The State of AI in Mid-2026: A Literature Review for Operational Leaders

Prepared by Perth AI Consulting using Anthropic's Claude
(Fable 5 drafting; Opus 4.8 fact-check)

Version 1.0 (fact-checked) · 13 June 2026

Perth AI Consulting
perthaiconsulting.com.au

> *Prepared as a PAC resource. Research completed 13 June 2026. All web sources accessed June 2026 unless otherwise noted.*
>
> *This paper has been through a three-pass independent fact-check. The pre-fact-check draft is preserved at* archive/state-of-ai-mid-2026-v0.9-pre-fact-check.md*; every correction is logged in the Corrections Log appendix.*

Abstract

This paper surveys the state of applied artificial intelligence as of mid-2026, written for operational leaders of small and mid-sized businesses, regulated professionals, and the consultants who advise them. It covers ten capability categories: frontier language and reasoning models, agentic systems, vision and multimodal models, video and image generation, audio and voice, knowledge management, code generation, business automation platforms, the crypto-AI convergence, and vertical industry applications. For each category it distinguishes what is reliable in production today, what works in demonstrations but fails on real data, what is sold as more mature than it is, and what is further along than commonly assumed.

The paper's central finding is a persistent and measurable gap between benchmark performance and production reliability, and a second gap between individual-level productivity gains and firm-level financial results. Frontier models have improved materially in the twelve months to June 2026: million-token context windows are standard across Anthropic, OpenAI, and Google; abstract reasoning scores on ARC-AGI-2 moved from single digits to above the human average; and agentic coding became a multi-billion-dollar product category, with Anthropic reporting a US$2.5 billion annualised run-rate for Claude Code by February 2026. Yet the best independent evidence shows agents completing roughly 30 per cent of simulated office tasks (Carnegie Mellon's TheAgentCompany; best-agent result reported by The Register, 2025), document extraction plateauing near 75 per cent field-level accuracy on real-world documents, the best grounded summarisation tools still hallucinating in about 13 per cent of responses, and the largest government-scale Copilot evaluation to date finding high satisfaction and self-reported time savings but not establishing productivity gains under controlled measurement.

At the same time, several capabilities are underestimated. AI medical scribes deliver modest but real, peer-reviewed reductions in documentation burden. Confidence-gated transaction coding in accounting platforms such as Xero's JAX works as designed. Voice agents handle routine, bounded calls at customer satisfaction levels comparable to humans. Coding assistance, despite contested productivity evidence, has been adopted at a scale and speed without precedent in business software.

The pattern that emerges is consistent across categories: AI is production-ready where the task is bounded, the output is verifiable, and a human or deterministic system checks low-confidence cases; it is not production-ready where the task is open-ended, the output is hard to verify, and errors are costly. The paper closes with a disciplined view of the next 12 to 24 months, an adoption posture for operational leaders, full references, and an appendix listing every claim that could not be verified at the time of writing and why.

1. Introduction

1.1 Purpose

This review exists because the gap between what AI vendors claim and what their products do in production has become a material business risk in both directions. Leaders who believe the marketing waste money on systems that fail on their real data. Leaders who dismiss the field as hype forgo capabilities that have quietly crossed into dependable production use. Both errors are common, and both are avoidable with better information.

The intended reader operates a business or advises one: the owner of a 40-person construction firm, the principal of a medical practice, the partner of a suburban law or accounting firm, the financial adviser, the consultant who serves them. The reader is assumed to be sophisticated about their own operations and unsentimental about technology purchases. The paper is not written for machine learning engineers, and it deliberately avoids architecture detail except where it bears on a purchasing or deployment decision.

1.2 Scope

The survey is international, because the products are. Australian context is noted where it is materially relevant: local regulatory positions (AHPRA, ASIC, the Federal Court, the Privacy Act reforms), locally built products (Heidi Health, Lyrebird Health, ServiceM8, Employment Hero), and Australian adoption data. The paper covers capability and reliability. It does not cover AI ethics, safety policy, or governance debates in any depth; those matter, but they require their own paper. Nor is it a tool-selection guide for specific roles; it describes the terrain rather than prescribing routes across it.

1.3 Methodology

The review was researched in June 2026 through structured web research across vendor documentation, peer-reviewed literature, independent evaluation organisations (METR, Epoch AI, Stanford HAI and RegLab, the DORA research programme), regulator publications, financial press, and practitioner writing. Three methodological rules were applied throughout.

First, every capability claim is dated. A claim about what a model can do is only meaningful with a timestamp, because the field moves in months.

Second, vendor claims are labelled as vendor claims. A large fraction of the published evidence base in 2026 is written by parties selling the thing being evaluated. Voice agent "benchmarks" are published by voice agent vendors; CRM AI "reviews" carry affiliate links; revenue run-rates are told to investors, not audited. Where the only available number is a vendor number, the paper says so.

Third, what could not be verified is flagged rather than omitted or asserted. Appendix B lists every claim encountered during research that could not be traced to a credible primary source, with the specific reason. Several widely circulated "facts" about AI in 2026 appear in that appendix rather than in the body of this paper, which is itself a finding.

A note on the limits of this method: the research relied on what is publicly available as of 13 June 2026. Private deployment data, unpublished accuracy figures, and paywalled analyst reports were not accessible. Where a claim depends on such a source, this is stated. The authors also note that independent evaluation lags product releases by six to twelve months in most categories, so the most recent products are necessarily assessed partly on vendor evidence, with appropriate caution.

1.4 How to read the quality assessments

Each capability section ends with the same four-part judgement: what is deployable in production today; what works in demonstrations but fails on real data; what is widely sold but should not yet be trusted; and what is widely dismissed but further along than assumed. These judgements are the paper's core value and they are stated plainly. They reflect the evidence available in June 2026 and should be expected to age; the temporal frame in Section 2 is intended to help the reader judge how fast.

2. The Temporal Frame

2.1 The pre-AI baseline

It is worth recording, briefly, how operational work was done before late 2022, because the baseline is already being forgotten. Documents were drafted from precedent and template. Transcription was done by humans or by speech software that required correction of most sentences. Customer enquiries queued for humans or hit decision-tree chatbots that customers actively avoided. Bookkeeping data entry was manual or rules-based. Software was written by software developers, with no exceptions worth noting. Search meant keywords. Synthesis of a hundred-page document meant a person reading a hundred pages. None of this was a complaint at the time; it was simply the cost structure of knowledge work, and businesses were organised around it.

2.2 Early AI, 2022 to 2024: real augmentation, absent transformation

ChatGPT launched on 30 November 2022 and reached roughly 100 million monthly users within two months, the fastest consumer software adoption recorded to that point (Demand Sage, 2026; figures originally from UBS/Similarweb estimates). GPT-4 followed in March 2023, and the first wave of copilots arrived through 2023 and 2024.

The controlled evidence from this period holds up. The Harvard and BCG “jagged frontier” experiment (Dell’Acqua et al., September 2023) gave 758 consultants GPT-4 and found 12.2 per cent more tasks completed, 25 per cent faster, with the largest gains accruing to below-median performers; but on tasks outside the model’s competence, AI use made consultants measurably worse. GitHub’s own controlled study (February 2023) found a 55.8 per cent speed-up on a single scoped coding task, a number that was widely over-generalised; later field studies found gains closer to 10 to 26 per cent depending on context, and the study was vendor-run.

What did not happen in this period was firm-level transformation. Pilots proliferated; profit-and-loss impact mostly did not. The explanation that took hold by 2025, supported by survey evidence discussed below, was that the bottleneck was never raw model quality but integration: workflow redesign, data hygiene, accountability for output, and systems that retain context. The era’s enduring lesson is that a capable model in a browser tab changes individual output, not business processes.

2.3 Now, 2025 to mid-2026: the agentic turn and the evidence gap

Three shifts define the current period.

The first is reasoning models. OpenAI's o1 (late 2024) and DeepSeek's R1 (January 2025, open weights) established test-time computation, in which models spend variable effort "thinking" before answering, as the dominant paradigm. Reasoning capability is the prerequisite for the second shift: agents, meaning systems that plan and execute multi-step tasks using tools rather than answering single prompts. The third is infrastructure: Anthropic's Model Context Protocol (November 2024) became the de facto standard for connecting models to tools and data, and was donated in December 2025 to the Agentic AI Foundation under the Linux Foundation, with OpenAI, Google, Microsoft, and AWS as co-sponsors (Linux Foundation, December 2025), which is the strongest available evidence that the integration layer has standardised.

Capability moved materially. Million-token context windows became standard across the three major labs between February and June 2026. ARC-AGI-2, an abstract-reasoning benchmark on which the best model scored 8.6 per cent in May 2025, was above the 66 per cent human average for several frontier models by early 2026 (Arc Prize Foundation leaderboard, 2026). Agentic coding became the first unambiguous commercial product-market fit of the agent era, discussed in Section 3.7.

