than because it was hidden. No item changed from unverifiable to refuted in the window (June's Pass 2 had already refuted item 10). The full item-by-item disposition is in Appendix B, which also opens **eight new items** (N1 to N8) for claims this edition encountered and could not verify.

Abstract

This is the second edition of Perth AI Consulting's quarterly evidence review of applied artificial intelligence, and the first update. It covers the evidence window from 14 June 2026 to 2 September 2026, holds the previous edition's frame fixed (ten capability categories, four judgements each, four adoption postures), and records how each of the forty judgements moved. The reader who has not seen the Mid-2026 edition will find every judgement restated in full, with the June evidence carried forward and cited, so this document stands on its own.

The central finding is stability with sharp local movement. Eleven of forty judgements moved, nine of them directionally. Four improved, all in infrastructure and regulation: open-weight models (DeepSeek V4-Flash, MIT-licensed, one-million-token context, at about 36 times lower input price than GPT-5.6 Sol on vendor-disclosed pricing), the Model Context Protocol (2026-07-28 became the current protocol version, with a formal lifecycle), agent-payment rails (Visa live in Europe; US Treasury stablecoin rulemaking), and Australia's regulatory floor (Fair Work Commission AI-disclosure rules from 20 October 2026; software as a medical device among the TGA's twelve compliance priorities for 2026-27). Five declined, all in measurement and durability: OpenAI's own 8 July 2026 analysis estimates that about 30 per cent of SWE-bench Pro tasks, the field's contamination-resistant coding benchmark, are broken, which also makes benchmark-derived coding claims less usable; a peer-reviewed taxonomy catalogues 33 failure modes in retrieval-augmented systems, twelve of them never empirically studied; OpenAI shut down its Atlas browser less than ten months after launch and Google retired five video and image models, the fourth flagship withdrawal in about a year; and daily volume on the x402 machine-payment rail fell 93 per cent year to date.

The two macro gaps June named are both confirmed by new data. McKinsey's State of AI survey (fieldwork 4 May to 8 June 2026, 1,719 respondents, published 25 August) found 80 per cent of individual users reporting productivity gains while the share of firms attributing at least some earnings impact to AI stayed flat at 37 per cent and "high performers" stayed at about 6 per cent (consultancy self-report). Stanford's Digital Economy Lab (12 August 2026, ADP payroll data through June 2026) found employment of 22-to-25-year-olds in the most AI-exposed occupations now about 19 per cent below where it would otherwise be, from 15 per cent on

Stanford's own prior estimate (the June edition carried 13 per cent), still operating through reduced hiring rather than separations. The economics overhang deepened: Alphabet and Amazon raised 2026 capital expenditure guidance in July, Meta narrowed its range upward, and Microsoft reported US$115.9 billion of property and equipment additions for its fiscal year.

As at September 2026, the adoption posture is unchanged in substance and sharper in one respect: the benchmark chart in the sales deck is now disputed by the labs themselves, so "which benchmark, measured by whom, on which subset" is no longer a specialist's question but the first thing a buyer should ask.

---

1. Introduction

1.1 Purpose

The purpose is unchanged from the previous edition: the gap between what AI vendors claim and what their products do in production is a material business risk in both directions, and better dated information reduces it. What is new is the time dimension. A single survey tells a reader where the field stands; a series tells them which way it is moving and how fast, and lets them see where the previous survey was wrong. This edition is written to be quoted on both counts.

1.2 Scope

The scope is the same ten capability categories, international with Australian context noted where material, covering capability and reliability rather than ethics, safety policy, or governance debates. The reader is the same: the owner of a mid-sized construction firm, the principal of a medical practice, the partner of a suburban law or accounting firm, the financial adviser, and the consultants who advise them.

The evidence window is 14 June 2026 to 2 September 2026. The previous edition's research closed on 13 June 2026; starting the day after leaves no gap. Anything the previous edition already covered is baseline evidence, even where it was re-reported in the window, and is labelled as such.

1.3 Methodology

The review was researched on 2 September 2026 through seven parallel research streams (frontier models and code; agents and business automation; vision and the medical, property, and construction verticals; video, image, audio, and music; knowledge management and the professional verticals; crypto and AI convergence; and a cross-cutting stream covering the previous edition's predictions, macroeconomics, employment, and Australian policy). Each stream ran between 21 and 35 distinct searches (a further 12 planned queries could not be run), fetched primary pages for anything load-bearing, and reported explicitly where searches

returned nothing meeting the source standard. Those nulls are the evidential basis for every "unchanged (no new evidence)" classification and are listed by query in Appendix A.

The three rules of the previous edition apply throughout: every capability claim is dated; vendor claims are labelled as vendor claims; and what could not be verified is flagged rather than omitted or asserted. A fourth rule is added for update editions: **carried-forward judgements are cited to the previous edition and to its primary sources, never restated as new.** Where this edition repeats a June figure, the sentence says so.

One limitation specific to this run must be stated plainly. The research session's web-search budget was exhausted before every stream finished, and several questions were closed by direct fetches of regulator and vendor pages rather than by search. The streams affected are listed in Appendix A with the queries that could not be run. Where that weakens a null result, the register says "no new evidence" rather than "confirmed", and the reader should weight it accordingly.

1.4 How to read a time-series edition

Each capability section has the same five parts: the June 2026 judgement (cited, with enough of the research detail to stand alone); what happened in the window (dated, source-labelled); a four-cell movement table; the updated September 2026 judgement, restated in full; and one self-contained, dated sentence that states the section's headline.

The movement classification is a claim like any other and carries a citation. Its meaning depends on the column. In the **deploy today** column, improved means more is dependable and declined means something dependable was withdrawn or shown unreliable. In the **demo-versus-production gap** column, improved means independent evidence narrowed the gap and declined means it widened. In the **oversold** column, improved means the oversold claim gained independent support and declined means further evidence that it is oversold. In the **underestimated** column, improved means further evidence the capability is further along than assumed and declined means it is less far along. The convention is that improved is always the reader's favour and declined is always more caution. **Unchanged** carries one of two notes: confirmed, where new evidence supports the June judgement, or no new evidence, where nothing of the required standard appeared and the June judgement is carried forward on its original evidence. **Reframed** means the judgement stands but the reason or the boundary of what counts moved.

2. The temporal frame

2.1 to 2.2 The pre-AI baseline and early AI, 2022 to 2024

Unchanged from the previous edition, which the reader should consult for the full account (Perth AI Consulting, 2026, §2.1 to 2.2). In one paragraph: the randomised evidence from 2023 found real individual gains (12.2 per cent more tasks and 25 per cent faster in the BCG experiment; a 55.8 per cent speed-up on one scoped coding task) with the largest gains accruing to less-experienced workers, and no firm-level transformation followed.

2.3 Now, from September 2026: the agentic turn, one quarter on

June characterised 2025 to mid-2026 as "the agentic turn and the evidence gap": agent products shipping faster than independent evaluation could assess them. One quarter on, the products kept shipping and the evaluation gap did not close. OpenAI launched ChatGPT Work on 9 July 2026, an agent layer that "takes an outcome, gathers information across connected apps and workflows, breaks the job into smaller steps, and completes them independently" with human approval gates, and on the same day announced it would shut its ten-month-old Atlas browser (TechCrunch, 2026a; InfoWorld, 2026). Anthropic extended Cowork to mobile and web from 8 July 2026 and published usage data from 1.2 million sessions in the last two weeks of May: business-process operations 33.4 per cent, content creation 16.4 per cent, software development 8.7 per cent (vendor telemetry, pre-window fieldwork, reported by TechCrunch, 2026d). Microsoft's Windows Agent Runtime, announced at Build in June, remained in preview through the window (Rework, 2026).

The independent evidence that did arrive points the same way as June. METR's expenditure-horizon study (21 July 2026) is the most important: it asks at what dollar budget an autonomous agent's improvement to a real research target (the NanoGPT training speedrun) equals what a human would achieve with the same money, and found horizons of US$0 to 3,300 across six high-expenditure runs, with GPT-5 and Opus 4.1 making no meaningful progress, GPT-5.5 and Opus 4.8 making modest progress at roughly US$2,300 and US$3,300, each run extended to about US$10,000 of spend, and only 50 to 70 per cent of agent contributions judged mergeable by the project maintainer (METR, 2026). No rerun of TheAgentCompany with current models appeared (Appendix B item 8).

The evidence-base pollution June described has a new form: the labs now dispute each other's benchmarks in public. On 8 July 2026, the day before releasing GPT-5.6, OpenAI published an analysis estimating that "~30% of the SWE-bench Pro tasks are broken", with an automated pipeline flagging 27.4 per cent and five human reviewers flagging 34.1 per cent (OpenAI, 2026b). Independent commentary noted the timing: Anthropic's Fable 5 had scored 80 per cent on SWE-bench Pro against GPT-5.6 Sol's 64.6 per cent (Willison, 2026a). Whatever the motive, the finding stands: the benchmark the June edition recommended as contamination-resistant is now, on one lab's evidence, substantially broken.

2.4 The next 12 to 24 months: announced versus speculative, restated for December 2026

These are the statements the December 2026 edition will score.

Scheduled or highly likely. The Privacy Act's automated-decision-making transparency obligations commence on 10 December 2026; OAIC guidance is expected before then but had not appeared by 2 September (OAIC, 2026a). The EU AI Act's transitional period for machine-readable marking of existing generative systems ends on 2 December 2026 (Morgan Lewis, 2026; law-firm summary, not independently fetched). The Sora API is discontinued on 24 September 2026 (OpenAI, 2026c). The Fair Work Commission's AI-disclosure requirements take effect on 20 October 2026 (Fair Work Commission, 2026). US Treasury's GENIUS Act stablecoin rulemaking closes its comment period on 19 October 2026 (Federal Register, 2026). Anthropic's IPO is reported as possible "as soon as this fall"; if it lists, the first audited frontier-lab accounts become public (Axios, 2026b). Google is expected to ship a Gemini 3.5 Pro flagship after missing it in July (TechCrunch, 2026b).

Plausible but unscheduled. A Colorado-style narrowing of other US state AI laws under federal pressure; a first ASIC action over AI-drafted advice documents (none by 2 September); an independent rerun of an office-task agent benchmark with 2026 models (none has appeared in eighteen months); Deloitte's or another consultancy's 2026-fieldwork enterprise survey.

To be discounted. Named next-generation flagships from any lab beyond the ones above (Gemini 4 pretraining is reported as under way, which is not a release); any transaction-volume claim for agentic commerce not backed by a published figure; and recursive self-improvement claims, now against METR's evidence rather than merely without it.

The disciplined base case for December 2026 is June's with two additions marked: continued incremental frontier releases (now: at falling cost), agents with somewhat longer horizons and somewhat better reliability inside supervised workflows, deeper vertical products, (now:) more regulatory specificity in Australia, and no scheduled discontinuity.

---

3. Capability survey

3.1 Large language and reasoning models

June 2026 judgement. Three labs defined the closed frontier (Claude Opus 4.8 and Fable 5, GPT-5.5, Gemini 3.1 Pro); million-token context had become standard; ARC-AGI-2 scores had passed the human average; and the credible open-weight frontier was Chinese and European (DeepSeek V4, Mistral Large 3, Qwen), with every 2026 flagship a sparse mixture-of-experts. Hallucination was not solved (Stanford RegLab's 17 to 33 per cent in legal tools marketed as hallucination-free), MMLU-Pro was saturated above 89 per cent (Epoch AI, April 2026), SWE-

bench Verified was suspected of contamination, and the contamination-resistant SWE-bench Pro sat "near 80 per cent" for the best models on aggregator-reported figures. Judgement: deploy text and document work with review, structured extraction with validation, and coding assistance; distrust anything quoted in benchmark points; treat "hallucination-free" and headline benchmark margins as oversold; and treat open-weight models as underestimated (Perth AI Consulting, 2026, §3.1).

What happened in the window. Three closed-lab flagship-class releases, one delay, and a busy open-weight quarter. OpenAI released the GPT-5.6 family on 9 July 2026 in three tiers, Sol (US$5/30 per million tokens), Terra (US$2.50/15), and Luna (US$1/6), with vendor-claimed leads on several agentic benchmarks (OpenAI, 2026a). Anthropic released Claude Opus 5 on 24 July 2026 at unchanged Opus pricing (US$5/25), positioned by Anthropic as approaching Fable 5 at half the price (Axios, 2026a; vendor positioning), and Claude Fable 5.1 and Mythos 5.1 on 1 September 2026 at unchanged Fable pricing (US$10/50) with cache reads cut 75 per cent to US$0.25 per million; the vendor-reported gains over Fable 5 include Terminal-Bench 4.0 at 55.8 per cent (from 42.0) and Humanity's Last Exam at 60.9 per cent without tools (Anthropic, 2026; vendor claims, no independent replication at the time of writing). Google released Gemini 3.6 Flash, 3.5 Flash-Lite, and a government-only 3.5 Flash Cyber on 21 July 2026 and did not release Gemini 3.5 Pro, which Bloomberg attributed to internal delays; Google said work on a Gemini 4 pretraining run had begun (TechCrunch, 2026b).

The open-weight frontier moved faster than the closed one on cost. DeepSeek open-sourced V4-Flash under the MIT licence on 31 July 2026: 284 billion total parameters, 13 billion active, a one-million-token context, priced at US$0.14 per million input tokens and US$0.28 output, about 36 times below GPT-5.6 Sol's input price (Open Source For You, 2026; specifications and pricing as disclosed by the vendor). Moonshot's Kimi K3 (announced 16 July 2026, 2.8 trillion parameters by the vendor's account, open weights promised for 27 July) was rated by the independent evaluator Artificial Analysis at 1,547 Elo on its private evaluation (Willison, 2026b). Zhipu's GLM-5.3 shipped on 14 August with weights to follow (MLQ, 2026). Two counter-signals: Meta's Muse Spark 1.1 (9 July 2026), Meta's first paid, API-only frontier model, confirmed its exit from open-weight frontier releases; it tops Scale AI's public SWE-bench Pro leaderboard at 61.5 per cent (Scale AI, 2026; its reported pricing was not fetched from a primary source and is carried in Appendix B, N7), and Alibaba's Qwen3.8-Flash-Next (26 August 2026) shipped under a "qwen-community-1.0" licence rather than Apache 2.0, a tightening commercial users should check (MarkTechPost, 2026). Mistral offered early access to an unnamed larger model to selected partners from July and released nothing publicly (TechTimes, 2026a).

On reliability, the window produced no new peer-reviewed hallucination measurement and no regulator action against "hallucination-free" marketing (Stream A and E nulls, Appendix A). What it produced instead was benchmark integrity evidence. OpenAI's 8 July analysis estimated about 30 per cent of SWE-bench Pro tasks are broken (OpenAI, 2026b). Scale AI's public SWE-bench Pro leaderboard (731 public tasks), fetched in the window, shows top scores between 46 and 61.5 per cent, each with a confidence interval of about plus or minus 3 to 3.6 points (Muse Spark 1.1 at 61.5, GPT-5.4 at 59.1, Muse Spark at 55.0, Claude Opus 4.6 at 51.9, Gemini 3.1 Pro at 46.1),

well below the 80 per cent and 64.6 per cent figures the labs quote for Fable 5 and GPT-5.6 on a non-public subset (Scale AI, 2026; Willison, 2026a). The June edition's "near 80 per cent" was, in retrospect, a vendor figure on a non-public subset rather than the public leaderboard, and this edition records that as a sharpening.