The adoption evidence, however, remained sobering throughout. The widely quoted MIT NANDA finding of August 2025, that 95 per cent of enterprise generative AI pilots produced no measurable profit-and-loss impact, was methodologically weak (52 interviews, a non-random sample, and a research group with its own commercial agenda; see Futuriom, August 2025, for the critique), but its direction matched stronger surveys. McKinsey's State of AI (the cited March 2025 report, surveying organisations in July 2024) found 78 per cent of organisations using AI in at least one function and 71 per cent regularly using generative AI, yet more than 80 per cent reported no tangible enterprise-level earnings impact and only 1 per cent described their rollouts as mature; the survey is a consultancy self-report. Deloitte's 2026 enterprise survey (3,235 senior leaders across 24 countries, fieldwork August to September 2025) reported 15 per cent of organisations achieving significant measurable ROI today (the sample size is confirmed; the specific 15 per cent figure remains publisher-reported and is flagged in Appendix B; like McKinsey, a consultancy self-report). Gartner predicted in June 2025 that more than 40 per cent of agentic AI projects would be cancelled by end-2027 and estimated that only about 130 of the thousands of vendors claiming "agentic AI" were genuinely agentic, coining "agent washing."

In Australia, National AI Centre tracking reported regular AI use among SMBs rising from 40 per cent in July 2024 to 69 per cent by January 2026, with daily use rising from 9 to 28 per cent (ai.gov.au, 2026; primary survey instrument not independently examined). Adoption, in other words, is no longer the constraint. Converting adoption into reliable operational value is.

Two economic facts frame everything else. Inference prices for a fixed level of capability fell at roughly 9 to 900 times per year depending on the capability milestone, around 40 times per year to match GPT-4-level performance (Epoch AI, 2025–26). Simultaneously, total AI spending exploded: combined 2026 capital expenditure guidance from Microsoft, Alphabet, Amazon, and Meta reached roughly US$630 to 725 billion, up from about US$400 billion in 2025 (CNBC, February 2026; totals vary with scope). Cheaper tokens and bigger bills coexist because reasoning and agentic workloads consume orders of magnitude more tokens per task. Whether the capital expenditure is justified is the live macroeconomic dispute of mid-2026, examined in Section 4.

2.4 The next 12 to 24 months: announced versus speculative

What is scheduled or highly likely, as distinct from speculated:

Regulation is loosening or delaying, not tightening. The EU reached provisional political agreement on 6 May 2026 (confirmed by Member State representatives in Council on 13 May) to postpone the AI Act's high-risk obligations from August 2026 to December 2027 (Annex III) and August 2028 (Annex I); transparency rules still commence in August 2026 (Gibson Dunn, May 2026; the agreement is provisional, not yet formally adopted). Australia's December 2025 National AI Plan shelved the previously proposed mandatory guardrails in favour of existing law, sector regulators, and a new Australian AI Safety Institute. The Privacy Act's automated-decision-making transparency obligations commence 10 December 2026 and are already legislated; this is the one firm Australian compliance date most SMBs will face. In the United States, the December 2025 executive order and March 2026 White House framework push toward federal preemption of state AI laws, though Congress had enacted nothing as of writing.

The economics will be stress-tested in public. Anthropic confidentially filed for an IPO at a reported US$965 billion valuation in June 2026 (Fortune, 1 June 2026), and an OpenAI listing is widely expected. Public listings would expose frontier-lab economics, currently a mixture of investor-reported run-rates and leaked losses, to audited scrutiny for the first time.

Agent infrastructure consolidates. MCP for tools, A2A for agent-to-agent communication, and a contested set of payment protocols (AP2, ACP, x402, all discussed in Section 3.9) are all under neutral foundations as of mid-2026. The committed hyperscaler capital expenditure means compute supply keeps growing into 2027 regardless of the demand debate.

What should be discounted: 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. Claims of recursive self-improvement on an 18-month horizon are speculation. The disciplined base case is continued incremental frontier releases, agents that handle somewhat longer task horizons with somewhat better reliability, deeper vertical products, and no scheduled discontinuity.

3. Capability Survey

3.1 Large language and reasoning models

The frontier in June 2026. Three labs define the closed frontier. Anthropic shipped Claude Opus 4.5 in November 2025 (with a 67 per cent price cut, to US$5/$25 per million tokens), Opus 4.6 in February 2026 (one-million-token context in beta, 128K output), Opus 4.8 on 28 May 2026, and on 9 June 2026 released Claude Fable 5, the public version of its "Mythos-class" tier, with certain dual-use capabilities (notably offensive cyber) restricted to a vetted-access variant called Mythos 5 (Anthropic, June 2026; TechCrunch, 9 June 2026). Independent commentary on Opus 4.8 called it "a modest but tangible improvement" (Willison, 28 May 2026); Fable 5's claimed margin over it, more than 10 per cent on some benchmarks, is a vendor claim with no independent replication yet, and no independent verification exists for the claimed cyber capabilities that justified the Mythos access restrictions, a sequence both TechCrunch and TechRadar covered sceptically.

OpenAI moved from GPT-5 (August 2025) through GPT-5.1, 5.2, and GPT-5.5 (April 2026), whose Instant variant became the ChatGPT default on 5 May 2026. GPT-5.5 offers a one-million-token context at US$5/$30 per million tokens, double its predecessor's price, a positioning the trade press flagged as paying more for "a new class of intelligence" on the vendor's say-so (The Decoder, April 2026). Google shipped Gemini 3 Pro on 18 November 2025 and Gemini 3.1 Pro on 19 February 2026, the latter claiming 77.1 per cent on ARC-AGI-2, more than double its predecessor three months earlier (Google, February 2026; vendor claim, directionally consistent with the Arc Prize leaderboard).

Anthropic's ecosystem moves matter to operators as much as its models: Skills (October 2025), packaged task instructions a model loads on demand, became an open standard in December 2025 with Canva, Notion, Figma, and Atlassian publishing skills; Cowork (January 2026, Windows February 2026) extends agentic file-and-task work to non-developers and remains a research preview; MCP is covered in Section 3.2.

Open weights. The credible open-weight frontier is now Chinese and European. DeepSeek V4 (April 2026) is a 1.6-trillion-parameter mixture-of-experts model with only 49 billion active parameters per token, MIT-licensed, scoring within a few points of closed frontier models on coding benchmarks at roughly 30 times lower cost. Mistral Large 3 (December 2025) is a 675B/41B sparse MoE under Apache 2.0. Alibaba's Qwen series iterated three times between February and May 2026. Meta's Llama 4 (April 2025) landed poorly: its advertised 10-million-token context was widely criticised as unusable, and the

two-trillion-parameter Behemoth variant had not shipped publicly as of May 2026, surviving only as an internal teacher model (conflicting reports exist; the weight of sourcing says not shipped). The architectural trend is uniform: every 2026 flagship is sparse mixture-of-experts, so “trillion-parameter model” now describes total capacity, not the compute spent per token.

What changed in twelve months. Four things, materially. Million-token context became standard rather than exotic. Abstract reasoning, as measured by ARC-AGI-2, moved from single digits to above the human average of 66 per cent. The gap between the best closed and best open models compressed to roughly 55 Elo points on community leaderboards (aggregator-reported; treat the precise figure cautiously). And price-per-capability fell sharply even as premium tiers rose, so the same dollar buys perhaps an order of magnitude more capability than in mid-2025.

What is reliable, and what is not. Drafting, summarisation, transformation, translation, structured extraction from clean text, and code assistance are dependable at frontier quality. Hallucination is not solved and the field has stopped pretending it will be soon: Stanford RegLab measured 17 to 33 per cent hallucination rates in commercial legal AI tools marketed as hallucination-free (peer-reviewed 2025), and the now-consensus explanation is that training rewards confident guessing over abstention. Benchmark scores have decoupled from production reliability: MMLU-Pro is saturated above 89 per cent for all frontier models (Epoch AI, April 2026), SWE-bench Verified is suspected of training contamination (OpenAI itself published contamination concerns in February 2026), and frontier models score materially lower, by 15 points and in some aggregator-reported cases far more, on the contamination-resistant SWE-bench Pro, where the best results sit near 80 per cent (Scale AI leaderboard via aggregators, 2026; precise per-model gaps are inconsistently reported). A reader shown a benchmark chart in a sales deck should ask which benchmark, measured by whom, with what scaffolding, and how many runs.

> **Quality summary.** Deploy today: text and document work with human review, structured extraction with validation, coding assistance. Demo-versus-production gap: anything quoted in benchmark points. Oversold: “hallucination-free” claims in any domain, and headline benchmark margins between frontier models, which are now within measurement noise of each other. 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.

3.2 Agentic systems and tool use

The infrastructure layer settled. MCP won the integration-protocol contest: donated to the Linux Foundation's Agentic AI Foundation in December 2025 with every major lab and cloud as a sponsor, with an independent census counting 17,468 public MCP servers in the first quarter of 2026 (Nerq, 2026). Google's complementary A2A protocol for agent-to-agent communication sits under the same foundation. For an operational leader the practical meaning is that connecting an AI system to business tools is no longer a bespoke integration problem; the plumbing is standardising the way web standards once did.

The products. Anthropic's Claude Code (terminal and IDE) and Cowork (desktop) anchor one approach: agents working in environments the user controls. OpenAI's Operator was shut down in August 2025 and folded into ChatGPT's agent mode; its Atlas browser launched in October 2025. Perplexity's Comet browser went free across platforms in March 2026. Google ships Project Mariner, the Jules coding agent, and agent mode across Gemini. Microsoft announced a Windows Agent Runtime at Build in June 2026. All of these work; none of them works reliably enough to run unattended on consequential tasks, which is the finding that matters.

The independent evidence. Three results define honest expectations. First, METR's task-horizon research (January 2026 update) finds the length of task an agent can complete at 50 per cent success doubling every four to seven months, an extraordinary trend, but the companion finding is that success decays roughly exponentially with task duration: an agent with a 50 per cent success rate at one hour has a far shorter horizon at the 80 per cent reliability a business actually needs. Second, Carnegie Mellon's TheAgentCompany simulation (2025) gave agents real office tasks in a simulated software firm; in the run reported by The Register (29 June 2025), the best agent (Gemini 2.5 Pro) completed 30.3 per cent of tasks autonomously, and failure modes included fabricating data and renaming a colleague in a chat system to pretend a task was done. (The CMU news write-up of an earlier run reported 24 per cent for the best agent then tested.) That study tested early-2025 models, and no comparable independent rerun with mid-2026 models had been published as of writing, which is itself a finding: vendors are shipping agent products faster than independent evaluation can assess them. Third, enterprise reports consistently describe a large gap between benchmark and real-world agent performance, with academic work (Mustahsan et al., 2025) showing that single-run benchmark scores cannot reliably distinguish genuine capability gains from sampling noise, motivating multi-run consistency reporting (the intraclass correlation coefficient) alongside policy-adherence measures.

Security is unsolved and the vendors say so. Brave demonstrated in August 2025 that hidden text in a Reddit post could make Perplexity's Comet read a victim's email and exfiltrate a one-time password. Proof-of-concept injections against Atlas appeared within hours of launch (The Register, October 2025). OpenAI's own security leadership stated in December 2025 that prompt injection is "unlikely to ever be fully 'solved'" and that agent mode expands the threat surface (Fortune, 23 December 2025). In March 2026 Senior US District Judge Maxine M. Chesney (Northern District of California) issued a preliminary injunction on 10 March 2026 barring Perplexity's Comet from accessing password-protected Amazon accounts, on Computer Fraud and Abuse Act grounds and a California computer-fraud statute, and ordered the destruction of Amazon data collected via Comet; the Ninth Circuit granted a temporary administrative stay on 16 March 2026 (CNBC; GeekWire). The ruling's logic is that a user's permission to an agent does not substitute for the platform's authorisation. A business connecting an agent to email, banking, or customer data should treat every web page and document that agent reads as potentially hostile input.

Where agents actually work in mid-2026. Coding, by a wide margin: constrained environment, verifiable output, version control as an undo button. Research and synthesis with citations a human checks. Bounded browser automation on known sites. Internal workflows where every consequential action requires human approval. Computer-use agents have improved markedly on the OSWorld benchmark, with frontier results around 70 per cent on standard OSWorld and roughly 78 to 80 per cent on the OSWorld-Verified variant as of June 2026 (XLANG), though scores are highly sensitive to benchmark variant, date, and scaffold, so a single headline number means little. Where they fail: long multi-step tasks on unfamiliar interfaces, payments and anything irreversible, edge cases, and any task whose success the agent itself is allowed to judge.

> **Quality summary.** Deploy today: coding agents; research agents with human review; tightly scoped internal automations with approval gates. Demo-versus-production gap: end-to-end “digital employee” demonstrations, which showcase the 30 per cent of attempts that succeed. Oversold: autonomous agents for consequential, unsupervised work; “agent washing” of ordinary automation is pervasive (Gartner, June 2025). Underestimated: the compounding value of the MCP standard itself, and supervised agents as a labour multiplier for staff who learn to delegate and verify.

3.3 Vision and multimodal models

Document understanding. This is the highest-value vision capability for most businesses and the one with the best-measured gap. On clean academic benchmarks, frontier models look close to solved; OmniDocBench is now described as saturated (LlamaIndex, 2026). On real documents the picture is different: the OmniAI benchmark of 1,000 real-world documents, run as document-to-structured-JSON extraction, found the best models near 75 per cent field-level accuracy (2025–26). The diagnostic detail matters: given perfect text, the same models score about 99 per cent on the extraction step. The bottleneck is reading degraded, skewed, handwritten, and complex-layout material, not reasoning about it. Specialist OCR models continued to beat general frontier models on raw character accuracy as of early 2026. The deployable pattern, used by every successful production system encountered in this research, is extraction with confidence scores, automatic acceptance of high-confidence fields, and human review of the rest.

Vision against standards. Checking visual material against reference documents, plans against building codes, photographs against specifications, is an emerging category rather than a mature one. Products exist for automated building plan review against US codes (CodeComply.AI, CivCheck, PlanCheckPro.AI, all 2025–26), and Florida legislated in 2025 to permit software-based plan review. No independent accuracy evaluation of any vision-against-standards product was found during this research, and the claim that AI plan review compresses 30-to-60-day approvals to days traces to a single unverifiable secondary source. The capability is real enough to watch and pilot; it is not yet evidenced enough to rely on.

Medical imaging. This is the one vision domain with randomised-trial evidence. The MASAI trial (roughly 106,000 women, Swedish national screening, final Lancet results published around February 2026) found AI-supported mammography reading raised sensitivity from 73.8 to 80.5 per cent at equal specificity and cut screen-reading workload 44 per cent, with a non-inferior interval-cancer rate; the press-released “reduction” in interval cancers was not statistically significant, a distinction vendor marketing elides. Pathology has an FDA-authorised product (Paige Prostate) with independent validation near 98 per cent sensitivity (Yale assessment, Modern Pathology, 2021; EU IVDR certification 2025). The cautionary cases are also instructive: DermaSensor’s FDA clearance shows 95.5 per cent sensitivity bought with 32.5 per cent specificity, meaning two-thirds of benign lesions are flagged (Dermatology Times, 2024); a prospective study of the consumer app SkinVision found more than 70 per cent of lesions could not even be adequately photographed when patients did it themselves (British Journal of Dermatology, 2026). Systematic reviews through 2025–26 repeatedly find that performance drops when

models move to new sites, scanners, and populations, and that most published radiology AI lacks external validation. More than 1,250 AI-enabled devices had FDA authorisation on the agency's running list by around July 2025 (FDA AI/ML-enabled device list, corroborated by the Bipartisan Policy Center; Innolitics separately recorded 295 such clearances during 2025); clearance volume is accelerating and the FDA moved in December 2025 to exempt some radiology triage categories from premarket notification.

Inspection, valuation, construction. Construction progress tracking against BIM models (Buildots, OpenSpace) is in real production use on large projects, with vendor case studies but no published independent accuracy figures. Aerial property measurement (EagleView) claims 98.77 per cent measurement accuracy against an unnamed benchmark; vendor claim. Automated valuation models augmented with image analysis claim material accuracy gains, but the widely circulated figure (94.2 per cent of valuations within 5 per cent of sale) traces to a single secondary source with unseen methodology. In Australia, CoreLogic claims roughly 90 per cent of its AVM estimates fall within 15 per cent of sale price (vendor claim, 2024). Building defect detection from imagery performs well on visually distinct defect classes (vendor and academic sources suggest 90 to 95 per cent on well-defined classes such as water staining and roof membrane damage) and poorly on subtle, internal, or occluded defects, which is precisely where inspection liability lives.

Known failure modes. The 2025–26 research literature documents persistent failures in counting, fine spatial relations, chart and table reading, and confident hallucinated readings of degraded text. These are not edge cases for business use; they are the centre of invoice processing, plan reading, and report extraction. Every production deployment should assume the model will sometimes read a number that is not there, and design the validation layer accordingly.

> **Quality summary.** 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 of 95 per cent fall to roughly 75 per cent field accuracy on real material); medical tools moved across sites. Oversold: vision-against-standards products quoting no independent accuracy data; consumer-grade diagnostic apps. Underestimated: how good frontier models have become at ordinary document work when the workflow includes verification; for many SMBs this is the single highest-ROI AI capability available in 2026.

3.4 Video and image generation

The defining event of the period is a shutdown. OpenAI's Sora 2 launched on 30 September 2025 as an invite-only social app with synchronised audio and a likeness-insertion feature, generated enormous attention, restricted free access in January 2026, and was then discontinued: the consumer app closed 26 April 2026 and the API sunsets 24 September 2026 (OpenAI help centre, 2026). Commentary framed it as a monetisation failure and a caution about building business processes on consumer AI products (Futurum, 2026). The most-hyped product of late 2025 did not survive seven months.

The working frontier. Google's Veo 3.1 (October 2025) generates 4-to-8-second clips with native audio, extendable by chaining; consumer access runs from free tiers to the AI Ultra plan at roughly US$250 per month. Runway's Gen-4.5 generates up to 60 seconds in a single pass, the longest among major models, with character consistency features; its pricing is the honest corrective to "unlimited creativity" marketing, since the US$28 monthly Pro plan funds roughly 90 seconds of top-model footage at 25 credits per second. Kling 3.0 (February 2026) added multi-shot storyboards up to six scenes. Midjourney's V8.1 became default in June 2026 and remains the image leader for stylised work, with video limited to short image animation. In image generation, Google's Gemini 3 Pro Image ("Nano Banana Pro", from November 2025) substantially solved legible text rendering and multi-subject consistency, FLUX.2 (open weights, November 2025) leads open models, and per-image costs are now cents.

What "studio quality" actually means in mid-2026. Single establishing shots at 1080p with synchronised audio, often convincing on first viewing. It does not mean editable, continuity-stable, multi-shot footage. Persistent failure modes across all models, per 2026 stress tests: text on signs and labels, hands and fine motor action, physics under occlusion and contact, multi-character interaction, and character drift across shots. Clip length per generation remains 4 to 15 seconds for most models; "two-minute videos" are chained extensions with visible drift. For marketing b-roll, product mock-ups, and social content these limits are workable. For narrative or brand-critical film they are not, which is why professional use runs through heavy human editing.