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	Five frontier-class releases; none changed what is dependable with review.	OpenAI (2026a); Anthropic (2026); Axios (2026a)
Gap	**Declined**	The contamination-resistant benchmark June recommended is now estimated by one lab to be about 30 per cent broken; public and vendor-quoted scores diverge by 20 to 30 points.	OpenAI (2026b); Scale AI (2026); Willison (2026a)
Oversold	Unchanged (confirmed)	No new hallucination study; benchmark margins between labs now publicly disputed by the labs themselves.	Willison (2026a)
Underestimated	**Improved**	DeepSeek V4-Flash and Kimi K3 extend the open-weight case on cost and context; Meta's exit and Qwen's licence change are counter-signals, not reversals.	Open Source For You (2026); Willison (2026b); MarkTechPost (2026)

September 2026 judgement. Deploy today: text and document work with human review, structured extraction with validation, and coding assistance, now at materially lower cost for bounded tasks on open-weight models. Demo-versus-production gap: anything quoted in benchmark points, and now specifically any SWE-bench figure not accompanied by the subset, the scaffold, and the number of runs. Oversold: "hallucination-free" claims in any domain, and headline benchmark margins between frontier models, which the labs themselves now dispute. Underestimated: open-weight models, which for many bounded business tasks match closed frontier quality at a fraction of the cost and can run under the business's own control; the licence terms now need reading.

Quotable. As at September 2026, the best public SWE-bench Pro scores sit between 46 and 62 per cent (Scale AI leaderboard) while the labs quote 64.6 and 80 per cent on a non-public subset, and OpenAI itself estimates about 30 per cent of the benchmark's tasks are broken (OpenAI, 8 July 2026); a benchmark figure without its subset, scaffold, and run count is not evidence.

3.2 Agentic systems and tool use

June 2026 judgement. The integration layer had settled on MCP (Linux Foundation, December 2025) with A2A beside it. Agent products worked but none reliably enough to run unattended on consequential tasks: METR's task-horizon research found the length of task an agent completes at 50 per cent success doubling every four to seven months while success decayed steeply with

duration; Carnegie Mellon's TheAgentCompany simulation (2025) had the best agent completing 30.3 per cent of office tasks autonomously, with no rerun on current models; prompt injection was "unlikely to ever be fully solved" by OpenAI's own assessment; and a US district court had enjoined Perplexity's Comet from accessing password-protected Amazon accounts on Computer Fraud and Abuse Act grounds (10 March 2026), stayed by the Ninth Circuit. Judgement: deploy coding agents, research agents with review, and scoped internal automations with approval gates; distrust "digital employee" demonstrations, which showcase the 30 per cent of attempts that succeed; treat autonomous consequential work and "agent washing" as oversold; and treat MCP's compounding value and supervised agents as a labour multiplier as underestimated (Perth AI Consulting, 2026, §3.2).

What happened in the window. The infrastructure matured on schedule. The Model Context Protocol's 2026-07-28 revision became the current protocol version (in MCP's own vocabulary a revision is Draft, Current, or Final, and Final is reserved for past versions that will not change), introducing a feature lifecycle policy (Active, Deprecated, Removed, with at least twelve months between deprecation and the earliest removal, subject to a ninety-day expedited exception), a conformance-suite requirement for standards-track proposals, and an extensions framework that now carries Tasks (asynchronous long-running operations), Skills over MCP, and MCP Apps (interactive elements inline in conversations) (Model Context Protocol, 2026a, 2026b; the Agentic AI Foundation's May post had announced the schedule; Agentic AI Foundation, 2026). Roots, Sampling, and Logging entered deprecation. No authoritative registry count exists: aggregator figures in the window ranged from about 2,000 (official registry) to 81,055 (Glama scrape), which is itself the finding (Appendix B items 6 and 7).

The legal picture changed. On 4 August 2026 the Ninth Circuit vacated the Comet preliminary injunction in Amazon.com Services v. Perplexity AI, holding that "it is the user who 'accesses' Amazon's computers, with the help of the Assistant", that "however advanced the Assistant currently is, it is a tool, not a person for statutory purposes", and that "Perplexity's servers never directly access Amazon's servers"; the panel limited the holding to the technology before it and said it did not "establish a new legal regime governing agentic AI" (Ninth Circuit, 2026). For operators, the CFAA theory against agent makers is narrower than June's snapshot implied; the platform-terms and contract theories are untouched, and nothing in the ruling bears on reliability.

The product landscape churned. OpenAI announced the Atlas sunset on 9 July and shut it down on 9 August 2026, less than ten months after launch (TechCrunch, 2026a; 36Kr, 2026) and launched ChatGPT Work with approval gates (InfoWorld, 2026). Google's standalone Project Mariner was folded into Gemini's agent mode around May 2026, a pre-window event the June edition missed (Appendix B, N2). Microsoft's Windows Agent Runtime stayed in preview (Rework, 2026). Anthropic's Cowork went to mobile and web from 8 July, with vendor telemetry from late May showing agent sessions dominated by business-process work rather than coding (TechCrunch, 2026d).

The independent evidence did not close the gap. No rerun of TheAgentCompany appeared. METR's expenditure-horizon study (21 July 2026) is the closest thing to a fresh independent

measurement and found autonomous agent optimisation "has so far had minimal effect on AI R&D progress" on its target, with 50 to 70 per cent of agent contributions mergeable (METR, 2026). OSWorld figures could not be re-verified against a live XLANG page; Anthropic's 1 September release reports 41.7 per cent on a new "OSWorld 2.0 (strict)" variant (Anthropic, 2026; vendor claim), which is not comparable to June's numbers. On security, OWASP published its GenAI LLM Top 10 for 2026 on 3 August, mapped to its agentic-applications framework (OWASP, 2026); no new dated MCP exploit or prompt-injection incident was located in the window, and NIST's agent standards work remained at initiative stage (Stream B nulls).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	ChatGPT Work (9 July) ships approval-gated; pre-window Cowork telemetry points the same way.	InfoWorld (2026); TechCrunch (2026d)
Gap	Unchanged (confirmed)	No rerun of TheAgentCompany; METR found minimal autonomous effect on a real research target.	METR (2026); Appendix B item 8
Oversold	**Reframed**	The Ninth Circuit narrowed one legal theory against agent makers; reliability and prompt-injection judgements unchanged.	Ninth Circuit (2026); OWASP (2026)
Underestimated	**Improved**	MCP's 2026-07-28 revision is the current version, with a twelve-month deprecation guarantee, conformance testing, and extensions; the standard is compounding as June predicted.	Model Context Protocol (2026a, 2026b)

September 2026 judgement. Deploy today: coding agents; research agents with human review; tightly scoped internal automations with approval gates; and now MCP-connected tools built against the current specification with its twelve-month deprecation guarantee. Demo-versus-production gap: end-to-end "digital employee" demonstrations, which still showcase the roughly 30 per cent of attempts that succeed on the only independent office-task evidence available (2025), and autonomous research or optimisation agents, which METR finds have minimal effect. Oversold: autonomous agents for consequential, unsupervised work; "agent washing" of ordinary automation; and, newly, any reading of the Ninth Circuit's ruling as a green light, since it addresses who accesses a computer, not whether the agent works. Underestimated: the compounding value of the MCP standard, now with a lifecycle policy that makes it safer to build on, and supervised agents as a labour multiplier for staff who learn to delegate and verify.

Quotable. As at September 2026, the only independent office-task evaluation of AI agents still shows about 30 per cent autonomous task completion (Carnegie Mellon, 2025, best-agent result), no rerun with current models has been published in eighteen months, and METR's July 2026 study found autonomous agents had "minimal effect" on real research progress; the MCP standard, meanwhile, published its current specification revision on 28 July 2026 with a twelve-month deprecation guarantee.

3.3 Vision and multimodal models

June 2026 judgement. Document understanding was the highest-value vision capability and the one with the best-measured gap: near-saturated on clean benchmarks (OmniDocBench), about 75 per cent field-level accuracy on the OmniAI benchmark of 1,000 real documents, with the same models scoring about 99 per cent given perfect text, so the bottleneck was reading degraded material, not reasoning about it. Vision-against-standards products (plan review against codes) had no independent accuracy evaluation. Medical imaging had randomised evidence (the MASAI mammography trial, roughly 106,000 women, sensitivity 73.8 to 80.5 per cent, screen-reading workload down 44 per cent) alongside cautionary cases (DermaSensor's 32.5 per cent specificity; SkinVision's photograph-quality failures) and persistent site-transfer degradation. More than 1,250 AI-enabled devices carried FDA authorisation as of mid-2025. Judgement: deploy extraction with confidence-gated review, construction progress capture, and regulator-cleared medical tools inside their indication; distrust handwriting and degraded documents and medical tools moved across sites; treat vision-against-standards products and consumer diagnostic apps as oversold; and treat ordinary document work with verification as the single highest-ROI capability for many SMBs (Perth AI Consulting, 2026, §3.3).

What happened in the window. Little that meets the standard, which is recorded rather than filled. No rerun of the OmniAI real-document benchmark was located (the benchmark page returned a 404), and no new independent document-extraction evaluation appeared. Extend, a document-processing vendor, published a July 2026 note arguing the same benchmark-to-production gap ("a model scoring 98% CER fails field extraction on multi-column tables") with its own product's figures (Extend, 2026; vendor claim, not comparable to OmniAI's method). In vision-against-standards, searches again found only vendor and trade content for CodeComply.AI, PlanCheckPro.AI, and the Australian NCC tools Defect Detector and Voltin; the absence of any independent evaluation is now two editions old. In medical imaging, industry commentary in August cited the FDA's count of AI-enabled devices passing 1,500 as of April 2026 (Clinical Trial Vanguard, 2026; the milestone predates the window), with the Epic sepsis model (33 per cent sensitivity in post-deployment analysis) offered as the caution that authorisation runs ahead of validation. Two systematic reviews on external validation surfaced in search but could not be fetched or dated (Appendix B, N4). No TGA enforcement action against a specific vision product was found. Vision failure-mode research (counting, chart reading, degraded-text hallucination) was under-searched because the budget ran out (Appendix A).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (no new evidence)	June judgement carried forward on its original evidence.	Perth AI Consulting (2026, §3.3)
Gap	Unchanged (no new evidence)	No rerun of the OmniAI real-document benchmark; roughly 75 per cent field accuracy stands as the last independent figure.	Perth AI Consulting (2026, §3.3); Extend (2026)

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Oversold	Unchanged (confirmed)	A second edition's search found no independent evaluation of any vision-against-standards product.	Stream C nulls (Appendix A)
Underestimated	Unchanged (no new evidence)	June judgement carried forward.	Perth AI Consulting (2026, §3.3)

September 2026 judgement. Deploy today: document extraction with confidence-gated human review; construction progress capture; specific regulator-cleared medical tools inside their cleared indication. Demo-versus-production gap: handwriting and degraded documents (clean-benchmark scores near 95 per cent fall to roughly 75 per cent field accuracy on real material, on the last independent measurement, 2025 to 2026); medical tools moved across sites, scanners, and populations. Oversold: vision-against-standards products quoting no independent accuracy data, now for a second consecutive quarter; consumer-grade diagnostic apps. Underestimated: how good frontier models have become at ordinary document work when the workflow includes verification; for many SMBs this remains the single highest-ROI AI capability available.

Quotable. As at September 2026, no independent accuracy evaluation of any AI plan-review or vision-against-standards product has been published, and the last independent measurement of document extraction on real-world material remains about 75 per cent field-level accuracy (OmniAI, 2025 to 2026), against about 99 per cent when the text is clean.

3.4 Video and image generation

June 2026 judgement. The defining event was a shutdown: OpenAI's Sora 2 consumer app closed on 26 April 2026 seven months after launch, with the API to sunset on 24 September 2026. The working frontier was Veo 3.1, Runway Gen-4.5 (up to 60 seconds in a pass), Kling 3.0, Midjourney V8.1, Gemini 3 Pro Image, and FLUX.2; "studio quality" meant single establishing shots, not continuity-stable multi-shot footage; clip length was 4 to 15 seconds for most models; and the zero-touch content farm had been demonetised by YouTube's July 2025 policy. Judgement: deploy image generation for marketing, short-form video with human selection, and AI-assisted pipelines; distrust anything needing continuity, on-screen text, or brand fidelity; treat "automated content engine" revenue, "studio quality" as a general claim, and any single product's durability as oversold; and treat the cost collapse of competent short-form content as underestimated (Perth AI Consulting, 2026, §3.4).

What happened in the window. The Sora API sunset date is unchanged at 24 September 2026 (OpenAI, 2026c), and the pipeline layer has already routed around it: Higgsfield's MCP roster, tracked through 11 August 2026, removed Sora and added Seedance 2.5, Kling 3, and Soul 2.0 (MCP.film, 2026; aggregator listing, corroborating detail only). Google's own Gemini API changelog records shutdown dates of 30 June 2026 for the Veo 2.0 and Veo 3.0 models and 17 August 2026 for the three Imagen 4 variants (Google, 2026b). ByteDance opened public API access to Seedance 2.5 on 16 July 2026, generating a continuous, unstitched 30-second clip in a single pass (Runway Gen-4.5's 60 seconds, pre-window, remains the longest), with the Motion

Picture Association and five studios holding unanswered cease-and-desist letters over Seedance 2.0 from February and no lawsuit yet filed (TechTimes, 2026b; duration is a vendor claim). Alibaba's video model, disclosed as "HappyHorse" (version 1.1), ranked second on Arena.ai's blind-comparison video leaderboards at 1,444 Elo (VentureBeat, 2026); the same report put Sora's economics at roughly US$1 million a day in cost against US$2.1 million total revenue (unattributed figures in a trade article, not independently sourced). xAI's Grok Imagine Video 1.5 reached general availability on 16 June 2026 at US$0.08 per second including audio (InVideo, 2026; secondary source, vendor pricing). No independent stress test of continuity, text, or physics failure modes appeared in the window; Seedance's 30 seconds is a duration claim, not a continuity result. In image generation, nothing in the window superseded Ideogram 4.0 (3 June, open weights, text rendering) or GPT Image 2 (April), both pre-window. Platform policy did not move on any primary source located: YouTube's, Meta's, and TikTok's synthetic-content pages either returned errors or undated text, and Senator Pocock's "My Face, My Rights" deepfake bill remained before the Senate with no verified progress (Stream D nulls; Appendix A).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	New models extend duration and cut cost without changing the human-curated, short-form boundary.	TechTimes (2026b); VentureBeat (2026)
Gap	Unchanged (no new evidence)	No independent stress test in the window; single-pass duration is a vendor claim.	Stream D nulls (Appendix A)
Oversold	**Declined**	Sora's sunset holds, Google shut down Veo 2.0, Veo 3.0, and Imagen 4, and pipelines have already dropped Sora; product durability is worse than June judged.	OpenAI (2026c); Google (2026b); MCP.film (2026)
Underestimated	Unchanged (no new evidence)	One new vendor price point (Grok Imagine, US$0.08 a second) is not a trend measurement.	InVideo (2026)

September 2026 judgement. Deploy today: image generation for marketing and mock-ups; short-form video for social and advertising with human selection from multiple generations; AI-assisted (not AI-run) video pipelines, built against a model roster the business expects to replace. Demo-versus-production gap: anything requiring shot-to-shot continuity, on-screen text, or precise brand fidelity; the showreel is the best of dozens of attempts, and a 30-second single pass is a duration, not a continuity guarantee. Oversold: "automated content engine" revenue claims; "studio quality" as a general proposition; and the durability of any single product, which the window demonstrated again with five model shutdowns at Google (Veo 2.0, Veo 3.0, and three Imagen 4 variants) and a pipeline that had already dropped the most-hyped product of 2025. Underestimated: how cheap and fast competent short-form visual content has become for businesses that previously could not afford it at all.

Quotable. As at September 2026, the Sora API is three weeks from its 24 September shutdown, Google shut down Veo 2.0 and Veo 3.0 on 30 June and Imagen 4 on 17 August, and the leading

automated-video pipeline has already removed Sora from its roster; a business building on any single video model should plan for its replacement within a year.