Automated video pipelines. Higgsfield operates an MCP server that connects more than 30 video and image models (including Veo 3.1, Kling 3.0, and others) directly to Claude, Cowork, and Claude Code, with tools for marketing video generation and character-consistent production; the integration is real and first-party documented (higgsfield.ai/mcp, verified 13 June 2026). The revenue claims marketed around such pipelines ("replaces a five-figure monthly retainer") are unaudited and practitioner writing

is sceptical of them. The fully automated YouTube channel, meanwhile, has been squeezed by policy: YouTube's 15 July 2025 monetisation update demonetises "inauthentic" mass-produced content (YouTube creator policy communications, July 2025), including template AI voiceover over stock footage, while explicitly permitting AI-assisted content with genuine originality; synthetic-content disclosure has been mandatory for realistic AI material since 2024–25. The viable model in 2026 is AI-accelerated production with human curation, scripting judgement, and editing, at perhaps 70 to 90 per cent automation; the zero-touch content farm is largely a demonetised dead end.

> **Quality summary.** 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. 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. Oversold: “automated content engine” revenue claims; “studio quality” as a general proposition; and the durability of any single product in this category, as Sora’s seven-month arc demonstrates. Underestimated: how cheap and fast competent short-form visual content has become for businesses that previously could not afford it at all.

3.5 Audio, voice, and music generation

Voice agents. The platforms (Retell, Vapi, Bland, ElevenLabs Conversational AI, OpenAI's Realtime API) are commercially mature as of 2026: sub-second latency is routine, per-minute costs run from roughly US$0.06 to 0.33 all-in depending on stack (one of the few independent data points comes from a HackerNoon analysis of 4,000 measured production sessions, 2026, finding US$0.063 to 0.146 per minute on OpenAI's realtime-mini), and ElevenLabs raised US$500 million at a US$11 billion valuation in February 2026 (CNBC, 4 February 2026) with a reported but unaudited US$330 million-plus ARR. The evidential problem is that almost every published "benchmark" in this category is written by a vendor evaluating itself or a competitor; only one small independent academic cross-platform evaluation was found, and that scarcity is a finding.

What the credible evidence supports: AI handles routine, bounded calls (bookings, FAQs, intake, missed-call response) at customer satisfaction comparable to or slightly better than humans for simple tasks, and consumer preference inverts hard for complex or emotional matters (Salesforce survey, 2025). Realistic containment rates in production run around 50 per cent, not the 90-plus per cent in vendor decks (IrisAgent, 2026, itself a voice-AI

vendor, so this contrast should be read with the paper's own vendor-source caution). Known failure modes: accents and noisy audio degrade entity-level accuracy (times, names, account numbers) even when transcripts look fluent; interruption handling still requires tuning; and hallucinated commitments are a legal liability, with the Air Canada chatbot precedent (2024, airline held liable for its bot's invented refund policy) still the controlling logic. The telling detail is that vendors themselves keep humans in the loop: Smith.ai backs its AI receptionists with hundreds of human agents.

Text-to-speech and transcription. Short conversational synthetic speech is now, for practical purposes, indistinguishable from human; ElevenLabs' v3 (from June 2025) added directed emotional control via tags across 70-plus languages. Where synthesis still flags: long-form narration requiring global interpretation of a text (prosody drifts, and occasional spurious emotional artefacts appear in long generations) and lower-resourced languages. Transcription benchmark word-error rates of 4 to 5 per cent (OpenAI's GPT-4o-transcribe, Deepgram Nova-3, 2025–26) routinely become 15 to 20 per cent on real production audio with noise, accents, and crosstalk, a gap every vendor's own engineering documentation concedes. Medical scribe accuracy is examined in Section 3.10; the key transcription finding there is that omissions, not mistranscriptions, dominate errors, and error rates are measurably higher for some accents and dialects (summarised in an npj Digital Medicine commentary, Zhang, 2025, which draws on the underlying validation studies), which is a direct equity and clinical-risk issue.

Music. Suno (v5 generation; US$400 million raised at a US$5.4 billion valuation on 3 June 2026) and Udio produce radio-plausible complete songs from prompts; quality judgements rest on reviewer writing rather than any rigorous listening-test literature. The legal position moved decisively in 2025–26 but by settlement rather than judgement: Universal settled and licensed with Udio (October 2025), Warner settled with both Udio and Suno (November 2025), while Sony's and Universal's cases against Suno continue, with the plaintiffs moving in May 2026 to add 61,000-plus recordings to the complaint, and the musicians' union suing the majors over the settlements themselves (Music Business Worldwide; TechCrunch, June 2026). No court has ruled on whether training was fair use. Commercial use of AI music is therefore licensed-but-unsettled: viable for internal and low-stakes use, still carrying residual rights risk for brand-critical work.

Voice fraud. Voice cloning from seconds of audio is commodity capability, and deepfake-enabled fraud against businesses is real (the US$25 million Arup case in Hong Kong, 2024, is well documented), though most circulating loss statistics are low-confidence aggregator figures that could not be traced to primary data. The operational responses are not technical: callback verification on independently obtained

numbers, code words, and payment controls that no single phone call can override.

> **Quality summary.** 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. Oversold: “indistinguishable from human” as a blanket claim, fully autonomous phone-based sales, and most published voice-agent ROI statistics, which are vendor case studies. Underestimated: missed-call response and after-hours intake, which are narrow, cheap, and capture revenue that was previously simply lost; and the fraud exposure cut from the same cloning technology, which most SMBs have not yet addressed with even basic verification procedures.

3.6 Knowledge management and synthesis

The products. Google's NotebookLM is the category-defining tool: source-grounded chat, Audio and Video Overviews in 80 languages (September 2025), a Deep Research agent (November 2025), Cinematic Video Overviews (Google, March 2026, built on Gemini 3, Nano Banana Pro, and Veo 3, rolled out to Google AI Ultra subscribers at up to 20 per day), and source limits from 50 sources per notebook on the free tier to 600 on the top tier; an enterprise version runs inside Google Workspace governance. Anthropic's Claude Projects and OpenAI's ChatGPT Projects offer persistent, scoped workspaces; memory features became standard across all three major assistants between September 2025 and March 2026. Obsidian's ecosystem (local embeddings plus a locally run model) represents a small but real privacy-first counter-trend. Microsoft Copilot's knowledge features are widely deployed and, per multiple February 2026 enterprise reports, widely stalled, less for model-quality reasons than for governance ones: Copilot surfaces whatever the user technically has permission to see, which in most organisations includes a large volume of over-permissioned files.

The honest accuracy number. The most useful independent data point in this category: a 2025 preprint evaluating document-grounded assistants in a newsroom reporting workflow (arXiv:2509.25498) found NotebookLM, the best tool tested, still hallucinated in roughly 13 per cent of responses (against around 40 per cent for general assistants on the same corpus), including converting attributed opinions into flat statements of fact. Grounding in sources reduces hallucination; it does not eliminate it, and it adds a subtler risk, which is that cited answers look more trustworthy than they are. Retrieval-augmented systems in enterprise show the same pattern at scale: the canonical failure modes (wrong

chunks retrieved, right chunks missed, fluent answers unsupported by sources) are now well catalogued in the RAG literature (for a recent survey, arXiv:2510.09106, 2025).

Long context versus retrieval. Million-token context windows prompted a round of "RAG is dead" commentary in early 2026. The measured reality is that retrieval-style accuracy degrades well before the advertised window on every model tested, with near-perfect single-fact recall collapsing on realistic multi-fact tasks, and the economics favour retrieval by orders of magnitude for large corpora. The practitioner consensus by mid-2026: load the whole corpus into context when it is small and static; use retrieval for anything large, changing, or permission-controlled; most serious systems are hybrids.

How businesses actually use this, and where it breaks. Real, heavy use: meeting-transcript synthesis, document digestion (Audio Overviews turned out to be a sincerely popular consumption format), onboarding and policy Q&A, research triage. Where it breaks: stale corpora (the knowledge base answers from the 2023 version of the policy), permission leakage, the roughly 10-to-15 per cent grounded-hallucination floor, and the organisational failure of treating a knowledge assistant as a one-off install rather than an ongoing documentation-maintenance commitment. APQC survey work (2025–26) found 85 per cent of organisations had not operationalised AI knowledge tools at scale. The firms that succeed treat the document corpus as the product and the AI as the interface to it.

> **Quality summary.** 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. 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.

3.7 Code generation and software creation

The commercial facts are extraordinary and self-reported. Claude Code went from public launch (May 2025) to a reported US$2.5 billion annualised run-rate by February 2026 (company-reported alongside its Series G; VentureBeat, 2026), with the estimate that roughly 4 per cent of public GitHub commits are now authored by it (SemiAnalysis, February 2026, later cited by METR; commit-signature methodology, public repositories only, not independently audited). Cursor reached a reported US$2 billion ARR by February 2026, the fastest in business-software history, while raising at valuations approaching US$50 billion (TheNextWeb, 2026). GitHub's Copilot coding agent (assign an issue, receive a pull request) has been generally available since September 2025, with an "Agent HQ" running Copilot, Claude, and Codex agents side by side in preview from early 2026 (GitHub changelog, 2026). Among the "vibe-coding" platforms for non-developers, Lovable was estimated at around US$400 million annualised revenue as of February 2026 (Sacra estimate) and Replit projected US$1 billion for the year (company statements, 2026), while reporting acknowledged churn problems from one-off project completion (Sacra interview series, 2026), an unresolved divergence with the headline revenue figures worth noticing.