3.5 Audio, voice, and music generation

June 2026 judgement. Voice-agent platforms were commercially mature (sub-second latency; US$0.06 to 0.33 per minute; ElevenLabs at a US$11 billion valuation with US$330 million-plus reported ARR), but almost every published benchmark was vendor-authored and only one small independent evaluation existed. Credible evidence supported bounded calls (bookings, FAQs, intake, missed-call response) at satisfaction comparable to humans, with realistic containment near 50 per cent rather than the 90-plus in vendor decks. Transcription word-error rates of 4 to 5 per cent on benchmarks became 15 to 20 per cent on real audio; medical-scribe errors were dominated by omissions and were higher for some accents. Music was licensed-but-unsettled (Universal and Warner settlements; Sony and Universal's Suno case continuing; no fair-use ruling). Voice-cloning fraud was real (the Arup case) but the aggregate statistics were untraceable. Judgement: deploy bounded after-hours and overflow voice handling with escalation, transcription with review, and TTS; distrust clean-audio claims and vendor containment rates 1.4 to 2 times delivered; treat "indistinguishable from human", autonomous phone sales, and vendor ROI statistics as oversold; and treat missed-call response and the fraud exposure as underestimated (Perth AI Consulting, 2026, §3.5).

What happened in the window. Commercial activity without evaluative evidence. Bland AI raised a US$50 million Series C on 16 June 2026 led by Dell Technologies Capital, reporting 3.5 million calls a week across 250-plus enterprise customers (Fortune, 2026; vendor figures). ElevenLabs' page reporting US$500 million ARR (originally 5 May, updated 31 August) and its 10 August changelog (Dubbing v2 across 90-plus languages, agent conversation-filtering, speech-to-text updates, no error rates given) are vendor material (ElevenLabs, 2026a, 2026b). Every source located on containment or accuracy was a vendor or vendor-adjacent comparison site; the scarcity of independent voice-agent evaluation that June called "a finding" persisted through a second quarter. No new independent word-error-rate benchmark on real-world audio and no new accent-equity study appeared. In music, the litigation moved without resolving: the American Federation of Musicians sued Universal and Warner in early June 2026 (reported 8 June; pre-window) over the Suno and Udio licences; as of 6 July 2026 Judge Saylor had not ruled on the plaintiffs' motion to add 61,026 recordings to the Suno complaint, Suno was resisting it citing a parallel denial of a 30,442-recording expansion in the Udio case, and the theoretical statutory exposure of the expansion was put at over US$9 billion against about US$84 million on the original 560 works (Music Business Worldwide, 2026a, 2026b). No court anywhere ruled on fair use for training in the window. On fraud, the ACCC's 5 June release (pre-window) reported A$248.3 million in combined Australian scam losses for January to March 2026 with no voice-cloning breakout, and no FBI IC3 2026 report had been published (ACCC, 2026a; Appendix B item 15). ASIC's 19 August report on dismantling 3,106 fraudulent crypto platforms in FY2026 explicitly cited deepfake endorsements of the Prime Minister and the finance commentators Tom Piotrowski and Alan Kohler (ASIC, 2026; §3.9).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (no new evidence)	June judgement carried forward; commercial growth is not reliability evidence.	Fortune (2026); Perth AI Consulting (2026, §3.5)
Gap	Unchanged (no new evidence)	Still no independent cross-platform voice-agent evaluation.	Stream D nulls (Appendix A)
Oversold	Unchanged (confirmed)	Scale claims (calls per week, ARR) continue to substitute for containment evidence.	Fortune (2026); ElevenLabs (2026a)
Underestimated	Unchanged (no new evidence)	No primary voice-cloning loss data in the window.	ACCC (2026a)

September 2026 judgement. Deploy today: voice agents for bounded after-hours and overflow handling with human escalation; transcription with review for anything consequential; TTS for IVR, content, and accessibility. Demo-versus-production gap: clean-audio accuracy claims; vendor containment rates roughly 1.4 to 2 times delivered rates, on evidence that has not been independently refreshed since 2025. Oversold: "indistinguishable from human" as a blanket claim, fully autonomous phone-based sales, and most published voice-agent ROI statistics, which remain vendor case studies. Underestimated: missed-call response and after-hours intake, which are narrow, cheap, and capture revenue that was previously lost; and the fraud exposure from the same cloning technology, which Australian regulators are now naming in scam reporting, and which most SMBs have not addressed with callback verification and payment controls.

Quotable. As at September 2026, no independent evaluation of voice-agent containment or accuracy has been published in the two quarters this series has looked, while the platforms report scale (3.5 million calls a week at Bland; US$500 million ARR at ElevenLabs, both vendor figures); the realistic planning assumption remains roughly half of routine calls fully contained.

3.6 Knowledge management and synthesis

June 2026 judgement. NotebookLM defined the category (source-grounded chat, Audio and Video Overviews, a Deep Research agent, enterprise deployment inside Workspace governance); Claude Projects and ChatGPT Projects offered persistent scoped workspaces; Microsoft Copilot's knowledge features were widely deployed and widely stalled on governance, because Copilot surfaces whatever a user technically has permission to see. The honest accuracy number was a 2025 newsroom study (arXiv:2509.25498) in which NotebookLM, the best tool tested, still hallucinated in roughly 13 per cent of responses against about 40 per cent for general assistants. Million-token context did not kill retrieval: accuracy degraded well before the advertised window on multi-fact tasks and the economics favoured retrieval for large corpora. June put the grounded-hallucination floor at roughly 10 to 15 per cent. APQC found 85 per cent of organisations had not operationalised AI knowledge tools at scale (vendor-sponsored survey).

Judgement: deploy grounded Q&A over curated, current document sets with citations checked; distrust enterprise search over an ungoverned file estate; treat "ask anything about your business" and the implication that grounding means accuracy as oversold; and treat NotebookLM-class tools for regulated professionals against bounded authoritative texts as underestimated (Perth AI Consulting, 2026, §3.6).

What happened in the window. The products kept adding grounding and memory. Google's NotebookLM upgrade (8 June 2026, pre-window, five days before June's research closed) added an agentic research layer with code execution and web-research source building, with a vendor-claimed 65 per cent-plus internal win rate over the prior version (Google, 2026a; vendor claim). Anthropic merged Claude's memory across chat and Cowork on 25 August 2026, with user-editable memories and sensitive categories excluded by default (TechCrunch, 2026e). Neither is accuracy evidence.

The accuracy evidence that arrived went the other way. A systematic taxonomy published in the TrustNLP workshop at ACL on 4 July 2026 catalogues 33 failure modes in retrieval-augmented generation across seven pipeline stages and finds that 12 of the 33 (36 per cent), including all eight agentic-RAG modes, still lack any peer-reviewed empirical investigation; the paper calls agentic RAG the fastest-growing deployment paradigm with the least scientific scrutiny (TrustNLP, 2026). A 5 July preprint proposing a span-level grounded-hallucination detector (GASP) reached a response-level AUC of about 0.73 on the RAGTruth benchmark, transferred to summarisation, and did not transfer to short-answer question answering (arXiv, 2026); a purpose-built detector that works on some task types and not others is evidence that the grounded-hallucination floor June described is not yet bounded, let alone falling. No independent follow-up to the newsroom study appeared, no new dated enterprise report on Copilot governance stalls was located, and no new long-context benchmark met the standard (Stream E nulls). Gartner reported on 1 September 2026 that only 22 per cent of organisations have scaled AI across multiple business units (Gartner, 2026; analyst self-report, not knowledge-tools specific).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	Anthropic's 25 August memory merge extends the deployable pattern; NotebookLM's upgrade is pre-window.	TechCrunch (2026e); Google (2026a)
Gap	Unchanged (no new evidence)	No new dated report on enterprise search governance.	Stream E nulls (Appendix A)
Oversold	**Declined**	A peer-reviewed taxonomy finds 36 per cent of catalogued RAG failure modes unstudied; a purpose-built detector fails to transfer across tasks. Grounding-means-accuracy is more oversold than June implied.	TrustNLP (2026); arXiv (2026)

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Underestimated	Unchanged (no new evidence)	June judgement carried forward.	Perth AI Consulting (2026, §3.6)

September 2026 judgement. Deploy today: grounded Q&A and synthesis over curated, current document sets, with citations checked for consequential decisions; meeting synthesis; audio digestion. Demo-versus-production gap: enterprise search over an ungoverned file estate; the demo corpus is clean, yours is not. Oversold: "ask anything about your business" positioning, and the implication that source-grounding means accuracy, which the July 2026 failure-mode literature shows is less well understood than the products imply, particularly once agents are added to the retrieval loop. Underestimated: NotebookLM-class tools for regulated professionals working against bounded authoritative texts (a standard, an act, a product disclosure statement), which is close to the ideal use case, provided the verification habit holds.

Quotable. As at September 2026, the peer-reviewed literature catalogues 33 distinct ways a retrieval-augmented AI system can fail, has empirically studied only 21 of them, and has studied none of the eight that involve agents (TrustNLP, July 2026); the last independent measurement of the best grounded tool still put its hallucination rate near 13 per cent (2025).

3.7 Code generation and software creation

June 2026 judgement. The commercial facts were extraordinary and self-reported: Claude Code at a reported US$2.5 billion run-rate (February 2026), Cursor at a reported US$2 billion ARR, roughly 4 per cent of public GitHub commits authored by Claude Code (SemiAnalysis, commit-signature method). The productivity evidence was contested and the best study said so: METR's July 2025 trial found experienced developers 19 per cent slower while believing themselves 20-plus per cent faster; its February 2026 follow-up found point estimates flipping to modest speedups with intervals too wide to fix a number. DORA 2025 found throughput and instability rising together; Stack Overflow's 2025 survey found 84 per cent using AI and distrust rising to 46 per cent; GitClear found duplication up roughly 50 per cent. Non-developers could ship prototypes and internal tools; not anything with authentication, payments, or adversarial users (Escape.tech's scan of 5,600 vibe-coded apps found over 2,000 high-impact vulnerabilities; vendor source). Judgement: deploy AI-assisted development with review and tests non-negotiable, and non-developer internal tools holding no sensitive data; distrust the demo app with no auth and no attackers; treat "anyone can ship production software" and benchmark-derived claims as oversold; and treat the fall in the cost of custom internal software as underestimated (Perth AI Consulting, 2026, §3.7).

What happened in the window. The revenue figures kept climbing and stayed self-reported. Anthropic's run-rate was reported at about US$65 billion at end-July (Bloomberg-sourced, not company-confirmed; TechCrunch, 2026c), against US$19 billion in March (company-reported; Yahoo Finance, 2026b). Sacra estimated Cursor at US$4 billion annualised in May (pre-window; Sacra, 2026) and reported a 16 June 2026 agreement for SpaceX to acquire Cursor for US$60

billion, which could not be corroborated from a second source and is carried as unverified (Appendix B, N6). The coding-agent products themselves iterated: GPT-5.6 Sol was launched on a vendor-claimed lead on the Artificial Analysis Coding Agent Index (OpenAI, 2026a), Fable 5.1 on vendor-claimed gains on Terminal-Bench and CursorBench (Anthropic, 2026), and GitHub's changelog showed incremental Copilot releases with no confirmation that Agent HQ left preview (GitHub, 2026).

The independent productivity evidence did not refresh. Stack Overflow's 2026 developer survey opened on 23 June 2026 "for human developers only" and had not published results by 2 September (Stack Overflow, 2026). No DORA 2026 report existed. GitClear's most recent update was dated 9 June (pre-window; GitClear's 2025 analysis is cited in Perth AI Consulting, 2026, §3.7). METR's expenditure-horizon study is the one new independent data point and concerns autonomous optimisation of a research codebase rather than assisted development; its finding of minimal effect and 50 to 70 per cent mergeable contributions is consistent with June's "verification tax" (METR, 2026). No new vibe-coding security scan and no new named breach attributed to an AI-built application were located (Stream A nulls).

What did move was benchmark integrity, and it moved against the sales deck. OpenAI's 8 July analysis estimated about 30 per cent of SWE-bench Pro tasks broken (27.4 per cent by pipeline, 34.1 per cent by five human reviewers) after having found SWE-bench Verified had "fundamental design and contamination issues" earlier in the year (OpenAI, 2026b). Scale AI's public leaderboard shows the best public score at 61.5 per cent (Muse Spark 1.1) with Claude Opus 4.6 at 51.9 and Gemini 3.1 Pro at 46.1, while labs quote 80 and 64.6 per cent for their newest models on a commercial subset (Scale AI, 2026; Willison, 2026a). Both SWE-bench variants June discussed are now disputed by at least one lab.

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	Products iterated inside review workflows; no evidence changed the boundary.	OpenAI (2026a); Anthropic (2026); GitHub (2026)
Gap	Unchanged (no new evidence)	No new security scan; survey results pending.	Stack Overflow (2026); Stream A nulls
Oversold	**Declined**	Benchmark-derived claims are now contested by the labs themselves; both SWE-bench variants disputed.	OpenAI (2026b); Scale AI (2026)
Underestimated	Unchanged (no new evidence)	Price falls (DeepSeek V4-Flash, Opus 5) are directionally consistent; no study measured the effect on internal-software cost.	Open Source For You (2026); Axios (2026a)

September 2026 judgement. Deploy today: AI-assisted development for professional teams, with code review and tests treated as non-negotiable; non-developer internal tools that hold no sensitive data. Demo-versus-production gap: the app that works in the demo has no auth, no

edge cases, and no attackers; and the developer who feels faster has, on the last randomised evidence, not been measured as faster. Oversold: "anyone can ship production software"; and benchmark-derived capability claims, which as of July 2026 the labs dispute among themselves on the field's flagship coding benchmark. Underestimated: the aggregate effect of coding agents on the cost of custom internal software, which has fallen far enough that processes too small to justify software in 2023 now justify it, provided someone competent owns security; the window's price cuts push further in that direction.

Quotable. As at September 2026, no randomised evidence has been published since February 2026 on whether AI coding tools make developers measurably faster, the 2026 Stack Overflow survey has not reported, and the one flagship coding benchmark the labs agree to use is estimated by OpenAI (8 July 2026) to have about 30 per cent broken tasks; buy on code review and test coverage, not on a benchmark chart.

3.8 Business automation platforms with AI built in

June 2026 judgement. Platforms sold autonomy and delivered supervised competence, and pricing said so: HubSpot moved Breeze agents to US$0.50 per resolved conversation, Intercom's Fin charged US$0.99 per resolution, and Salesforce restructured Agentforce pricing twice. Salesforce showed the clearest marketing-to-reality gap (about 3,000 paid customers against 5,000 deals, first-year cost of ownership US$150,000 to 425,000 on partner analyses); HubSpot claimed 65 per cent resolution (vendor); Intercom claimed 65 to 70 while its own case studies clustered at 42 to 53 and an independent January 2026 evaluation of complex tickets found 24 per cent best case. The Microsoft 365 Copilot evidence was two studies: the UK GDS cross-government trial (about 20,000 civil servants, high satisfaction, self-reported 26 minutes a day saved, no independent productivity measurement) and the much smaller DBT sub-study (about 300 analysed users, no measurable gain, weaker spreadsheet performance despite feeling faster). Judgement: deploy missed-call text-back, after-hours intake, FAQ-grade deflection on a maintained knowledge base, and workflow automation with approval steps; distrust resolution rates from simple traffic and "AI employee" demos; treat enterprise agent suites as plug-and-play and all published ROI figures as oversold; and treat the boring automations and outcome pricing as underestimated (Perth AI Consulting, 2026, §3.8).

What happened in the window. One large number and no new measurement. Salesforce reported on 26 August 2026 that Agentforce annual recurring revenue "exceeded $1.5 billion", up more than 240 per cent year on year, and combined Agentforce and Data 360 ARR of nearly US$3.9 billion, with no customer count and no cost-of-ownership figure disclosed (Salesforce, 2026; investor filing, growth rate off a small base). No new independent evaluation of customer-service AI resolution rates appeared; the only in-window figures were aggregator claims without a stated method (Stream B nulls). HubSpot's April 2026 outcome-pricing move was re-reported repeatedly but is baseline evidence; no new per-outcome pricing or reversal was found. GoHighLevel again had no independent audit; the horizontal tools (Zapier Agents, Make, n8n) surfaced only comparison content. For Microsoft 365 Copilot, the Australian Government's whole-of-government trial evaluation (about a third of participants using it daily; predominantly

summarising and rewriting; self-reported savings of "up to an hour" on those tasks; 64 per cent of managers perceiving uplift) was located but its publication date could not be recovered from the fetched page, and it is believed to predate the June edition (Digital Transformation Agency, n.d.; Appendix B, N3). No new AI-receptionist evidence appeared beyond a press-release wire claim by an Australian entrant (Appendix B item 16).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (no new evidence)	June judgement carried forward.	Perth AI Consulting (2026, §3.8)
Gap	Unchanged (no new evidence)	No independent resolution-rate measurement in the window.	Stream B nulls (Appendix A)
Oversold	Unchanged (confirmed)	Agentforce ARR reported with no customer count or cost figure; the pattern June described continues.	Salesforce (2026)
Underestimated	Unchanged (confirmed)	Outcome pricing stands; no reversal.	Perth AI Consulting (2026, §3.8)

September 2026 judgement. Deploy today: missed-call text-back and after-hours intake; FAQ-grade customer service deflection on a well-maintained knowledge base; workflow automation with human approval steps. Demo-versus-production gap: resolution rates quoted from simple-traffic mixes; "AI employee" demonstrations that conceal the throttles, the configuration burden, and the escalation tail; and vendor revenue growth quoted without customer counts. Oversold: enterprise agent suites as plug-and-play (the data engineering is the project); virtually all published ROI numbers in this category, none independently audited, now including a US$1.5 billion ARR figure that says nothing about delivered resolution. Underestimated: the boring automations, which is where the dependable money is, and outcome-based pricing as a buyer's tool, since a vendor willing to charge per resolution is making a falsifiable claim.

Quotable. As at September 2026, Salesforce reports Agentforce revenue above US$1.5 billion a year (26 August 2026) and no independent measurement of any enterprise customer-service agent's resolution rate has been published since January 2026; the buyer's protection is still the same, pay per resolution and measure the escalation tail.

3.9 Crypto and AI convergence

June 2026 judgement. Infrastructure consolidating impressively, demand small, one flagship consumer experiment already shut. NEAR ran the most coherent stack (confidential compute; Intents at about US$20 billion cumulative on-chain volume, overwhelmingly swaps and bridges; a marketplace with no published metrics). OpenAI sunset Instant Checkout on 24 March 2026 with about 30 merchants ever live and Walmart reporting three times lower in-chat conversion. The

rails consolidated anyway: AP2 to the FIDO Alliance; Visa and Mastercard programmes in pilot; Visa's roughly US$7 billion stablecoin-settlement run-rate (settlement, not agent volume); Stripe's Tempo live with a Machine Payments Protocol; x402 at 165 million transactions and about US$50 million cumulative by late April, "real and tiny". Bittensor's US$43 million quarterly revenue stood against an independent US$3 to 15 million annual estimate; Fetch.ai's ASI chain was unlaunched. Judgement: deploy nothing, watch, with stablecoin settlement the adjacent capability; distrust agent-to-agent commerce demos; treat token volume and "the agent economy is here" as oversold; and treat the rails and confidential compute as underestimated (Perth AI Consulting, 2026, §3.9).

What happened in the window. The rails moved to production and the volume moved to nothing. Visa Intelligent Commerce went live in Europe on 2 July 2026 (announced at the Visa Payments Forum in Paris, reported 6 July) with more than 30 issuing banks (vendor-announced count) (Barclays, HSBC UK, BBVA, ING, Commerzbank, Nordea, Revolut, Klarna) and named merchants (lastminute.com, Frasers, Cleverbridge, BrickDepot), with authentication through Visa Payment Passkeys and a Trusted Agent Protocol; no transaction volume was published (Industry Spread, 2026). Mastercard's Agent Pay for Machines (10 June, pre-window) and the Visa and OpenAI integration announcement (12 June, pre-window, no launch date) sit at the boundary and are noted in the "missed" list above. No successor to Instant Checkout shipped. US Treasury published its GENIUS Act rulemaking on payment-stablecoin issuance on 18 August 2026 with comments to 19 October; the regime is proposed, not final (Federal Register, 2026). On the one rail with hard numbers, x402's daily settlement volume had fallen 93 per cent year to date and 55 per cent in three months as of 13 August 2026, to a seven-day average of about US$41,800 and a latest daily figure near US$28,400, from a late-2025 peak that repeatedly approached US$800,000 and occasionally exceeded US$1 million a day, a fall of about 97 per cent from that peak (Yahoo Finance, 2026a; Helios Analytics figures). NEAR Intents reached "over $22 billion" cumulative by 27 July (NEAR's own framing, quoted in the Nansen report), and the Nansen quarterly report gave the first quantified figure for the confidential routing feature: about US$30 million TVL, with 42 per cent of near.com swap volume routed privately by default in Q2 (Nansen, 2026; analytics aggregator, not vendor). Stripe's Tempo published no throughput; Bittensor's figures were only re-reported; the ASI chain remained at testnet with mainnet "late 2026 or early 2027" (Laika Labs, 2026). ASIC reported on 19 August 2026 that it had dismantled 3,106 fraudulent crypto investment platforms in FY2026, up about 30 per cent, with generative-AI deepfakes of the Prime Minister and the finance commentators Tom Piotrowski and Alan Kohler cited in more than A$7.4 million of reported impersonation-scam losses (ASIC, 2026).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	Still nothing for a typical reader; the one measured machine-payment rail collapsed.	Yahoo Finance (2026a)

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Gap	**Reframed**	Visa's programme is live with named issuers and merchants (2 July) but no published volume; no consumer-scale successor to Instant Checkout. The gap is now "live but unmeasured" rather than "pilot".	Industry Spread (2026)
Oversold	**Declined**	x402 daily volume down 93 per cent year to date (about 97 per cent from its late-2025 peak); "the agent economy is here" is more oversold than June judged.	Yahoo Finance (2026a)
Underestimated	**Improved**	Rails live (Visa, 2 July), rulemaking under way (GENIUS Act), and a first quantified confidential-compute adoption figure (NEAR).	Industry Spread (2026); Federal Register (2026); Nansen (2026)

September 2026 judgement. Deploy today: for almost all PAC readers, nothing; this is a watch category. Stablecoin settlement for international payments remains the adjacent capability with real near-term relevance, now with a US federal rulebook in draft. Demo-versus-production gap: agent-initiated card payments are live in Europe with named banks and merchants and no published volume; agent-to-agent commerce demos still stand against the absence of any consumer-scale deployment that has survived contact with merchants. Oversold: token projects citing volume that is speculation or subsidy; "the agent economy is here", against a machine-payment rail whose daily volume fell from a late-2025 peak near US$800,000 to about US$30,000 to 40,000 by August. Underestimated: the rails themselves; the incumbents moved from pilot to production inside a quarter, and confidential compute for AI workloads now has a first measured adoption figure, which is small and real.

Quotable. As at September 2026, Visa's agent-payment programme is live with more than 30 European banks and no published transaction volume (2 July 2026), while daily settlement on the x402 machine-payment rail has fallen 93 per cent year to date to roughly US$30,000 to 40,000 (13 August 2026); the rails are ready and the demand is not.

3.10 Vertical AI applications

June 2026 judgement. Vertical products were where the most defensible value sat and where Australian products were most visible. Clinical scribes were the most validated category (Heidi Health's US$65 million Series B; independent evidence of real but modest benefit, about 16 minutes saved per eight hours in a multi-site study, a 21 per cent burnout reduction in JAMA cohorts, and omissions dominating errors at 86 per cent in one validation); AHPRA, RACGP, and the TGA (30 January 2026, scribes that interpret conversations are regulated medical devices) had published the rules. Legal AI was a strong accelerant whose output courts now required practitioners to verify (Federal Court GPN-AI, 16 April 2026; Stanford RegLab's 17 and 33 per cent hallucination rates). Financial advice had 74 per cent of practices using or planning to use AI (up from 45 per cent a year earlier) against a global figure near 45 per cent with AI policies, and no ASIC enforcement yet. Real estate: listing copy solved, valuation the frontier, RICS's standard

mandatory from 9 March 2026. Construction: progress tracking in production on vendor evidence only. Accounting: confidence-gated reconciliation (Xero JAX in open beta) was the right pattern and no independent benchmark existed. HR: competent drafts, a liability unreviewed. Trades: real time-savers, none autonomous. Judgement: deploy scribes under compliant consent and review, legal research with mandatory citation verification, confidence-gated bookkeeping, and quote drafting; distrust vendor time-saved claims; treat "hallucination-free" professional tools and autonomous compliance documents as oversold; and treat modest validated savings in high-frequency workflows and Australia's already-published regulatory floor as underestimated (Perth AI Consulting, 2026, §3.10).

What happened in the window. The regulators kept writing, and the products kept scaling on vendor numbers. On the regulatory floor: Clayton Utz's 24 July 2026 note on the TGA's scribe position reports that software as a medical device has been "identified as one of 12 priority focus areas for the TGA's compliance and enforcement activities in 2026 and 2027" and flags "feature creep" (a scribe adding diagnostic or treatment suggestions may trigger new ARTG obligations) (Clayton Utz, 2026; law-firm summary; the TGA's own priorities page was not fetched, Appendix B, N8). The Fair Work Commission published AI-use requirements on 26 August 2026, effective 20 October 2026: parties must disclose generative-AI use in case documents, verify citations, keep witness statements in the witness's own words, keep confidential material out of public tools (naming ChatGPT, Claude, Copilot, and Gemini), with non-compliant documents liable to be disregarded and false evidence carrying up to twelve months' imprisonment (Fair Work Commission, 2026; SmartCompany, 2026). The OAIC updated its facial-recognition guidance on 29 July 2026 to reflect the Administrative Review Tribunal's Bunnings decision; the guide, first published in November 2024, says entities "should conduct a formal, structured and documented risk assessment prior to implementing FRT" and adopt a privacy-by-design approach (OAIC, 2026c). The EU's Article 50 transparency obligations commenced on 2 August (§2.4). No new Federal Court or state practice note, no new AHPRA or RACGP guidance, and no ASIC statement on AI-assisted advice was located (Stream E nulls, with the ASIC null weakened by the search-budget limit; Appendix A).

On the incident record: Damien Charlotin's database of AI hallucination cases passed 2,000 entries (2,005 when fetched on 2 September 2026, most recent entry 31 August), with the United States at 1,375 and Australia second at 110, including two Australian entries inside the window (MFRD and National Disability Insurance Agency, Administrative Review Tribunal, 26 August; Heywood v State of Queensland (Department of Education), Queensland Industrial Relations Commission, 24 August) (Charlotin, 2026). No further Australian solicitor sanction beyond the 2025 case was located.

On the products: Harvey was reported on 7 August 2026 to be raising US$500 million at a US$15.5 billion valuation with annualised revenue above US$350 million, up from US$190 million in January (PYMNTS, 2026; reported, not closed). Heidi Health raised no Series C in the window (Appendix B item 24); Abridge's last events (Eli Lilly investment, NVIDIA partnership) fell on 11 June, pre-window. No new peer-reviewed scribe evidence on time, burnout, or accents appeared. Xero announced at Xerocon Denver (post dated 19 August 2026) that auto bank

reconciliation had processed "more than 100 million transactions" since launch and was joined by XeroForce (a natural-language agent builder, in early access with general availability promised later in 2026), a month-end automation agent, and live Xero data inside Microsoft 365 Copilot and ChatGPT (Xero, 2026; vendor claims; auto reconciliation itself carries no stated exit from beta and no independent accuracy benchmark, Appendix B item 25). REA Group's FY26 results (6 August 2026) name a Consumer AI Assistant available to all members, Campaign Assist for Ignite customers, and AI tools for Mortgage Choice brokers, which resolves the June edition's inability to verify any dated REA generative feature (Kalkine, 2026; Appendix B item 27); nothing was found for Domain. Employment Hero's most recent figures (A$300 million-plus ARR, three "agentic" HR products, 17 February 2026) predate the window and supersede the A$2 billion 2023 valuation June carried as its only figure (Employment Hero, 2026; vendor). No dated trades-platform changelog entry was fetchable; ServiceM8's and Simpro's update pages returned errors, and the sub-category remains sourced from review sites (Stream E nulls).

Movement.

JUDGEMENT	CLASSIFICATION	REASON	SOURCE
Deploy today	Unchanged (confirmed)	Confidence-gated reconciliation scaled (vendor figure); tribunal rules formalise the verification workflow.	Xero (2026); Fair Work Commission (2026)
Gap	Unchanged (no new evidence)	No new controlled study of time saved in any vertical.	Stream C and E nulls (Appendix A)
Oversold	Unchanged (confirmed)	The hallucination-incident record passed 2,000 cases with Australian entries in the window.	Charlotin (2026)
Underestimated	**Improved**	FWC disclosure rules (new, dated, with penalties), the TGA's listing of software as a medical device among twelve compliance priorities, and the OAIC's post-Bunnings update: the regulatory floor is more specific than in June.	Fair Work Commission (2026); Clayton Utz (2026); OAIC (2026c)

September 2026 judgement. Deploy today: AI scribes under AHPRA/RACGP-compliant consent and review workflows, on ARTG-included products, with the TGA now listing the category among its compliance priorities; legal research acceleration with mandatory citation verification, which from 20 October 2026 is a disclosure obligation before the Fair Work Commission as well as the courts; confidence-gated bookkeeping automation; quote drafting in trade platforms. Demo-versus-production gap: time-saved claims (vendor surveys say 40 to 60 per cent; the controlled studies, none refreshed this quarter, say minutes per day that are nonetheless worth having). Oversold: "hallucination-free" professional tools, against an incident record that passed 2,000 cases in August 2026 with Australia second in the world; autonomous compliance documents of any kind (SOAs, building reports, HR policies, tribunal filings). Underestimated: the compounding value of modest, validated savings inside high-frequency workflows, and the degree to which Australian regulators have now published the rules of the road, which are documented, achievable, dated, and mostly ignored.

Quotable. As at September 2026, Australian regulators moved three times in one quarter (the Fair Work Commission's AI-disclosure requirements, announced 26 August and effective 20 October; the TGA's listing of software as a medical device among its 2026-27 compliance priorities; the OAIC's 29 July update to its facial-recognition guidance), and the global database of court cases involving AI-fabricated material passed 2,000 entries with Australia second at 110 (checked 2 September 2026).

---

4. Discussion

The two gaps held, and both got fresh numbers. June's macro picture was a pair of gaps: benchmark-to-production and individual-to-firm. The window tested both. On benchmark-to-production, the field's own flagship coding benchmark was found by one of its two leading vendors to be about 30 per cent broken (OpenAI, 2026b), the public leaderboard and the vendor-quoted scores on it diverged by 20 to 30 points (Scale AI, 2026; Willison, 2026a), and METR's attempt to measure what autonomous agents contribute to a real research target, in dollars, found the answer was close to nothing at current capability (METR, 2026). On individual-to-firm, McKinsey's 2026 State of AI survey, in the field from 4 May to 8 June 2026 with 1,719 respondents in 97 countries, found 80 per cent of respondents reporting that AI improved their individual productivity, 37 per cent attributing at least some EBIT impact to AI (flat on the prior year), and the share of "high performers" flat at about 6 per cent (McKinsey, 2026; consultancy self-report). Forty per cent of large organisations reported scaling agents, up from 27 per cent; among smaller organisations the figure stayed at 22 per cent. Three years and one quarter into the era, the explanation June gave has not needed revision: the binding constraints are workflow redesign, data quality, verification, and accountability, none of which a better model fixes by itself.

The evidence base is polluted, and the labs have joined in. June's warning was about vendor benchmarks, affiliate reviews, and SEO content. The window added a category: labs auditing each other's chosen benchmarks at moments that suit them. OpenAI's SWE-bench Pro analysis was published the day before its own flagship release and the day after a competitor's better score on that benchmark (Willison, 2026a). The analysis may be entirely right; the point for a buyer is that a benchmark's validity is now itself a contested vendor claim, and the practical rule from June, ask which benchmark, measured by whom, with what scaffolding, and how many runs, gains a fifth question: on which subset. The independent institutions June named (METR, Epoch AI, Stanford, DORA, government evaluations) published one major result in the window (METR) and otherwise lagged as June said they would; Stack Overflow's survey and DORA's report, the two annual independent snapshots of developer practice, both fall after this window.