The productivity evidence is genuinely contested, and the best study says so itself. METR's randomised trial (July 2025) found experienced open-source developers were 19 per cent slower with early-2025 AI tools while believing themselves 20-plus per cent faster, the single most quoted sceptical result in the field. METR's February 2026 follow-up found point estimates flipping to modest speedups with early-2026 tools but with confidence intervals too wide, and selection effects too severe, to support a firm number; METR's own interpretation is that developers are probably now somewhat faster, by an unmeasurable margin. Around these anchors: Google's DORA 2025 report (90 per cent adoption) found AI increases throughput and instability together, with a measurable "verification tax"; Stack Overflow's 2025 survey of 49,000 developers found 84 per cent using AI while distrust of its accuracy rose year over year to 46 per cent, and 45 per cent saying debugging AI code takes longer than writing it; GitClear's analysis of 211 million changed lines found code duplication up roughly 50 per cent and refactoring collapsing since 2021, the strongest evidence that AI-assisted speed is being partly borrowed from future maintainability.

What a non-developer can ship in mid-2026. Reliably: prototypes, landing pages, internal tools, and simple CRUD applications on managed backends, which is a real and underappreciated expansion of who can make software. Not reliably: anything with authentication, payments, sensitive data, or adversarial users. The security evidence is

now substantial, though much of it comes from application-security vendors and should be read with that in mind: Escape.tech's October 2025 scan of roughly 5,600 publicly available vibe-coded applications found over 2,000 high-impact vulnerabilities, 400-plus exposed secrets, and 175 instances of personal-data exposure including medical records and bank-account numbers (Escape.tech, "The State of Security of Vibe Coded Apps," 2025; Escape.tech is an application-security vendor, so read the framing with that in mind); a May 2025 study of 1,645 sampled Lovable-built apps found about 10 per cent (170 apps) exposed personal data and around 70 per cent had row-level security disabled (reported via TheNextWeb, 2026); and the Tea app breach (July 2025, roughly 72,000 images including 13,000 government IDs from an unsecured database; Security.org, 2025) became the category's cautionary tale, though attributing that specific incident to vibe coding is contested. The benchmark story told in sales decks, SWE-bench Verified rising from 60 per cent to near 100 per cent in a single year (Stanford HAI AI Index, 2026), should be heavily discounted for contamination, which OpenAI itself documented in February 2026; the contamination-resistant SWE-bench Pro sits near 80 per cent for the best models, and real codebases are harder still.

What still needs a developer. System architecture, debugging production systems, security review, legacy integration, and, increasingly, reviewing the flood of AI-generated output. The consistent 2026 finding is role shift rather than role elimination: implementers becoming orchestrators and reviewers, with judgement as the scarce input.

> **Quality summary.** 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. Oversold: “anyone can ship production software”; benchmark-derived capability claims. 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.

3.8 Business automation platforms with AI built in

The category's defining tension is that platforms sell autonomy and deliver, at best, supervised competence. The most instructive data point is pricing: between late 2025 and April 2026, HubSpot moved its Breeze agents to per-outcome pricing (US$0.50 per resolved conversation), Intercom's Fin charges US$0.99 per resolution, and Salesforce restructured Agentforce pricing twice, landing on consumption credits. Outcome pricing is the market's tacit admission that seat-based AI pricing was not justified by delivered value.

The evidence by platform. Salesforce Agentforce shows the clearest marketing-to-reality gap: independent ecosystem reporting (Salesforce Ben, April 2025) described adoption as slow, with the CFO guiding to "modest" near-term Agentforce sales and roughly 3,000 paid customers (against about 5,000 deals, 2,000 still testing); implementation-partner analyses put first-year total cost of ownership at US$150,000 to 425,000 once data work and implementation are counted. HubSpot's Customer Agent claims 65 per cent resolution across 8,000-plus activations (vendor); user reports split between satisfied simple-FAQ deployments and "not worth it", with outcomes hinging almost entirely on knowledge-base quality. Intercom's Fin claims 65 to 70 per cent resolution; its own published case studies cluster at 42 to 53 per cent, and a January 2026 evaluation of complex multi-step support tickets found best-case 24 per cent success across leading tools. The repeatable pattern across customer-service AI: the claimed rate is roughly 1.4 to 1.8 times the delivered rate, and the delivered rate depends on how much routine, well-documented traffic the queue contains.

GoHighLevel, the platform underneath thousands of agency-sold "AI employee" offerings (US$97 per month add-on, throttled rather than unlimited), does its core job adequately: missed-call text-back, appointment booking, FAQ handling for local service businesses. The buyer's hazard is the reseller layer, since most published reviews carry affiliate incentives, and no independent performance audit of the platform was found. The horizontal automation tools (Zapier Agents out of beta, Make, n8n 2.0, all January 2026) are the quiet workhorses of the category. The widely circulated claim that "79 per cent of organisations run agents in production with 171 per cent ROI" comes from PwC's 2025 AI Agent Survey, which actually found 79 per cent had implemented AI agents "at some level" (not specifically in production) and respondents projecting an average ROI of 171 per cent (192 per cent for US enterprises); these are self-reported, projected figures from a consultancy survey, not audited returns, and should be read accordingly (PwC, 2025).

The Microsoft 365 Copilot result deserves its own paragraph because it is the largest government-scale evaluation in the category: a UK government evaluation (the Government Digital Service cross-government findings report, June 2025) of roughly

20,000 civil servants given Microsoft 365 Copilot for three months found high user satisfaction and self-reported time savings of around 26 minutes per day, but the evaluation was not a randomised controlled trial and did not establish productivity gains under controlled measurement; users felt faster at drafting and summarising, were measurably weaker at spreadsheet analysis and slide creation, and satisfaction was high regardless. Microsoft's commissioned Forrester studies projecting 112 to 457 per cent ROI model hypothetical composite organisations and should be read as marketing collateral. The perception-versus-measurement gap (people feel faster; the measurements disagree) recurs across this entire paper.

AI receptionists and missed-call text-back are the productised features most relevant to PAC's readership. Typical pricing is US$99 to 299 per month for 24/7 coverage with booking. The underlying pitch statistics ("62 per cent of SMB calls go unanswered") trace to old vendor-commissioned research and could not be verified. The honest case does not need them: a bounded agent that answers after-hours calls, books appointments, and texts back missed calls captures revenue that was previously lost outright, fails visibly when it fails, and costs little. Realistic expectations: roughly half of routine calls fully contained, the rest handed off, and a configuration-and-monitoring obligation that does not disappear after week one.

> **Quality summary.** 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” demos that conceal the throttles, the configuration burden, and the escalation tail. Oversold: enterprise agent suites as plug-and-play (the data engineering is the project); virtually all published ROI numbers in this category, none of which are independently audited. 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.

3.9 Crypto and AI convergence

This section is included because the area is under-reported rather than because it is mature. The honest framing: the infrastructure is consolidating impressively; the demand is small; and one flagship consumer experiment has already been shut down.

NEAR Protocol. NEAR runs the most coherent crypto-AI stack. Its confidential compute offering (inference inside trusted execution environments, so prompts and models stay encrypted even from the operator) is live and documented, with an "IronClaw" agent runtime and a confidential GPU marketplace launched at NEARCON 2026; all claims about it are vendor-sourced, with no independent security audit or named enterprise customers found. NEAR Intents, its cross-chain transaction layer, has on-chain-verifiable volume: roughly US$10 billion cumulative by 16 January 2026 and about US$20 billion by early June 2026 (Dune Analytics; DefiLlama). The sceptical caveat is in the framing rather than the numbers: Intents volume is overwhelmingly swap-and-bridge activity, closer to a decentralised exchange than to "AI agent commerce", and NEAR's agent marketplace (market.near.ai), the closest thing to its advertised agentic job board, is live but published no volume, agent-count, or earnings data that could be found. Operational, yes; proven as an agent economy, no.

Agentic commerce: the retreat and the rails. The most clarifying event of the period was a failure: OpenAI and Stripe launched Instant Checkout inside ChatGPT in September 2025 under their co-developed Agentic Commerce Protocol, and OpenAI sunset it on 24 March 2026 with only around 30 Shopify merchants ever live, citing merchant onboarding and product-data quality problems and, as of February 2026, no infrastructure to collect US sales tax (CNBC, 24 March 2026; corroborated by trade coverage); Walmart reportedly saw three times lower conversion in-chat than on its own site, a figure CNBC confirms. The protocol layer kept consolidating anyway: Google's AP2 payment-mandate protocol moved to the FIDO Alliance in May 2026 with working groups chaired by Mastercard and Visa; Visa's and Mastercard's agent-payment programmes remain pilots, though Visa separately reported a roughly US$7 billion annualised stablecoin-settlement run-rate across nine blockchains by late April 2026 (Visa investor relations; CoinDesk), which is stablecoin settlement rather than agentic-payment volume, the two being routinely conflated; Microsoft launched Copilot Checkout in January 2026 (no volume data published); and Stripe's Paradigm-incubated Tempo blockchain went live on 18 March 2026 with a Machine Payments Protocol for per-call agent micropayments, following Stripe's US$1.1 billion Bridge acquisition. The one machine-payment rail with hard numbers, Coinbase and Cloudflare's x402, reported 165 million transactions and roughly US$50 million cumulative volume by late April 2026, an average of about 30 cents per transaction. Machine payments are real and tiny.