What is shifting fastest. June named three things: agentic coding, cost, and infrastructure standardisation. Cost moved fastest in the window: DeepSeek V4-Flash at US$0.14 per million input tokens with a one-million-token context and an MIT licence (Open Source For You, 2026),

Anthropic's 75 per cent cache-read cut (Anthropic, 2026), and OpenAI's three-tier pricing from US$1 (OpenAI, 2026a) mean a bounded task that cost a dollar in June costs cents in September. Infrastructure standardisation moved on schedule: a new current MCP specification with a deprecation guarantee (Model Context Protocol, 2026a), Visa's agent payments live in Europe (Industry Spread, 2026), the EU's transparency obligations in force (European Commission, 2026b). Agentic coding kept compounding commercially (a reported US$65 billion Anthropic run-rate, Bloomberg-sourced and unconfirmed; TechCrunch, 2026c) without any new independent productivity measurement, which is the same shape as June.

What is still hard. The same four things, with sharper evidence on two. Reliability at the tail: unchanged, and unmeasured this quarter in most categories. Verification: the RAG failure-mode taxonomy (TrustNLP, 2026) makes the problem look larger, not smaller, and a purpose-built detector that transfers to some tasks and not others (arXiv, 2026) is a reminder that "grounded" is a design choice, not a guarantee. Security for agents: no new incident of note, OWASP's 2026 guidance published (OWASP, 2026), and the Ninth Circuit's ruling (Ninth Circuit, 2026) shifting one legal risk from the agent maker to the user, which for a business deploying an agent under its own account is not reassuring. Measurement: the perception gap appears again in McKinsey's 80 per cent individual gains against 6 per cent firm-level high performers, and in the Australian Government's own Copilot evaluation, where "up to an hour" of self-reported savings sits beside a third of participants using the tool daily (Digital Transformation Agency, n.d.).

The economics overhang deepened. Alphabet raised its 2026 capital expenditure guidance to about US$205 billion on 22 July (CNBC, 2026; primary page not fetched, figure corroborated across secondary reports), Amazon to about US$220 billion on 30 July (Seeking Alpha, 2026; headline only), and Meta narrowed its range upward to US$130 to 145 billion on 29 July, with second-quarter capex of US$31.1 billion against US$17.0 billion the prior year (Investing.com, 2026). Microsoft reported US$35.8 billion of property and equipment additions for the June quarter and US$115.9 billion for fiscal 2026, against US$64.6 billion in fiscal 2025 (Microsoft, 2026). Those four figures sum to roughly US$671 to 686 billion, inside June's US$630 to 725 billion range for 2026 hyperscaler capex and in its upper half, with the caveat that Microsoft's figure is a fiscal year to June and the others are calendar-year guidance, so the sum is indicative only; the direction, two raised and one narrowed upward, is what matters. Bloomberg reported fresh circular-financing concern around roughly US$750 billion of Nvidia-linked deals on 27 July and ran two bubble-warning columns in late August (Bloomberg, 2026a, 2026b; headlines captured, not fetched in full); no realised correction occurred in the window. The frontier-lab accounts remain unaudited: Anthropic's IPO was reported as targeted for September or October at about US$2 trillion (Axios, 2026b), and Microsoft's own accounts recorded net gains of US$4.96 billion (US$4,963 million) on its OpenAI investments for fiscal 2026 (Microsoft, 2026), a figure that says something about the accounting treatment of OpenAI's recapitalisation and nothing reliable about OpenAI's operating losses (Appendix B item 28). June's operational conclusion stands and gained a fourth data point: build on capabilities and standards, rent products, because Sora, Operator, Instant Checkout, and now Atlas were all shut down within a year by the best-resourced labs in the field.

Employment, briefly, and worse. Stanford's Digital Economy Lab updated its "Canaries" analysis on 12 August 2026 with ADP payroll data through June 2026: employment among 22-to-25-year-olds in highly AI-exposed occupations "now stands about 19% below where it would be if it had kept pace" with less-exposed peers, from 15 per cent on Stanford's own prior estimate (the June edition carried the earlier 13 per cent figure), and "the adjustment appears to operate primarily through reduced hiring of young workers rather than increased separations" (Stanford Digital Economy Lab, 2026). McKinsey's respondents expecting job cuts from AI rose from 32 to 39 per cent, while only 14 per cent reported actual workforce reductions over the past year (McKinsey, 2026). The reading for operational leaders is unchanged and more urgent: entry-level work that consists of producing routine first drafts is being absorbed, and the development pathway for juniors needs rebuilding around verification, judgement, and client work.

A note on the series. Eleven of forty cells moved in a quarter, nine directionally. If that rate holds, the frame will need its first structural revision in about two years, not two quarters, and the categories that will need it first are the ones where "unchanged (no new evidence)" recurs: audio and voice, and business automation, where the evidence base is so thin that the register cannot tell movement from silence. The December edition should treat a third consecutive "no new evidence" in any cell as a finding about the category, not about the quarter.

---

5. Implications for operational leaders

The adoption posture for September 2026 is stated in full, because the reader should not need June to act on it. Where a capability moved between postures in the window, the sentence says so.

Deploy now, with verification built in. The capabilities that clear the production bar still share a shape: bounded task, verifiable output, human or deterministic check on low-confidence cases. In practice: document extraction with confidence gating; meeting and consult documentation under professional review, on regulator-included products; grounded knowledge assistants over curated corpora; coding assistance inside professional review processes; missed-call and after-hours voice handling with escalation; confidence-gated bookkeeping, now with a second Australian platform (MYOB) alongside Xero; drafting acceleration everywhere, with the draft never leaving the building unreviewed. New this quarter: MCP-connected tools can be built against the current specification with a twelve-month deprecation guarantee, and bounded tasks that were marginal on cost in June should be re-priced against open-weight models. Most of PAC's readership remains under-deployed on all of this.

Pilot deliberately, with falsifiable success metrics. Supervised agents for internal multi-step workflows; customer-service deflection where the knowledge base is current and someone owns it; vision-against-standards tools in parallel with existing review, not instead of it; and, new this quarter, agent-initiated card payments where a bank offers them, as a pilot with spending limits and merchant restrictions, not as a channel. Set the metric before the pilot (containment rate,

error rate against a human baseline, minutes saved measured rather than reported) and prefer vendors on outcome pricing.

Wait on unsupervised agents for consequential actions; autonomous compliance documents; agent-initiated payments (moved in part: a bank-offered, limit-bound pilot is now in the pilot posture above, and that is the only boundary this edition moves between postures); anything whose pitch depends on a vendor benchmark, which as of July 2026 includes the field's flagship coding benchmark, or an unaudited ROI figure; and any single video model as a production dependency. The cost of waiting fell again this quarter.

Walk away from "hallucination-free" claims (the incident record passed 2,000 cases in August); "AI employee" propositions sold without containment-rate evidence; token-denominated AI investment pitches (the one measured machine-payment rail lost 93 per cent of its daily volume this year); any vendor who cannot answer what happens when the system is wrong; and, new this quarter, any benchmark chart that does not name its subset.

Three structural recommendations, restated. First, treat data and documentation as the asset; every category surveyed still shows deployment success tracking corpus quality more closely than model choice, and the July failure-mode literature explains why. Second, build the verification habit into roles; the perception gap appeared again in McKinsey's survey and the Australian Government's Copilot evaluation, and individuals remain unreliable narrators of their own productivity. Third, mind the regulatory floor, which in Australia moved three times this quarter: the Fair Work Commission's AI-disclosure requirements from 20 October 2026, the TGA's listing of software as a medical device among its twelve compliance priorities for 2026-27, and the OAIC's 29 July update to its facial-recognition guidance; the Privacy Act's automated-decision-making transparency obligations commence on 10 December 2026, and the OAIC's guidance had not been published by 2 September. None of these prohibits adoption; all of them assume a human remains accountable, which is also the correct engineering assumption.

---

6. Conclusion

The honest summary of September 2026 is that the field spent a quarter proving June right in almost every category and wrong in the direction of too much optimism in the few where it moved. Eleven of forty judgements changed, nine of them directionally. The four that improved are all infrastructure: open weights got cheaper and longer, the integration standard got a current specification with a deprecation guarantee, the payment rails went live in Europe, and Australian regulators wrote more rules. The five that declined are all about trust in measurement and in products: the flagship coding benchmark is disputed by its own users, the failure modes of grounded systems are more numerous than studied, a fourth flagship product was shut down by a best-resourced lab, the machine-payment rail with real numbers lost most of them, and benchmark margins between labs became something the labs argue about in public.

For the enthusiast, the quarter delivered a US$65 billion reported run-rate, a 30-second single-pass video, a US$0.14-per-million-token model with a million-token context, and a court ruling that agents are tools rather than trespassers. For the sceptic, it delivered a 30 per cent broken benchmark, a 19 per cent employment gap for the youngest workers in exposed occupations, a 93 per cent collapse in machine-payment volume, and a survey in which 80 per cent of people feel more productive and 6 per cent of firms can show it. Both lists are true, and the quarter did not move the boundary between them.

This edition carries a date, and it will be scored in December. The statements in §2.4 are the ones to check. The sections most likely to age fastest are, on this quarter's evidence, cost and the regulatory floor, both of which moved on schedule; the section most likely to be wrong is agents, because it has not been independently measured in eighteen months and the products are moving without it.

References

36Kr. (2026, August 10). ChatGPT Atlas officially shut down less than 10 months after launch [News article]. https://eu.36kr.com/en/p/3933002676534663

Agentic AI Foundation. (2026, May 27). MCP is growing up [Blog post]. https://aaif.io/blog/mcp-is-growing-up

Anthropic. (2026, September 1). Claude Fable 5.1 and Claude Mythos 5.1 [Vendor announcement]. https://www.anthropic.com/claude-fable-and-mythos-5-1

arXiv. (2026, July 5). GASP: Grounding-aware sensitivity by perturbation for span-level hallucination detection in retrieval-augmented generation [Preprint, arXiv:2607.04223]. https://arxiv.org/abs/2607.04223

ASIC. (2026, August 19). ASIC dismantles 3,106 fraudulent crypto investment platforms in FY2026 [Regulator report, as reported by Bitcoin.com News]. https://news.bitcoin.com/regulation-and-legal/asic-pulls-3106-crypto-scams-as-ai-networks-target-aussie-savings/

Australian Competition and Consumer Commission (ACCC). (2026a, June 5). Thousands of scam websites taken down as online scams continue to cost Australians [Media release]. https://www.accc.gov.au/media-release/thousands-of-scam-websites-taken-down-as-online-scams-continue-to-cost-australians

Axios. (2026a, July 24). Anthropic releases new model, Claude Opus 5 [News article]. https://www.axios.com/2026/07/24/anthropic-releases-new-model-opus-5

Axios. (2026b, August 17). Anthropic revenue run-rate, IPO and OpenAI [News article]. https://www.axios.com/2026/08/17/anthropic-revenue-run-rate-ipo-openai

Bloomberg. (2026a, July 27). Nvidia's $750 billion in deals reignite circular AI fears [News article; headline captured, body not fetched]. https://www.bloomberg.com/news/articles/2026-07-27/nvidia-s-750-billion-deals-revive-fear-of-ai-circular-financing

Bloomberg. (2026b, August 20). US stock market bubble: AI boom won't prevent a sharp correction [Opinion column; headline captured, body not fetched]. https://www.bloomberg.com/opinion/articles/2026-08-20/us-stock-market-bubble-ai-boom-won-t-prevent-a-sharp-correction

Charlotin, D. (2026). AI hallucination cases database [Community-maintained database; 2,005 entries as at 31 August 2026]. https://www.damiencharlotin.com/hallucinations/

Clayton Utz. (2026, July 24). Healthcare software: TGA clarifies when AI will be regulated as a medical device [Law-firm commentary]. https://www.claytonutz.com/insights/2026/july/healthcare-software-tga-clarifies-when-ai-will-be-regulated-as-a-medical-device

Clinical Trial Vanguard. (2026, August 18). FDA has authorized 1,500 AI medical devices. The evidence for most of them is still catching up [Industry commentary]. https://www.clinicaltrialvanguard.com/opinion/fda-has-authorized-1500-ai-medical-devices-the-evidence-for-most-of-them-is-still-catching-up/

CNBC. (2026, July 22). Alphabet Q2 2026 earnings [News article; figure corroborated via secondary headlines, primary page returned 403]. https://www.cnbc.com/2026/07/22/google-earnings-q2-goog-live-updates.html

Cooley. (2026, August 3). EU AI Act transparency obligations take effect 2 August 2026 [Law-firm commentary]. https://www.cooley.com/news/insight/2026/2026-08-03-eu-ai-act-transparency-obligations-take-effect-2-august-2026

Digital Transformation Agency. (n.d.). Whole-of-government Microsoft 365 Copilot trial: Summary of evaluation findings, executive summary [Government evaluation; publication date not recoverable from fetched page]. https://www.digital.gov.au/initiatives/copilot-trial/summary-evaluation-findings/cts-executive-summary

ElevenLabs. (2026a, May 5, updated August 31). $500M ARR and new investors [Vendor blog post; self-interested]. https://elevenlabs.io/blog/500m-arr-and-new-investors

ElevenLabs. (2026b, August 10). Changelog [Vendor documentation]. https://elevenlabs.io/docs/changelog/2026/8/10

Employment Hero. (2026, February 17). At A$300M ARR, AI-powered Employment Hero enters next phase as SEEK Growth Fund launches stake sale [Vendor press release; self-interested]. https://employmenthero.com/blog/at-a300m-arr-ai-powered-employment-hero-enters-next-phase-as-seek-growth-fund-launches-stake-sale/

European Commission. (2026a). Regulatory framework for AI [Official policy page; records AI Omnibus Regulation entry into force 27 July 2026]. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission. (2026b, July 31). Commission starts enforcing AI Act rules and new transparency requirements from 2 August [Press release]. https://digital-strategy.ec.europa.eu/en/news/commission-starts-enforcing-ai-act-rules-and-new-transparency-requirements-2-august

Extend. (2026, July 7). OCR benchmarks and real-world documents [Vendor article; self-interested]. https://www.extend.ai/resources/ocr-benchmarks-real-world-documents

Fair Work Commission. (2026, August 26). Use of AI in Commission cases [Regulator guidance; requirements effective 20 October 2026]. https://www.fwc.gov.au/about-us/news-and-media/news/use-ai-commission-cases

Federal Register. (2026, August 18). GENIUS Act regulations on payment stablecoin issuance, offer and sale (Notice of proposed rulemaking, 91 FR 53368). https://www.federalregister.gov/documents/2026/08/18/2026-16796/genius-act-regulations-on-payment-stablecoin-issuance-offer-and-sale

FelloAI. (2026). AI model discontinuation tracker [Aggregator; used only to corroborate the Atlas shutdown, which 36Kr independently reports]. https://felloai.com

Fortune. (2026, June 16). Voice AI startup Bland raises $50 million after being rejected by 180 investors [News article]. https://fortune.com/2026/06/16/voice-ai-bland-50-million-after-being-rejected-by-180-investors/

Gartner. (2026, September 1). Gartner survey finds only 22% of organizations have successfully scaled AI across multiple business units [Analyst press release; self-report]. https://www.gartner.com/en/newsroom/press-releases

GitHub. (2026). GitHub changelog [Vendor documentation]. https://github.blog/changelog/

Google. (2026a, June 8). Better research with NotebookLM [Vendor blog post]. https://blog.google/innovation-and-ai/products/notebooklm/better-research-notebooklm/

Google. (2026b). Gemini API changelog [Vendor documentation; records shutdown dates for Veo 2.0, Veo 3.0 (30 June 2026) and Imagen 4 variants (17 August 2026)]. https://ai.google.dev/gemini-api/docs/changelog