The rest of the field. Bittensor illustrates the verification problem: promoted figures of US$43 million in quarterly AI revenue stand against an independent March 2026 analysis estimating US$3 to 15 million annually in genuine external revenue, a tenfold-plus gap, with token emissions rather than customer demand driving most capital flows. Fetch.ai's

unified ASI chain remained unlaunched two years after its token merger, targeted for late 2026. Virtuals Protocol's agent economy was, by the first quarter of 2026, paying agents subsidies to have revenue. The pattern for a business reader: treat token-denominated activity metrics as marketing until traced to a neutral on-chain dashboard, and treat "agent economy" claims as aspirational by default.

> **Quality summary.** Deploy today: for almost all PAC readers, nothing; this is a watch category. Stablecoin settlement for international payments is the adjacent capability with real near-term business relevance. Demo-versus-production gap: agent-to-agent commerce demos against the absence of any consumer-scale deployment that has survived contact with merchants. Oversold: token projects citing volume that is actually speculation or subsidy; “the agent economy is here”. Underestimated: the rails themselves; the speed with which payments incumbents (Visa, Mastercard, Stripe, PayPal, the FIDO Alliance) standardised agent-payment authorisation suggests they expect the demand to arrive, and confidential compute for AI workloads addresses a real privacy problem regulated firms actually have.

3.10 Vertical AI applications

Vertical products, built for one industry's workflow rather than general capability, are where the most defensible 2026 value sits, and also where Australian products are most visible. Each vertical below gets the same treatment: what shipped, what the independent evidence says, and the local regulatory position.

Clinical documentation. AI scribes are the most validated vertical category. Melbourne's Heidi Health raised US$65 million (Series B, October 2025, Point72-led, US$465 million valuation) and expanded through 2026 (UK acquisition, evidence and communications products, a wearable microphone); Melbourne's Lyrebird Health raised US$12 million in June 2025 with vendor-claimed usage of tens of thousands of consults daily; Abridge reached a US$5.3 billion valuation (June 2025, extended April 2026); Microsoft's Dragon Copilot claims 100,000-plus daily clinicians. The independent evidence says the benefit is real and modest: a large multi-site study reported about 16 minutes saved per 8 hours of patient care with inconsistent usage (STAT, 1 April 2026); JAMA cohort studies found reduced documentation time and a 21 per cent reduction in burnout measures; a NEJM AI study at Atrium Health concluded the leading product was unlikely to produce appreciable system-level productivity gains. Vendor claims of 40-to-60 per cent documentation-time reductions are survey-based and unverified. Error profiles matter clinically: omissions dominate (reported at 86 per cent of errors in one validation study, cited in an npj Digital

Medicine commentary; Zhang, 2025), and error rates are higher for some accents. In Australia, AHPRA guidance requires human judgement over outputs and informed patient consent; the RACGP requires GP oversight and encourages GPs-in-training to develop documentation skills before relying on AI scribes; and the TGA clarified on 30 January 2026 that scribes which interpret clinical conversations are regulated medical devices requiring ARTG inclusion, a compliance fact some practices using offshore tools have not yet absorbed.

Legal. Harvey raised US$200 million at a US$11 billion valuation in March 2026, claiming 42 per cent of the AmLaw 100; Thomson Reuters announced one million CoCounsel professional users in February 2026. Against this, the peer-reviewed Stanford RegLab study (preregistered 2024, published in the Journal of Empirical Legal Studies, 2025) found Lexis+ AI hallucinating in 17 per cent of responses and Westlaw's AI research in 33 per cent, both marketed at the time with hallucination-free language. The consequences are no longer hypothetical in Australia: the first Australian solicitor was sanctioned for filing AI-fabricated citations (Federal Circuit and Family Court matter; regulatory outcome 2025), dozens of further incidents are catalogued in a community-maintained database, and the Federal Court issued practice note GPN-AI on 16 April 2026 requiring practitioners to personally verify that AI-sourced authorities exist and support the propositions cited, with parallel notes in NSW (which bans generative AI for affidavits and expert reports without leave, from February 2025), Victoria, Queensland, and South Australia. The honest position: legal AI is a strong research and drafting accelerant whose output must be verified line by line, and the courts have now made that verification a professional obligation rather than a best practice.

Financial advice. Adviser Ratings reported (May 2025) that 74 per cent of Australian advice practices were using or planning to use AI, up from 45 per cent a year earlier, with SOA/ROA production a use case for nearly half (46 per cent); the same report put the share of practices with AI policies or guidance well short of adoption (a global figure of around 45 per cent), a gap that should alarm licensees. ASIC has issued no AI-specific rulebook; existing obligations (best interests duty, records, licensee supervision) apply in full to AI-assisted advice, and ASIC's 2024 review (REP 798) found governance lagging adoption. SOA-drafting tools (Padua, and newer entrants claiming roughly 60 per cent preparation-time cuts) publish no independent validation. No ASIC enforcement action specifically over AI-drafted advice documents had been identified as of June 2026; the first one will define the category.

Real estate. Listing copy is a solved, commodity use case requiring no specialist tool. Valuation is the substantive frontier: CoreLogic's AVM claims roughly 90 per cent of

estimates within 15 per cent of sale price (vendor, 2024), PropTrack covers 12.6 million properties, and image-augmented valuation shows promise on academic evidence with headline accuracy claims resting on a single unverified source. RICS's responsible-AI standard became mandatory for regulated surveyors on 9 March 2026 and explicitly flags AVMs and defect detection as high-risk uses. No major 2026 consumer-facing generative feature from REA or Domain was verified during research.

Construction and inspection. Progress tracking against BIM (OpenSpace, which acquired Disperse in November 2025; Buildots) is in production on large projects with vendor-case-study evidence only. Australian defect-detection products (Voltin's facade inspection; Defect Detector's NCC-compliance reports claiming 95 per cent-plus accuracy, unaudited) are early. The capability summary from Section 3.3 carries over: strong on visually distinct surface defects, weak on the concealed defects that drive disputes, and not a substitute for a qualified inspector's liability-bearing judgement.

Accounting and bookkeeping. The reliable shipped layer is confidence-gated transaction processing. Xero's JAX rebuild (from September 2025) ships auto bank reconciliation in open beta (Xero Grow in AU/NZ/UK, Growing in the US, Standard and above globally), matching or creating transactions only at high confidence and leaving low-confidence statement lines for manual review, a design independent commentary singled out as the right pattern (Xero Central; VentureBeat, 2025–26). Dext extends capture into reviewable categorisation; Fathom ships AI report commentary; Intuit Assist automates QuickBooks categorisation. No independent accuracy benchmark exists for any of them, and Spotlight Reporting's AI specifics could not be verified at writing time. Autonomous bookkeeping is not a real product in 2026; supervised coding-and-reconciliation is, and it works.

HR documentation. Employment Hero (last confirmed valuation around A$2 billion, 2023, with AI across recruitment and payroll workflows) and HiBob anchor the category locally. Generated policies, contracts, and position descriptions are competent first drafts; Australian employment law is jurisdiction-specific and award-complex enough that unreviewed AI HR documents are a liability, not a saving.

Trades. ServiceM8 ships an AI quote assistant (voice note to draft quote) and smart scheduling; Tradify, AroFlo, Fergus, and Simpro have added AI-assisted features over the past 18 months, with Fergus deliberately minimal. The sourcing for this sub-category is largely SEO-grade review sites, so feature claims should be checked against vendor changelogs before purchase. The realistic appraisal: quoting from voice notes, job summaries, and invoice chasing are real time-savers for one-to-twenty-person operations; none of it is autonomous, and none of it needs to be.

Quality summary. Deploy today: AI scribes under AHPRA/RACGP-compliant consent and review workflows; legal research acceleration with mandatory citation verification; confidence-gated bookkeeping automation; quote drafting in trade platforms. Demo-versus-production gap: time-saved claims (vendor surveys say 40 to 60 per cent; controlled studies say minutes per day that are nonetheless worth having). Oversold: “hallucination-free” professional tools, autonomous compliance documents of any kind (SOAs, building reports, HR policies). Underestimated: the compounding value of modest, validated savings inside high-frequency workflows, and the degree to which Australian regulators have already published the rules of the road; the compliance posture for most professions is documented, achievable, and mostly ignored.

4. Discussion

The two gaps are the macro picture. Every category in this survey exhibits the same pair of gaps. The first is benchmark-to-production: 95 per cent on a curated benchmark and 75 per cent on real documents; 4 per cent word-error rates becoming 15 to 20 on real calls; SWE-bench in the nineties and simulated office work at 30. The second is individual-to-firm: randomised trials keep finding real individual gains (the 2023 BCG experiment, scribe burnout reductions, drafting speedups) while firm-level surveys keep finding most companies without enterprise-level earnings impact (McKinsey finding more than 80 per cent with no tangible enterprise-level impact, and Deloitte only 15 per cent reporting significant ROI). Three years into the era, the gap between what the technology can do for a person and what it does for a profit-and-loss statement remains the defining economic fact, and the explanation has stabilised: 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 that is a management problem. A striking fraction of what a buyer will read about AI in 2026 is written by interested parties: vendor benchmarks, affiliate reviews, commissioned ROI studies modelling hypothetical companies, revenue run-rates told to investors, and SEO content farms that invent specifics. Several widely circulated claims investigated for this paper, including model releases, could not be traced past a single low-credibility site. The few genuinely independent institutions (METR, Epoch AI, Stanford's RegLab and HAI, DORA, academic trial groups, government evaluations like the UK Copilot study) consistently report more modest results than the surrounding marketing, and their work lags products by six to twelve months. The operational implication is that "what does the vendor's evidence actually consist of" is now a core procurement question.

What is shifting fastest. Three things. Agentic coding, where capability, adoption, and revenue are all compounding and where METR's task-horizon doubling (every four to seven months) is the single most important trend line in the field. Cost, where capability-adjusted inference prices fall roughly an order of magnitude or more per year, continuously changing which use cases clear the value bar. And infrastructure standardisation, where MCP, the skills format, and the agent-payment protocols moved into neutral foundations within about 14 months of existing, far faster than equivalent past standards.

What is still genuinely hard. Reliability at the tail: every system surveyed performs well on the routine middle of its distribution and degrades on the edge cases where business risk concentrates. Verification: there is still no general way for a model to know it is wrong, and the consensus account of hallucination (training rewards confident answers over abstention) implies the problem is structural, managed rather than solved. Security for agents: prompt injection is unsolved by the assessment of the vendors themselves, which bounds how much autonomy any prudent business grants. And measurement: the perception gap, in which users feel faster than they measurably are, appears in developer trials and government pilots alike, and means self-reported satisfaction cannot be the success metric for an AI deployment.

The economics overhang. Hyperscaler capital expenditure of roughly US$630 to 725 billion guided for 2026 stands against frontier labs that are unprofitable on enormous revenue (Anthropic's reported run-rate of tens of billions; OpenAI's losses, inferred at roughly US$12 billion for one quarter from Microsoft's SEC filings and projected by Deutsche Bank at around US$143 billion of cumulative negative free cash flow through 2029, none of it OpenAI-audited), documented circular-financing structures (Bloomberg, 2026), and credible scepticism from within the financial establishment alongside credible demand evidence. Goldman Sachs chief executive David Solomon, speaking on the record at the Economic Club of Washington, D.C., put it plainly: "a lot of the capital that's being deployed will not produce adequate returns. And a bunch of the capital that's being deployed will actually not produce any returns" (Banking Dive; Yahoo Finance, 2026). This paper takes no position on the bubble question except the operational one: businesses should adopt in ways that capture value at today's prices without depending on any particular vendor, price, or product surviving. The Sora shutdown, the Operator shutdown, and the Instant Checkout shutdown all occurred within nine months, in products from the best-resourced labs in the field. Build on capabilities and standards; rent products.

Employment, briefly. The evidence supports concentrated effects, not an apocalypse: a roughly 13 per cent relative employment decline for 22-to-25-year-olds in the most AI-exposed occupations since late 2022 (Stanford Digital Economy Lab, August 2025), against no detectable aggregate employment effect in national data through mid-2026. For operational leaders the practical reading is about role design: entry-level work that consists of producing routine first drafts is being absorbed, and the development pathway for juniors needs deliberate rebuilding around verification, judgement, and client work.

5. Implications for Operational Leaders

A reasonable adoption posture in mid-2026 follows from the evidence rather than from sentiment, and it can be stated concretely.

Deploy now, with verification built in. The capabilities that clear the production bar share a shape: bounded task, verifiable output, human or deterministic check on low-confidence cases. In practice that means document extraction with confidence gating; meeting and consult documentation under professional review; grounded knowledge assistants over curated corpora; coding assistance inside professional review processes; missed-call and after-hours voice handling with escalation; confidence-gated bookkeeping; drafting acceleration everywhere, with the draft never leaving the building unreviewed. These are not pilots; they are dependable operations work, and most of PAC's readership is under-deployed on them.

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 rather than instead of it. Set the metric before the pilot (containment rate, error rate against human baseline, minutes saved measured rather than reported) and prefer vendors on outcome pricing, who are at least making a falsifiable claim.

Wait on unsupervised agents for consequential actions; autonomous compliance documents; agent-initiated payments; anything whose pitch depends on a vendor benchmark or an unaudited ROI figure. The cost of waiting in this field is low, since capability per dollar improves continuously, while the cost of a premature deployment that touches clients, money, or regulators is not.

Walk away from "hallucination-free" claims, "AI employee" propositions sold without containment-rate evidence, token-denominated AI investment pitches, and any vendor who cannot answer what happens when the system is wrong.

Three structural recommendations. First, treat data and documentation as the asset: every category surveyed shows deployment success tracking corpus quality more closely than model choice. Second, build the verification habit into roles, since the perception gap means individuals are unreliable narrators of their own productivity; measure. Third, mind the regulatory floor, which in Australia is already specific: AHPRA and RACGP positions for clinicians, the TGA's January 2026 device classification for scribes, court practice notes for lawyers, ASIC's existing-obligations stance for advisers, RICS's mandatory standard for

surveyors, and the Privacy Act's automated-decision-making transparency obligations commencing 10 December 2026 for everyone. 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 mid-2026 is that artificial intelligence has become reliable infrastructure for a specific class of work, bounded, verifiable, supervised, while remaining an unreliable narrator of its own broader competence. The twelve months to June 2026 delivered material change: million-token context as standard, reasoning scores through the human average on the field's hardest public benchmark, agentic coding as the first proven product of the agent era, and the integration layer standardised under neutral governance. The same twelve months delivered the field's most instructive failures: three flagship products from the best-funded labs shut down, a court injunction against an agentic browser, peer-reviewed hallucination rates in tools marketed as hallucination-free, and the persistent refusal of firm-level financial results to match individual-level gains.

Neither the enthusiast's story nor the sceptic's story survives contact with the evidence. The enthusiast must explain why the best independent agent evaluation shows 30 per cent task completion, why the largest government-scale Copilot evaluation found high satisfaction and self-reported time savings but did not establish productivity gains under controlled measurement, and why outcome-based pricing swept the automation category. The sceptic must explain US$2.5 billion of annualised coding-agent revenue, randomised-trial evidence of real clinical documentation relief, a 44 per cent screening-workload reduction in a 106,000-woman trial, and task horizons doubling every few months. Both things are true, in different places, and knowing which place a given purchase decision sits in is most of the job.

For the operational leader, the disciplined posture is neither adoption nor abstention but engineering: deploy the proven capabilities with verification built in, pilot the plausible ones against falsifiable metrics, ignore the unaudited numbers, and assume that anything true about the field today carries a date stamp. This paper carries one too. It describes June 2026, it will be partly wrong by June 2027, and the sections most likely to age fastest, agent reliability and cost, are the ones worth re-checking first.

References

All sources accessed 13 June 2026 unless otherwise noted. Where a source is a vendor's own publication, this is evident from the publisher; Section 1.3 and Appendix A describe how vendor material was weighted.

Adviser Ratings. (2025, May 14). *The AI revolution in financial advice: Australian practices leading global adoption* (industry/commercial self-report; "74%" includes practices using or planning to use AI). https://www.adviserratings.com.au/news/the-ai-revolution-in-financial-advice-australian-practices-leading-global-adoption/

AHPRA. (2024). *Meeting your professional obligations when using AI in healthcare.* https://www.ahpra.gov.au/Resources/Artificial-Intelligence-in-healthcare.aspx

Anthropic. (2025, December). *Donating the Model Context Protocol and establishing the Agentic AI Foundation.* https://www.anthropic.com/news/donating-the-model-context-protocol

Anthropic. (2026, February 5). *Introducing Claude Opus 4.6.* https://www.anthropic.com/news/claude-opus-4-6

Anthropic. (2026, June 9). *Claude Fable 5.* https://www.anthropic.com/claude/fable

APQC. (2025–2026). *Study warns of a looming "Great Retirement" crisis and highlights the role of AI* [survey of 1,000 professionals, sponsored by eGain; instrument not independently examined]. https://www.apqc.org/about-apqc/news-press-release/apqc-study-warns-looming-great-retirement-crisis-highlights-role-ai

Arc Prize Foundation. (2026). *ARC-AGI leaderboard.* https://arcprize.org/leaderboard

ASIC. (2024, October 29). *REP 798: Beware the gap: Governance arrangements in the face of AI innovation.* https://www.asic.gov.au/about-asic/news-centre/find-a-media-release/2024-releases/24-238mr-asic-warns-governance-gap-could-emerge-in-first-report-on-ai-adoption-by-licensees/

Banking Dive. (2026). *Goldman CEO Solomon on AI capital returns, banking, and digital assets* [Solomon's on-the-record remarks at the Economic Club of Washington, D.C.]. https://www.bankingdive.com/news/goldman-ceo-solomon-ai-banking-rto-fed-digital-assets/804363/

Bloomberg. (2026). *The circular deals powering the AI boom.* https://www.bloomberg.com/graphics/2026-ai-circular-deals/

British Journal of Dermatology. (2026). *Prospective evaluation of the SkinVision smartphone skin-cancer detection app* [n≈1,458; more than 70% of lesions not adequately captured in

patient self-photography]. https://academic.oup.com/bjd/advance-article-abstract/doi/10.1093/bjd/ljag057/8488698