Industry Spread. (2026, July 6). Visa agentic payments live with 30 European issuers [Trade press]. https://theindustryspread.com/visa-agentic-payments-live-30-european-issuers/

InfoWorld. (2026, July 9). OpenAI launches ChatGPT Work as it broadens GPT-5.6 rollout [Trade press]. https://www.infoworld.com/article/4195478/openai-launches-chatgpt-work-as-it-broadens-gpt-5-6-rollout.html

Investing.com. (2026, July 29). Meta Q2 2026 slides: Revenue surges 28% as AI spending pressures margins [Financial press]. https://www.investing.com/news/company-news/meta-q2-2026-slides-revenue-surges-28-as-ai-spending-pressures-margins-93CH-4821943

InVideo. (2026). Grok Imagine AI video generator [Secondary vendor-adjacent article; xAI pricing as reported]. https://invideo.io/blog/grok-imagine-ai-generator/

Kalkine. (2026, August 6). REA Group reports FY26 results with revenue growth, AI product expansion and strategic acquisitions [Financial press]. https://kalkine.com.au/news/technology/rea-group-asx-rea-reports-fy26-results-with-revenue-growth-ai-product-expansion-and-strategic-acquisitions

Laika Labs. (2026, April 21). ASI:Chain testnet launch milestone [Vendor-adjacent news]. https://laikalabs.ai/news/asi-chain-testnet-launch-milestone

MarkTechPost. (2026, August 26). Alibaba's Qwen team releases Qwen3.8-Flash-Next [Trade press]. https://www.marktechpost.com/2026/08/26/alibabas-qwen-team-releases-qwen3-8-flash-next-a-125b-multimodal-moe-with-6b-active-parameters-previewing-the-qwen4-architecture/

McKinsey & Company. (2026, August 25). The state of AI in 2026 [Consultancy survey; self-report; fieldwork 4 May to 8 June 2026, 1,719 respondents]. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

MCP.film. (2026). Higgsfield MCP [Aggregator listing; roster verified through 11 August 2026]. https://mcp.film/mcps/higgsfield/

METR. (2026, July 21). Expenditure horizon [Independent evaluation with published method]. https://metr.org/blog/2026-07-21-expenditure-horizon/

Microsoft. (2026, July 29). Microsoft fiscal year 2026 fourth quarter earnings press release [Investor filing]. https://www.microsoft.com/en-us/investor/earnings/fy-2026-q4/press-release-webcast

MLQ. (2026, August 14). Zhipu releases GLM-5.3 through its coding service with weights still two weeks away [Trade press]. https://mlq.ai/news/zhipu-releases-glm-53-through-its-coding-service-with-weights-still-two-weeks-away/

Model Context Protocol. (2026a). Specification 2026-07-28 [Standards document; current protocol version]. https://modelcontextprotocol.io/specification/2026-07-28

Model Context Protocol. (2026b). Versioning and feature lifecycle [Standards documentation]. https://modelcontextprotocol.io/specification/versioning

Morgan Lewis. (2026, August 12). EU AI Act's transparency rules: What went into effect on 2 August [Law-firm commentary]. https://www.morganlewis.com/blogs/sourcingatmorganlewis/2026/08/eu-ai-acts-transparency-rules-what-went-into-effect-on-2-august

Music Business Worldwide. (2026a, June 8). Musicians' union sues UMG and Warner Music alleging member recordings were licensed to Suno and Udio without compensation or credit [Trade press]. https://www.musicbusinessworldwide.com/musicians-union-sues-umg-and-

warner-music-alleging-member-recordings-were-licensed-to-suno-and-udio-without-compensation-or-credit/

Music Business Worldwide. (2026b, July 6). After a judge denied Sony's move to expand its Udio case, Suno asks court to reject UMG and Sony's bid to add 61k recordings [Trade press]. https://www.musicbusinessworldwide.com/after-a-judge-denied-sonys-move-to-expand-its-udio-case-suno-asks-court-to-reject-umg-and-sonys-bid-to-add-61k-recordings/

Nansen. (2026, July 27). NEAR quarterly report Q2 2026 [Analytics report]. https://nansen.ai/post/nansen-near-quarterly-report-q2-2026

Nextgov. (2026, June 4). Lawmakers propose AI framework that would preempt state laws for 3 years [Trade press]. https://www.nextgov.com/artificial-intelligence/2026/06/lawmakers-propose-ai-framework-would-preempt-state-laws-3-years/413975/

Ninth Circuit. (2026, August 4). Amazon.com Services, LLC v. Perplexity AI, Inc., No. 26-1444 (9th Cir.) [Court opinion]. https://cdn.ca9.uscourts.gov/datastore/opinions/2026/08/04/26-1444.pdf

Office of the Australian Information Commissioner (OAIC). (2026a). Consultation on guidance for transparency in automated decision-making [Regulator consultation; closed 15 June 2026]. https://www.oaic.gov.au/engage-with-us/consultations/consultation-on-guidance-for-transparency-in-automated-decision-making

Office of the Australian Information Commissioner (OAIC). (2026b). Newsroom [Regulator media releases, checked to 2 September 2026]. https://www.oaic.gov.au/newsroom

Office of the Australian Information Commissioner (OAIC). (2026c, July 29). Facial recognition technology: A guide to assessing the privacy risks [Regulator guidance]. https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/organisations/facial-recognition-technology-a-guide-to-assessing-the-privacy-risks

Open Source For You. (2026, August 3). DeepSeek open-sources V4-Flash [Trade press]. https://www.opensourceforu.com/2026/08/deepseek-open-sources-v4-flash/

OpenAI. (2026a, July 9). Introducing GPT-5.6 [Vendor announcement]. https://openai.com/index/gpt-5-6/

OpenAI. (2026b, July 8). Separating signal from noise: Coding evaluations [Vendor research note; self-interested]. https://openai.com/index/separating-signal-from-noise-coding-evaluations/

OpenAI. (2026c). What to know about the Sora discontinuation [Vendor help centre]. https://help.openai.com/en/articles/20001152-what-to-know-about-the-sora-discontinuation

OWASP. (2026, August 3). OWASP GenAI LLM Top 10 2026 [Industry standards body]. https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

Perth AI Consulting. (2026). The State of AI in mid-2026: A literature review for operational leaders (Version 1.1). https://perthaiconsulting.com.au/resources/state-of-ai/mid-2026/ (PDF:

https://perthaiconsulting.com.au/resources/state-of-ai-mid-2026-literature-review-v1.1.pdf). Primary sources cited in this edition only as carried forward from the previous edition (Epoch AI; Carnegie Mellon's TheAgentCompany; METR's 2025 and February 2026 trials; Stanford RegLab; XLANG's OSWorld; the OmniAI benchmark; arXiv:2509.25498; SemiAnalysis; Escape.tech; DORA 2025; Stack Overflow 2025; GitClear; APQC; the Linux Foundation's MCP announcement; the Federal Court's GPN-AI) are listed in that edition's reference list and are not repeated here.

PYMNTS. (2026, August 7). Harvey targets $15.5 billion valuation as revenue surges past $350 million [Trade press]. https://www.pymnts.com/news/investment-tracker/2026/harvey-targets-15-5-billion-valuation-as-revenue-surges-past-350-million/

Rework. (2026). Microsoft Build 2026: Windows agent platform and store, what CTOs need to know [Trade commentary]. https://resources.rework.com/news/ai-at-work/microsoft-build-2026-windows-agent-platform-store-cto

Sacra. (2026). Cursor [Independent estimate; acquisition report unverified]. https://sacra.com/c/cursor/

Salesforce. (2026, August 26). Salesforce delivers record second quarter fiscal 2027 results [Investor release]. https://investor.salesforce.com/news/news-details/2026/Salesforce-Delivers-Record-Second-Quarter-Fiscal-2027-Results/default.aspx

Scale AI. (2026). SWE-bench Pro public leaderboard [Independent evaluator leaderboard, fetched 2 September 2026]. https://labs.scale.com/leaderboard/swe_bench_pro_public

Seeking Alpha. (2026, July 30). Amazon outlines Q3 net sales of $197B-$202B while lifting 2026 cash capex to about $220B [Financial press; headline captured, body not fetched]. https://seekingalpha.com

Skadden. (2026, June). Colorado repeals and replaces its AI Act [Law-firm commentary]. https://www.skadden.com/insights/publications/2026/06/colorado-repeals-and-replaces-its-ai-act

SmartCompany. (2026, August 26). Fair Work Commission cracks down on dodgy generative AI in cases [Business press]. https://www.smartcompany.com.au/artificial-intelligence/fair-work-commission-cracks-down-dodgy-generative-ai-cases/

Stack Overflow. (2026, June 23). The 2026 Developer Survey is now open, for human developers only [Survey announcement]. https://stackoverflow.blog/2026/06/23/the-2026-developer-survey-is-now-open-for-human-developers-only/

Stanford Digital Economy Lab. (2026, August 12). Canaries in the coal mine: August 2026 update [Independent research with published method]. https://digitaleconomy.stanford.edu/news/canariesaug26/

TechCrunch. (2026a, July 9). OpenAI is shutting down Atlas, but its AI browser ambitions are still growing [Technology press]. https://techcrunch.com/2026/07/09/openai-is-shutting-down-atlas-but-its-ai-browser-ambitions-are-still-growing/

TechCrunch. (2026b, July 21). Google releases three new Gemini models, but no 3.5 Pro [Technology press]. https://techcrunch.com/2026/07/21/google-releases-three-new-gemini-models-but-no-3-5-pro/

TechCrunch. (2026c, August 17). Anthropic's annualized revenue surges to $65B [Technology press; Bloomberg-sourced, not company-confirmed]. https://techcrunch.com/2026/08/17/anthropics-annualized-revenue-surges-to-65b/

TechCrunch. (2026d, July 7). The coding agent wars are spilling into the rest of the office: Claude Cowork [Technology press]. https://techcrunch.com/2026/07/07/the-coding-agent-wars-are-spilling-into-the-rest-of-the-office-claude-cowork/

TechCrunch. (2026e, August 25). Claude Cowork finally remembers what you told the app in chat [Technology press]. https://techcrunch.com/2026/08/25/claude-cowork-finally-remembers-what-you-told-the-app-in-chat/

TechTimes. (2026a, July 6). Mistral AI targets frontier gap with open-weight model entering July early access [Technology press]. https://www.techtimes.com/articles/319798/20260706/mistral-ai-targets-frontier-gap-open-weight-model-entering-july-early-access.htm

TechTimes. (2026b, July 16). Seedance 2.5 API live: ByteDance's 30-second AI video carries unresolved copyright risk [Technology press]. https://www.techtimes.com/articles/320683/20260716/seedance-25-api-live-bytedances-30-second-ai-video-carries-unresolved-copyright-risk.htm

TrustNLP. (2026, July 4). A systematic taxonomy of failure modes in retrieval-augmented generation systems [Peer-reviewed workshop paper, ACL 2026 TrustNLP]. https://aclanthology.org/2026.trustnlp-main.27.pdf

VentureBeat. (2026, June 22). Alibaba's AI video model rises to No. 2 in global rankings as OpenAI's Sora and ByteDance's Seedance fall away [Technology press]. https://venturebeat.com/technology/alibabas-ai-video-model-rises-to-no-2-in-global-rankings-as-openais-sora-and-bytedances-seedance-fall-away

Willison, S. (2026a, July 9). GPT-5.6 [Independent commentary]. https://simonwillison.net/2026/Jul/9/gpt-5-6/

Willison, S. (2026b, July 16). Kimi K3 [Independent commentary]. https://simonwillison.net/2026/Jul/16/kimi-k3/

Xero. (2026, August 19, updated August 27). Xerocon Denver 2026: Agentic workflows for the AI era [Vendor blog post; self-interested]. https://blog.xero.com/product-updates/xerocon-denver-2026-agentic-workflows-ai-era/

Yahoo Finance. (2026a, August 13). x402 settlement volume plunges 93% [Financial press, citing Helios Analytics]. https://finance.yahoo.com/markets/crypto/articles/x402-settlement-volume-plunges-93-105710906.html

Yahoo Finance. (2026b, March 4). Anthropic ARR surges to $19 billion [Financial press; company-reported figure]. https://finance.yahoo.com/news/anthropic-arr-surges-19-billion-151028403.html

Appendix A: Methodology and verification approach

Research process. This edition was researched on 2 September 2026 through seven parallel research streams, each briefed with the previous edition's text for its sections, the previous edition's Appendix B, the evidence window (14 June to 2 September 2026), and the source hierarchy below. Streams were instructed to fetch the primary page for anything load-bearing, to date and source-label every claim, to label vendor-originated figures as vendor claims, and to report explicitly, with the queries run, wherever a question returned nothing meeting the standard. Those nulls are the evidential basis for every "unchanged (no new evidence)" classification in the movement register. The coordinator drafted from the streams' raw notes without intermediate summarisation, then filled the most load-bearing gaps by direct fetch of primary pages (the Microsoft FY2026 earnings release, the MCP specification, OpenAI's SWE-bench Pro analysis, the Ninth Circuit opinion, the Stanford Digital Economy Lab update, the McKinsey survey page, METR's expenditure-horizon post, Anthropic's Fable 5.1 page).

Source hierarchy. Profile 1 from the series' method (established field): peer-reviewed studies and randomised trials; independent evaluators with published method (METR, Epoch AI, Stanford HAI and RegLab, DORA, Arc Prize, Scale AI's public leaderboards); government and regulator publications; externally auditable data (on-chain figures, SEC and ASX filings, public dashboards); reputable press; vendor documentation and announcements, used for what a product is and claims, never for how well it works; and aggregator or SEO content, used only to locate primary sources. Where a figure could be traced only to an aggregator, it appears in Appendix B rather than the body.

Movement classifications. Each of the forty judgement cells was classified by the coordinator from the streams' proposed classifications and the evidence behind them. The rules: a cell is improved or declined only on evidence dated inside the window and meeting the hierarchy above, with the direction defined per column as set out in §1.4 and Appendix D; a cell is unchanged (confirmed) only where in-window evidence of the required standard supports the June judgement (pre-window evidence, however recent, does not confirm); a cell is unchanged (no new evidence) where the streams' searches returned nothing meeting the standard, and the queries are listed below so the null is checkable; a cell is reframed where the judgement stands but the reason or boundary moved. Every classification carries a citation in the section's movement table.

Date discipline and vendor labelling. Every capability, performance, and commercial claim carries a date. Anything the previous edition already covered is labelled baseline evidence even where re-reported in the window; three pre-window items that the previous edition missed are

listed separately in the front matter rather than absorbed. Vendor figures (benchmark scores, ARR, call volumes, transaction counts, time savings, resolution rates) are labelled in the body text where they appear.

Known limitations. (1) The research session's shared web-search budget was exhausted before every stream finished. Twelve planned queries could not be run (listed below as "not executed"); the streams affected closed those questions by direct fetch of regulator and vendor pages, which is a weaker discovery method. The nulls most weakened are: ASIC statements on AI-assisted advice (§3.10); Adviser Ratings' 2026 adoption data (§3.10); vision failure-mode research (§3.3); Australian AI Safety Institute status (§2.4); graduate hiring data (§4); Australian government AI procurement (§4); Stripe and PayPal agentic-commerce adoption (§3.9). (2) Paywalled analyst reports, private deployment data, and two systematic reviews that could not be fetched were not accessible. (3) Independent evaluation continues to lag product release; the newest models in this edition are assessed on vendor evidence, stated as such. (4) Two lab benchmark disputes fell inside the window; where the labs disagree on a benchmark's validity, this edition reports both positions and treats the benchmark as contested. (5) The authors used AI systems (Claude Fable 5.1 for research coordination and drafting; a separate Claude Opus pass for the independent fact-check) with human direction of scope, structure, and judgement; this is disclosed in the spirit of the disclosure standards the paper discusses.

Queries run, by stream. Listed so that every null in the register can be re-run.

Stream A (frontier models and code), 28 queries: Anthropic Claude new model release July 2026; OpenAI GPT-5.6 release August 2026; Google Gemini 3.5 release 2026; xAI Grok new model release 2026; Anthropic Claude Opus 5 pricing context window per million tokens; ARC-AGI-2 leaderboard GPT-5.6 Gemini Opus 5 score August 2026; "Gemini 3.5 Pro" release date August 2026; Anthropic Claude Fable 5.1 Mythos 5.1 release; Epoch AI benchmark frontier models August 2026; DeepSeek V4 new model release 2026 R2; Qwen 3.5 Alibaba release 2026; Moonshot Kimi K2 new release 2026; Mistral Large 3 update new model July August 2026; DeepSeek V4 released date 2026 parameters license MIT; Qwen new model release August 2026; GLM Zhipu new model release 2026; Meta Llama new model release 2026 Muse Spark update; Claude Code ARR revenue 2026 Anthropic; Cursor ARR revenue 2026 valuation; METR study coding agents 2026 follow-up productivity; Stack Overflow 2026 developer survey results AI; GitHub Copilot Agent HQ update 2026 general availability; Replit Lovable revenue ARR update July August 2026; OpenAI Codex update 2026 Google Jules update; Stack Overflow 2026 developer survey results published July results; GLM-5.3 release date Zhipu Z.ai; Meta Muse Spark 1.1 July 9 2026 pricing context window; SpaceX acquire Cursor Anysphere $60 billion June 2026 (not executed: budget exhausted).

Stream B (agents and business automation), 35 queries: METR task horizon update 2026 doubling time; TheAgentCompany benchmark rerun 2026 CMU; OSWorld-Verified leaderboard August 2026; Claude Cowork launch 2026 Anthropic; Amazon Perplexity Comet Ninth Circuit ruling 2026; Microsoft Windows Agent Runtime ship 2026; GAIA benchmark leaderboard 2026 agent; Salesforce Agentforce Q2 FY27 earnings customers 2026; MCP registry server count

2026 official; Salesforce Ben Agentforce adoption report 2026; OWASP agentic AI security top 10 2026; NIST AI agent security guidance 2026; prompt injection agent browser exploit 2026 ChatGPT Atlas Comet; APS Australian Public Service Copilot trial results 2026; digital.gov.au Copilot trial evaluation report published date 2026; tau-bench BrowseComp agent benchmark results 2026; HubSpot Breeze Customer Agent pricing update 2026; AI receptionist Australia missed call 2026 startup; Zapier Agents n8n Make.com update 2026; "Project Mariner" OR "Gemini agent mode" update July August 2026; ChatGPT agent mode update July 2026 OpenAI; Intercom Fin resolution rate independent evaluation 2026; MCP spec release 2026 Linux Foundation Agentic AI Foundation; "ChatGPT Work" OpenAI launch GPT-5.6 announcement; Google Project Mariner shut down retired 2026; Atlas browser OpenAI sunset discontinued 2026; Zendesk AI resolution rate 2026 news; GoHighLevel independent audit review 2026; MCP official registry 2000 servers modelcontextprotocol.io count; Mustahsan multi-run consistency benchmark agent 2026 follow-up; Windows Agent Store GA 2026 shipped; MCP server exploit disclosed vulnerability 2026; Windows Agent Store general availability launched 2026; Comet Perplexity free platforms adoption users 2026; "OSWorld" score August 2026 Claude Gemini GPT-5.6 top.

Stream C (vision; medical, property, construction verticals), 38 queries: OmniAI benchmark document extraction 2026 update; OmniDocBench 2026 new results; Heidi Health funding 2026 Series C ARR; TGA AI scribe ARTG clarification January 2026; MASAI trial mammography AI 2026 follow-up study; FDA AI-enabled medical devices list count 2026; "OmniAI" OCR benchmark leaderboard update 2026 field accuracy; Defect Detector NCC compliance Australia 2026; Voltin facade inspection AI Australia 2026; Florida AI plan review software 2026 code compliance; CodeComply.AI PlanCheckPro accuracy evaluation 2026; Buildots OpenSpace construction progress tracking accuracy 2026; Abridge funding 2026 valuation Series E; Microsoft Dragon Copilot 2026 update clinicians study; AI scribe peer-reviewed study burnout time saved 2026; REA Group Domain generative AI feature 2026; Heidi Health 2026 news wearable microphone Series; Lyrebird Health 2026 funding update; RACGP AHPRA AI scribe guidance update 2026; multimodal model benchmark counting chart reading hallucination 2026 evaluation; Heidi Health Series C 2026 valuation raise; "Abridge" April 2026 extended funding valuation; EagleView aerial measurement accuracy 2026; specialist OCR model release 2026 character accuracy benchmark; RICS AVM standard 2026 update responsible AI surveyors; pathology AI dermatology AI diagnostic study 2026 external validation systematic review; CoreLogic PropTrack AVM accuracy 2026 Australia; document extraction LLM accuracy real documents study August 2026; FDA artificial intelligence enabled medical devices list count August 2026; DermaSensor SkinVision 2026 study accuracy; radiology AI external validation systematic review 2026 site generalization; "FDA" AI device list "1,300" OR "1,400" OR "1,500" 2026 authorized; TGA "digital scribe" ARTG requirement compliance July August 2026; Australia AI vision plan review NCC compliance tool 2026 (not executed: budget exhausted); vision language model counting benchmark 2026 arxiv failure (not executed: budget exhausted); chart QA benchmark 2026 multimodal models results (not executed: budget exhausted); Heidi Health Series C valuation billion 2026 (not executed: budget exhausted); PropTrack AI valuation feature 2026 REA (not executed: budget exhausted).

Stream D (video, image, audio, music), 33 queries: Sora API sunset September 2026 OpenAI update; Google Veo 3.1 pricing update July August 2026; Runway Gen-4.5 update August 2026; Kling 3.0 update July 2026 new features; ByteDance Seedance Alibaba Wan video model 2026; Midjourney video model 2026 release; Alibaba Wan HappyHorse video model June 2026 Artificial Analysis; Seedance 2.0 international rollout shelved copyright 2026; Gemini 3 Pro Image Nano Banana Pro update pricing August 2026; FLUX.2 update Black Forest Labs 2026; Ideogram new model 2026 release; OpenAI image model GPT-image update 2026; "Ideogram 4.0" release date 2026; YouTube monetization policy synthetic content update 2026; eSafety Commissioner deepfake synthetic media 2026; Australia deepfake legislation 2026 passed; ACMA synthetic media misinformation 2026; Meta AI generated content label policy 2026 update; TikTok synthetic media AI content policy 2026; Suno Sony Universal lawsuit ruling 2026; musicians union lawsuit majors Suno Udio settlement 2026; AI training copyright fair use ruling court 2026; ElevenLabs funding valuation 2026 ARR; FCC AI voice robocall ruling 2026; Scamwatch ACCC voice cloning scam losses 2026; Bland AI funding round valuation 2026; voice agent containment rate independent study 2026; transcription word error rate benchmark accent 2026 study; Higgsfield MCP update 2026; xAI Grok Imagine video model 2026; Luma Pika AI video 2026 update; ElevenLabs news August 2026; Vapi Retell Bland AI funding 2026.

Stream E (knowledge management; legal, advice, accounting, HR, trades), 31 queries: NotebookLM new features July 2026; Claude Projects ChatGPT Projects memory update August 2026; Microsoft Copilot knowledge governance enterprise report 2026; RAG failure mode benchmark study 2026; arXiv 2509.25498 follow-up citation grounded assistant hallucination; long context vs RAG benchmark needle multi-fact 2026; Obsidian local AI update 2026; APQC knowledge management AI survey 2026; Gartner enterprise search knowledge AI adoption 2026 survey; McKinsey state of AI 2026 knowledge management; Harvey legal AI funding valuation 2026; CoCounsel Lexis+ AI Westlaw new hallucination study 2026; Australian lawyer sanctioned AI fake citations 2026; "AI hallucination cases" database count 2026 lawyer; Federal Court Australia GPN-AI practice note amendment 2026; Law Society WA Legal Practice Board AI guidance 2026; Xero JAX auto bank reconciliation general availability 2026; Intuit Assist QuickBooks new feature 2026; Dext Fathom Spotlight Reporting AI feature 2026; MYOB AI feature 2026 announcement; Employment Hero valuation 2026 funding AI features; Fair Work Commission Fair Work Ombudsman AI guidance HR 2026; Australia case AI generated HR document unfair dismissal 2026; ServiceM8 Tradify AroFlo Fergus Simpro AI feature 2026 changelog; ASIC AI financial advice enforcement 2026 (not executed: budget exhausted; regulator site fetched directly); Adviser Ratings AI adoption survey 2026 financial advisers (not executed: budget exhausted; regulator site fetched directly); ASIC AI-drafted Statement of Advice SOA enforcement action (not executed: budget exhausted); OAIC AI privacy guidance automated decision-making 2026 (search budget exhausted; fell back to direct WebFetch of oaic.gov.au newsroom).

Stream F (crypto and AI convergence), 25 queries: NEAR Intents volume Dune Analytics 2026; NEAR IronClaw confidential compute audit August 2026; market.near.ai agent marketplace metrics 2026; x402 Coinbase Cloudflare transactions volume August 2026; Stripe Tempo

Machine Payments Protocol update 2026; Visa Intelligent Commerce Mastercard Agent Pay production 2026; "Agent Pay" Mastercard pilot production merchants 2026; GENIUS Act stablecoin implementation 2026; RBA stablecoin Treasury Australia regulation 2026; AUDD Australian stablecoin 2026; Fetch.ai ASI chain launch 2026; Bittensor revenue analysis August 2026; Virtuals Protocol agent economy revenue 2026; SEC ASIC AUSTRAC AI agent payments crypto regulation 2026; NEAR Intents cumulative volume "$30 billion" OR "$25 billion" 2026; NEAR AI enterprise customer confidential compute IronClaw partnership; OpenAI Instant Checkout successor agentic commerce 2026; Amazon "Buy for Me" Shopify PayPal agentic checkout update 2026; OpenAI ChatGPT commerce shopify Etsy checkout July August 2026; AP2 FIDO Alliance working group agentic payments update July 2026; Microsoft Copilot Checkout volume 2026; (not executed: budget exhausted) Stripe Agentic Commerce Protocol merchants adoption August 2026; (not executed: budget exhausted) agentic commerce report merchant conversion data 2026; (not executed: budget exhausted) PayPal agentic commerce AI checkout 2026 launch; (not executed: budget exhausted) Coinbase x402 foundation dashboard Cloudflare agent payments July 2026.

Stream G (cross-cutting frame), 28 queries: EU AI Act high-risk obligations postponement formally adopted 2026; EU AI Act transparency obligations August 2026 commence; Australian AI Safety Institute established 2026; Privacy Act automated decision-making December 2026 OAIC guidance; US federal preemption state AI laws Congress 2026; Anthropic IPO S-1 filing valuation 2026; Anthropic IPO public filing August 2026; OpenAI IPO listing 2026 progress; Microsoft Alphabet Amazon Meta Q2 2026 earnings capex guidance July 2026; GPT-5.6 OR "GPT-5.5" OR next flagship model announced July August 2026; Gemini 3 next flagship model release August 2026; Claude Opus 4.5 OR Claude 5 release date 2026; Anthropic revenue run-rate 2026 annualized; OpenAI Microsoft SEC filing losses Q2 2026; AI bubble market correction August 2026; circular financing AI Nvidia OpenAI Anthropic 2026 Bloomberg; Microsoft earnings July 2026 capex fiscal Q4 guidance; Alphabet Q2 2026 earnings capital expenditure guidance raised; Meta Q2 2026 earnings capex AI spending guidance; Amazon Q2 2026 earnings AWS capex AI; McKinsey State of AI 2026 report survey enterprise; Stanford Digital Economy Lab Canaries update 2026 employment AI; OpenAI Anthropic Google product shutdown discontinued July August 2026; Colorado AI Act effective date 2026 delayed; California AI law SB 53 effective 2026 Texas TRAIGA effective January 2026; National AI Centre Australia SMB adoption survey 2026; Australian government AI procurement APS policy 2026 (not executed: budget exhausted); graduate hiring AI 2026 report entry-level jobs (not executed: budget exhausted).

Appendix B: What could not be verified at writing time

This appendix is the series' standing watchlist. Part 1 revisits, item by item, the thirty claims the previous edition could not verify to its standard as at 13 June 2026 (numbering as in that edition's Appendix B; the previous edition's own second fact-check pass had already resolved

some in place, and that status is noted). Part 2 opens new items encountered in this window. Presence here is not an assertion that a claim is false.

Part 1: June's thirty items, revisited

#	JUNE ITEM (ABBREVIATED)	JUNE STATUS	SEPTEMBER 2026 DISPOSITION
1	Parameter counts for current closed models	Unverifiable; SEO only	**Unchanged.** No vendor disclosed parameters in the window; Kimi K3's 2.8 trillion is a vendor figure for an open-weight model.
2	Existence of "GPT-5.6" and "Gemini 3.2 Pro"	Split: 5.6 anticipated, 3.2 Pro not real	**Now verified (half).** GPT-5.6 shipped 9 July 2026 (OpenAI, 2026a). Gemini 3.2 Pro still does not exist; Google shipped no Pro-tier model in the window (TechCrunch, 2026b).
3	Llama 4 Behemoth shipped?	Resolved in June's Pass 2: not shipped	**Closed; no longer relevant.** Meta's Muse Spark 1.1 confirms a closed, API-only line; the open-weight Behemoth question is moot.
4	Mythos 5 / Fable 5 cyber capabilities	Vendor claim, no replication	**Unchanged.** No independent replication; Google's Gemini 3.5 Flash Cyber (21 July) is a parallel, equally unverifiable vendor claim.
5	"Claude Code writes 4% of public GitHub commits"	Traces to SemiAnalysis; no replication	**Unchanged.** No independent replication located.
6	"97 million monthly MCP SDK downloads"	Untraceable	**Unchanged.** No primary download figure in the window.
7	Nerq "17,468 public MCP servers"	Self-published census	**Unchanged, and worse.** Counts in the window ranged from about 2,000 (official registry) to 81,055 (Glama scrape); no authoritative census exists.
8	2026 rerun of TheAgentCompany	Confirmed not to exist	**Unchanged.** Still no rerun; re-checked.
9	OSWorld scores for current models	Resolved in Pass 2 as a variant artefact	**Weakened.** The June figures (about 70 standard, 78 to 80 Verified) could not be re-verified against a live XLANG page; the only fetchable primary page was dated July 2025 with lower scores, and vendors now report against "OSWorld 2.0 (strict)" (Anthropic, 2026). The June figures are carried with that caveat.
10	UF plan-review "30 to 60 days to 3 to 5 days"	Refuted in June's Pass 2	**Unchanged (remains refuted).** No new evidence.
11	"AVM 2.0" 94.2 per cent within 5 per cent	Orphaned figure	**Unchanged.** No primary source located.
12	Post-January-2026 Veo pricing; "Veo 3.1 Lite"	SEO only	**Unchanged.** Only aggregator cost calculators found; no Google pricing page fetched.
13	Sora 2 maximum clip length	Resolved in Pass 2 as a timeline	**Closed; no longer relevant.** Sora is discontinued (API sunset 24 September 2026).

#	JUNE ITEM (ABBREVIATED)	JUNE STATUS	SEPTEMBER 2026 DISPOSITION
14	Higgsfield "replaces a $5K/month retainer"	Vendor marketing copy	**Unchanged.** The MCP integration remains live and updated (Sora removed, 11 August); no economic evidence.
15	Aggregate voice-fraud statistics	Untraceable; Arup case verified	**Unchanged.** ACCC's June release carries no voice-cloning breakout; no FBI IC3 2026 report published.
16	AI receptionist ROI statistics	Vendor-commissioned lineage	**Unchanged.** One Australian entrant's press-wire claim (answrr) continues the pattern.
17	2025 ChatGPT memory-wipe incidents	Partially corrected	**Unchanged.** Not researched this window; not load-bearing.
18	Pinecone 12-deployment RAG study	No primary publication	**Unchanged.** Not re-examined; not load-bearing.
19	Salesforce 2025 consumer-preference survey	Vendor source, substance corroborated	**Unchanged.** No new Salesforce State of Service report located.
20	APQC "85 per cent not operationalised"	Traced to a sponsored survey	**Unchanged.** APQC's 2026 report (7 January, pre-window) was fetched only partially; the instrument was not examined.
21	NEAR Agent Market volumes	No published data	**Unchanged.** Still no published marketplace metrics; NEAR Intents volume (US$22 billion cumulative) and confidential routing (US$30 million TVL) are separately reported (Nansen, 2026).
22	Tempo / Machine Payments Protocol throughput	Vendor capacity claim only	**Unchanged.** No throughput figure, vendor or neutral, in the window.
23	Bittensor "$43M quarterly revenue"	Both sides verified; dispute real	**Unchanged.** Only re-reporting of the same figures; no new quarter's data.
24	Heidi Health Series C; US$21.9M ARR	No Series C; ARR figure not located	**Unchanged.** No Series C by 2 September 2026; the ARR figure was again not located.
25	Independent benchmarks for Xero JAX, Intuit Assist, Dext, ServiceM8	Confirmed absence	**Unchanged.** Xero's "100 million transactions" (19 August) is a scale figure, not an accuracy figure; MYOB's suite is equally unbenchmarked.
26	Spotlight Reporting 2026 AI features	Not located	**Unchanged.** Not located.
27	REA Group or Domain 2026 generative features	Imprecisely dated only	**Now verified (REA).** REA's FY26 results (6 August 2026) name a Consumer AI Assistant, Campaign Assist, and Mortgage Choice broker tools shipped in the year (Kalkine, 2026). Domain remains unverified.

#	JUNE ITEM (ABBREVIATED)	JUNE STATUS	SEPTEMBER 2026 DISPOSITION
28	OpenAI ~US$12B quarterly loss; ~US$143B cumulative burn	Inferred from Microsoft filings and Deutsche Bank	**Complicated.** Microsoft's FY2026 release records net gains on OpenAI investments of US$4,963 million for the year and US$480 million for the June quarter (Microsoft, 2026). The gains reflect the accounting treatment of Microsoft's stake after OpenAI's recapitalisation and say nothing reliable about OpenAI's operating losses, which remain unaudited; the June framing ("losses inferred from Microsoft filings") should not be repeated without that caveat.
29	NAIC SMB adoption figures (40 to 69 per cent regular use)	Confirmed against the Dec 2025 to Feb 2026 summary	**Weakened.** The NAIC's published dataset for that same fieldwork (7 May 2026) reports 43 to 44 per cent with "some level of AI adoption"; one dataset is reported two incompatible ways and the instrument has not been examined. Neither figure should be quoted as a time series until it is.
30	Deloitte "15 per cent with significant ROI"	Publisher-reported; sample confirmed	**Unchanged.** No 2026 Deloitte edition located; McKinsey's 6 per cent "high performers" (25 August 2026) is directionally consistent but does not verify the Deloitte figure.

Count. Two now verified (2, 27); one complicated (28); two weakened (9, 29); two closed as no longer relevant (3, 13); twenty-three unchanged. No item moved from unverifiable to refuted in the window.

Part 2: New items opened in this edition

#	CLAIM	WHY IT COULD NOT BE VERIFIED	WHAT WOULD RESOLVE IT
N1	Operational status of the Australian AI Safety Institute (staffing, first outputs)	The industry.gov.au pages could not be fetched (robots); search results predated the window.	A direct check of industry.gov.au and any Senate Estimates testimony.
N2	Google's Project Mariner discontinued around 4 to 7 May 2026	Press reports across three outlets agree; no primary Google announcement located; the DeepMind Mariner page no longer mentions the product.	A dated Google or DeepMind post. The correction to June §3.2 is carried on the press reports.
N3	Publication date of the Australian Government's whole-of-government Microsoft 365 Copilot trial evaluation	The DTA pages fetched carry no date; the trial and evaluation are believed to predate the June edition.	A dated DTA release. Until then the evaluation is cited as undated context, not window evidence.
N4	Two systematic reviews on external validation of diagnostic AI (a melanoma meta-analysis on PubMed; a radiology generalisability review at DOI 10.1097/MS9.0000000000004166)	CAPTCHA and paywall on fetch; dates and findings unconfirmed.	Institutional access.

#	CLAIM	WHY IT COULD NOT BE VERIFIED	WHAT WOULD RESOLVE IT
N5	Google's retirement of Veo 2.0, Veo 3.0 (30 June 2026) and Imagen 4 (17 August 2026)	Opened on an aggregator tracker; **closed during the fact-check** against Google's Gemini API changelog (Google, 2026b). Retained for the record.	Resolved.
N6	SpaceX agreement to acquire Cursor (Anysphere) for US$60 billion, 16 June 2026	Single-source (Sacra); not corroborated before the search budget ran out.	A second independent report or a company statement. Not used in the body.
N7	Meta Muse Spark 1.1 API pricing (reported US$1.25 input / US$4.25 output per million tokens)	Reported in search snippets from major outlets; no primary page fetched.	Meta's pricing page. The body states only that the model is Meta's first paid API model.
N8	The TGA's own statement listing software as a medical device among twelve compliance and enforcement priorities for 2026-27	Carried on a law-firm summary (Clayton Utz, 2026); the TGA priorities page was not fetched.	A fetch of the TGA's published compliance priorities. The §3.10 classification would survive on the FWC rules alone.

Appendix C: Corrections log (v0.9 to v1.0 fact-check pass)

The pre-fact-check draft is preserved at `state-of-ai-september-2026-literature-review-v0.9-draft.md`. An independent fact-check (a fresh Claude Opus pass that had not seen the drafting, working from the draft, the previous edition's text, and direct fetches of every cited URL; web search was unavailable to it) produced 118 findings: 61 verified, 26 partially verified, 18 pending, 13 refuted. Every refuted and partially verified finding was acted on as follows. The findings log is preserved beside this file as `state-of-ai-september-2026-factcheck-findings.md`.

Refuted (13), all corrected:

1. Three quotations did not match their sources. The OAIC facial-recognition guidance does not contain "a precautionary approach to the deployment of FRT is required under Australian law"; replaced with the guide's actual wording on a "formal, structured and documented risk assessment", and the guide correctly described as first published November 2024 and updated 29 July 2026. The Ninth Circuit opinion says "Perplexity's servers never directly access Amazon's servers", not "does not directly communicate with Amazon's servers"; corrected, and the "tool, not a person" quotation extended to its full sentence. The TGA position is "identified as one of 12 priority focus areas for the TGA's compliance and enforcement activities in 2026 and 2027", not "a priority enforcement focus area for 2026-2027"; corrected everywhere, including the §3.10 movement reason, the register, the

abstract, §5, and the §3.10 quotable sentence, and the TGA primary source opened as Appendix B item N8.

2. MCP "final release". MCP's versioning vocabulary reserves "Final" for past, unchanging revisions; 2026-07-28 is the current version. Corrected in the register, §2.4 scorecard, abstract, §3.2 (four places), §4, §5, §6, and the §3.2 quotable; the ninety-day expedited-removal exception to the twelve-month deprecation rule added; the AAIF May post re-labelled as the schedule announcement and the specification's own versioning page added as a source.
3. Two Australian regulator figures given in US dollars. ACCC's A$248.3 million (§3.5) and ASIC's A$7.4 million (§3.5, §3.9) corrected to Australian dollars.
4. x402 "93 per cent from its late-2025 peak". The 93 per cent is year to date; the fall from the peak is about 97 per cent, and the peak "repeatedly approached US$800,000 and occasionally exceeded US$1 million". Corrected in the abstract, register, §3.9 body, movement table, judgement, quotable, and §5.
5. Capex arithmetic. The draft's own figures (about US$671 to 686 billion) sit inside June's US$630 to 725 billion range, not "at or above its top end"; §4 rewritten with the sum, the fiscal-versus-calendar caveat, and Meta's guidance correctly described as narrowed upward rather than raised (abstract also corrected).
6. "Nine of forty cells moved". Four improved plus five declined plus two reframed is eleven; "nine" was the directional count. Corrected in the register summary, abstract, §4, and §6 to "eleven moved, nine directionally".
7. Seedance 2.5 "the longest single-pass duration recorded in this series". Runway Gen-4.5 (60 seconds) is longer and is in the same section; corrected.
8. Watchlist mischaracterised June's Appendix B. The draft implied all thirty items were open; June's Pass 2 had already promoted twelve, refuted three, and retained fourteen with a resolved status. The front-matter summary and Appendix B rewritten to record June's own statuses, items 3 and 13 re-described as closed by June and now moot rather than newly irrelevant, and "no item was refuted in the window" qualified with item 10's prior refutation.

Partially verified (26), corrected where the error was specific: Visa's European go-live re-dated to 2 July (reported 6 July) in six places; Cowork mobile and web re-dated to 8 July and its telemetry re-labelled as late-May fieldwork; Atlas "less than ten months"; Xero post re-dated to 19 August, "since launch" not "since beta", XeroForce's promised general availability noted; the Hollywood cease-and-desist letters correctly attributed to the MPA and five studios over Seedance 2.0 (February); HappyHorse's ranking correctly attributed to Arena.ai and versioned 1.1; Stanford's prior-year comparator corrected to 15 per cent (with June's 13 noted) in the abstract and §4; Microsoft's figure harmonised to US$4.96 billion (US$4,963 million); the Scale AI leaderboard given with its confidence intervals, its public-set size, and the fifth entry (Muse Spark, 55.0); "commercial subset" replaced with "non-public subset"; the §3.1 quotable's "65" restored to 64.6; DeepSeek's specifications and pricing, Kimi K3's parameter count, Opus 5's positioning, Visa's issuer count, and NEAR's cumulative volume labelled as vendor-originated; the

36-times price comparison given its comparator (GPT-5.6 Sol); the Epic sepsis "never validated in a randomised trial" clause removed; the EU political-agreement date noted as 7 May on the Commission's page against June's 6 and 13 May; the Commission press release re-dated 31 July; the AFM suit re-dated to "early June (reported 8 June)"; the Charlotin fetch date (2 September) separated from the most recent case date (31 August) and the two Australian cases named; June's "using or planning to use" restored in the financial-advice sentence; June's "20-plus per cent faster" restored in §3.7 and its "10 to 15 per cent floor" in §3.6; the §5 Wait posture restored to June's wording with the one boundary move (limit-bound agent payments to pilot) stated explicitly; the "base case" sentence marked for its two additions; the ASIC-named commentators added; the Meta pricing removed from the body and opened as N7; the register's §3.4 underestimated cell downgraded from "confirmed" to "no new evidence" (one vendor price point is not a trend), changing the split to 14 confirmed and 15 no new evidence.

Pending (18), disposition: Alphabet and Amazon capex figures retained with their sourcing caveats stated in the text; Bloomberg columns retained as headlines only; Fable 5.1's 1 September date retained (the vendor page says September 2026; the date is that of the announcement as observed); GLM-5.3, Mistral early access, Nextgov, Skadden, Cooley, Morgan Lewis, DTA, Stack Overflow, Claude memory merge, and Sora-related items retained as cited with the source class stated; the Muse Spark pricing removed (N7); the Mariner correction retained on press reports (N2); the OmniAI benchmark page 404 recorded (§3.3); the "extending to split payments" clause removed from the Xero sentence; the MCP registry count range retained as an illustration of the absence of a census.

Additions the fact-check supported: Google's own changelog confirms the Veo and Imagen shutdown dates (Appendix B N5 closed; §3.4 re-sourced); the TrustNLP finding that all eight agentic RAG modes are unstudied added to §3.6 and its quotable; McKinsey's 14 per cent actual-workforce-reduction figure added to §4 beside the 39 per cent expectation; METR's per-model horizons and spend added to §2.3.

Not changed, with reasons: §2.3 and §3.2 both narrate METR, Atlas, and ChatGPT Work; the duplication is deliberate so that each section stands alone. The §3.2 and §3.6 "deploy today" cells remain "confirmed" on in-window evidence (ChatGPT Work, 9 July; Claude's memory merge, 25 August) after the pre-window items (Cowork telemetry; NotebookLM) were re-labelled as supporting context only.

Appendix D: Series continuity

The fixed frame. The series measures movement against a frame that is held still between editions: ten capability sections (§3.1 to §3.10) with the names and numbering of the Mid-2026 edition; four judgements per section (deploy today; demo-versus-production gap; oversold; underestimated); four adoption postures (deploy now with verification built in; pilot deliberately with falsifiable metrics; wait; walk away); and two standing hypotheses (the benchmark-to-

production gap and the individual-to-firm gap). A section, judgement, or posture is renamed, split, merged, or reordered only between editions, with the rationale recorded in the metadata block's "structural change" row and a mapping from old cells to new. This edition made no structural change.

Classification definitions. Improved and declined are defined per column so that improved is always in the reader's favour and declined always warrants more caution: in the deploy-today column, improved means more is dependable and declined means something dependable was withdrawn or shown unreliable; in the gap column, improved means independent evidence narrowed the gap and declined means it widened; in the oversold column, improved means the oversold claim gained independent support and declined means further evidence that it is oversold; in the underestimated column, improved means further evidence the capability is further along than assumed and declined means it is less far along. Unchanged (confirmed) requires in-window evidence of the required standard; unchanged (no new evidence) requires the queries run to be listed in Appendix A; reframed means the judgement stands but the reason or boundary changed. A classification is a claim and carries a citation.

Carry-forward rule. Where no in-window evidence of the required standard exists, the previous edition's judgement is restated in full with its research detail, cited to the previous edition and, where it bears on a decision, to the primary source the previous edition used. Carried-forward text is never presented as new. A vendor claim carried forward remains a vendor claim.

Scorecard rule. Each edition's §2.4 is written to be scored, item by item, by the next edition, with the four outcomes shipped as announced, shipped differently, did not ship, and still open. Each edition also names the section it expects to age fastest and the next edition says whether it did.

Watchlist rule. Appendix B carries every unverified item forward with its disposition until it is verified, refuted, or retired as no longer relevant, and opens new items each edition. Numbering from the Mid-2026 edition is preserved; new items are prefixed N and numbered per edition.

Silence rule. A third consecutive "unchanged (no new evidence)" in any cell is to be treated by the next edition as a finding about the category's evidence base, stated in the section, not as a quiet quarter.

Site mapping. The forty register cells map one-to-one onto the verdict table on the series hub (section ids `s3-1` to `s3-10`; cell keys `deploy`, `gap`, `oversold`, `under`), and the classification drives the cell's trend colour. The hand-off note beside this file carries the mapping for this edition.

Appendix E: Iteration history

VERSION	DATE	STATUS	CHANGES
v0.9	2 September 2026	Archived	Seven-stream research draft. Archived at `state-of-ai-september-2026-literature-review-v0.9-draft.md`.
Fact-check pass	2 September 2026	Complete	Independent verification of every citation, in-window numerical claim, carried-forward baseline claim, the movement register's forty cells and count, internal consistency, vendor-claim labelling, and style. 118 findings: 61 verified, 26 partially verified, 18 pending, 13 refuted. Log preserved at `state-of-ai-september-2026-factcheck-findings.md`.
v1.0	2 September 2026	**Published (fact-checked)**	All 13 refuted findings corrected; 26 partial findings corrected or labelled; 18 pending items retained with their source class stated or moved to Appendix B; register split revised to 14 confirmed and 15 no new evidence; Appendix B N5 closed on a primary source; four evidence additions the fact-check supported.

Provenance. This paper is published by Perth AI Consulting as the second edition of its quarterly evidence review and the first in the series to be built as an update against a fixed frame. It was researched and drafted with AI assistance (Claude Fable 5.1, coordinating seven research streams), independently fact-checked by a separate model pass that had not seen the drafting and that traced every cited source, corrected against that log, and reviewed by a human before publication. The same standard applies to the paper as the paper applies to the field: AI-produced content is a draft requiring verification before it becomes load-bearing, and the corrections log above is the record of what that verification found.