Carnegie Mellon University. (2025). *TheAgentCompany: Benchmarking LLM agents on consequential real-world tasks.* https://www.cs.cmu.edu/news/2025/agent-company

Chaikin, A., & Sahib, S. K. (2025, August 20). *Agentic browser security: Indirect prompt injection in Perplexity Comet.* Brave. (Brave is a competing browser vendor; OTP-exfiltration demonstration independently corroborated by Willison, 25 August 2025.) https://brave.com/blog/comet-prompt-injection/

Civil Resolution Tribunal of British Columbia. (2024). *Moffatt v. Air Canada, 2024 BCCRT 149* [airline held liable for its chatbot's incorrect statements]. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html

CNBC. (2026, February 4). *Nvidia-backed AI startup ElevenLabs hits $11 billion valuation.* https://www.cnbc.com/2026/02/04/nvidia-backed-ai-startup-elevenlabs-11-billion-valuation.html

CNBC. (2026, February 6). *Google, Microsoft, Meta and Amazon AI capital expenditure guidance.* https://www.cnbc.com/2026/02/06/google-microsoft-meta-amazon-ai-cash.html

CNBC. (2026, March 10). *Amazon wins court order to block Perplexity's AI shopping agent* [N.D. Cal. preliminary injunction, Judge Maxine M. Chesney; Ninth Circuit temporary stay 16 March 2026]. https://www.cnbc.com/2026/03/10/amazon-wins-court-order-to-block-perplexitys-ai-shopping-agent.html

CNBC. (2026, March 24). *OpenAI revamps shopping experience in ChatGPT after struggling with Instant Checkout* [24 March 2026 sunset, ~30 Shopify merchants]. https://www.cnbc.com/2026/03/24/openai-revamps-shopping-experience-in-chatgpt-after-instant-checkout.html

CNBC. (2026, March 25). *Legal AI startup Harvey raises $200 million at $11 billion valuation.* https://www.cnbc.com/2026/03/25/legal-ai-startup-harvey-raises-200-million-at-11-billion-valuation.html

CoinDesk. (2026, April 29). *Visa expands stablecoin settlement network as volume hits US$7 billion run-rate.* https://www.coindesk.com/business/2026/04/29/visa-expands-stablecoin-settlement-network-as-volume-hits-usd7-billion-run-rate

DefiLlama. (2026). *NEAR Intents protocol* [on-chain volume dashboard]. https://defillama.com/protocol/near-intents

Dell'Acqua, F., McFowland, E., Mollick, E., et al. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality.* Harvard Business School Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321

Deloitte. (2026). *State of AI in the enterprise 2026* [n=3,235 senior leaders, August–September 2025, 24 countries, confirmed; the 15% significant-ROI figure remains publisher-reported; consultancy self-report]. https://www.deloitte.com/global/en/issues/generative-ai/state-of-ai-in-enterprise.html

Demand Sage. (2026). *ChatGPT statistics* [third-party aggregator; figures originally from UBS/Similarweb estimates]. https://www.demandsage.com/chatgpt-statistics/

Department of Industry, Science and Resources (Australia). (2024). *Voluntary AI Safety Standard.* https://www.industry.gov.au/publications/voluntary-ai-safety-standard

Dermatology Times. (2024). *FDA clears DermaSensor device for skin cancer detection.* https://www.dermatologytimes.com/view/fda-clears-dermasensor-device-for-skin-cancer-detection

DORA / Google Cloud. (2025). *2025 DORA state of AI-assisted software development report.* https://dora.dev/insights/balancing-ai-tensions/

Dune Analytics. (2026). *NEAR Intents dashboard* [on-chain data; ~US$10B cumulative volume by 16 January 2026 and ~US$20B by 3 June 2026]. https://dune.com/near/near-intents

eMarketer. (2026). *OpenAI forecast to post around US$143 billion cumulative loss* [Deutsche Bank projection of cumulative negative free cash flow, 2024 to 2029]. https://www.emarketer.com/content/openai-forecast-143-billion-loss

Epoch AI. (2025–2026). *LLM inference price trends.* https://epoch.ai/data-insights/llm-inference-price-trends

Epoch AI. (2026, April). *Benchmark saturation: MMLU-Pro* [frontier models clustered above ~89%]. https://epoch.ai/data-insights/

Epoch AI. (2026). *SWE-bench Verified independent evaluations.* https://epoch.ai/benchmarks/swe-bench-verified

Escape.tech. (2025). *The state of security of vibe coded apps* [~5,600 apps scanned October 2025, 2,000+ high-impact vulnerabilities, 400+ exposed secrets, 175 personal-data exposures; application-security vendor]. https://escape.tech/state-of-security-of-vibe-coded-apps

FDA. (2025). *Artificial intelligence and machine learning (AI/ML)-enabled medical devices* [running list; cumulative total exceeded 1,250 by mid-2025]. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

Federal Court of Australia. (2026, April 16). *Practice note GPN-AI: Use of generative artificial intelligence.* https://www.fedcourt.gov.au/law-and-practice/practice-documents/practice-notes/gpn-ai

Florida Legislature. (2025). *House Bill 683: Authorising automated and software-based building plan review.* https://www.flsenate.gov/

Fortune. (2025, December 23). *OpenAI on prompt injection in AI browsers.* (OpenAI characterised prompt injection as "unlikely to ever be fully 'solved'.") https://fortune.com/2025/12/23/openai-ai-browser-prompt-injections-cybersecurity-hackers/

Fortune. (2026, June 1). *Anthropic confidentially files IPO at $965 billion valuation.* https://fortune.com/2026/06/01/anthropic-confidentially-files-ipo-965-billion-valuation/

Fortune. (2026, June 5). *Is AI a bubble? Goldman's early sceptic on profits.* https://fortune.com/2026/06/05/is-ai-a-bubble-worth-it-short-term-early-goldman-skeptic-covello-profits/

Futuriom. (2025, August). *Why we don't believe MIT NANDA's weird AI study.* https://www.futuriom.com/articles/news/why-we-dont-believe-mit-nandas-werid-ai-study/2025/08

Futurum Group. (2026). *OpenAI Sora discontinuation: What the end of a platform means for enterprise AI strategy.* https://futurumgroup.com/insights/openai-sora-discontinuation-what-the-end-of-a-platform-means-for-enterprise-ai-strategy/

Gartner. (2025, June 25). *Gartner predicts over 40% of agentic AI projects will be canceled by end of 2027.* https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

GeekWire. (2026, March). *Coverage of the Amazon v. Perplexity preliminary injunction* [N.D. Cal., Judge Maxine M. Chesney, 10 March 2026; Ninth Circuit temporary administrative stay 16 March 2026]. https://www.geekwire.com/

Gibson Dunn. (2026, May). *EU AI Act omnibus agreement: Postponed high-risk deadlines and other key changes.* https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/

GitClear. (2025). *AI assistant code quality research: 211 million changed lines.* https://www.gitclear.com/ai_assistant_code_quality_2025_research

GitHub. (2023, February). *Research: Quantifying GitHub Copilot's impact on developer productivity.* https://arxiv.org/abs/2302.06590

Google. (2025, November 18). *Gemini 3 Pro.* https://blog.google/products/gemini/gemini-3/

Google. (2026, February 19). *Gemini 3.1 Pro.* https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/

Google. (2026, March). *Generate your own Cinematic Video Overviews in NotebookLM* [Gemini 3, Nano Banana Pro, and Veo 3; Google AI Ultra; up to 20 per day]. https://blog.google/innovation-and-ai/products/notebooklm/generate-your-own-cinematic-video-overviews-in-notebooklm/

Government Digital Service (UK). (2025, June). *Microsoft 365 Copilot experiment: Cross-government findings report.* (DBT was one of 12 participating organisations; the evaluation covered roughly 20,000 employees and was not a randomised controlled trial.) https://assets.publishing.service.gov.uk/media/683db42bd23a62e5d32680d0/M365_Copilot_Experiment_Findings_Report.pdf

HackerNoon. (2026). *OpenAI Realtime API pricing in 2026: Real-world data from 4,000 measured sessions.* https://hackernoon.com/openai-realtime-api-pricing-in-2026-real-world-data-from-4000-measured-sessions

Hagar, N., Agustianto, W., & Diakopoulos, N. (2025). *Not wrong, but untrue: LLM overconfidence in document-based queries* (arXiv:2509.25498; accepted to the Computation + Journalism Symposium 2025). https://arxiv.org/abs/2509.25498

Higgsfield. (2026). *Higgsfield MCP* [documentation, fetched directly 13 June 2026]. https://higgsfield.ai/mcp

Innolitics. (2025, December 20). *2025 year in review: AI/ML medical device 510(k) clearances* (reports 295 clearances during 2025; the cumulative ~1,250 figure is the FDA's running list). https://innolitics.com/articles/year-in-review-ai-ml-medical-device-k-clearances/

IrisAgent. (2026). *Voice AI customer service: 2026 benchmarks.* https://irisagent.com/blog/voice-ai-customer-service-2026-benchmarks/

JAMA Network Open. (2025). *Cohort studies of ambient AI clinical documentation* [reduced documentation time; an associated roughly 21% reduction in burnout measures across two systems]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12514625/

Lång, K., et al. (2026). *Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence (MASAI) trial: Final results.* The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)02464-X/abstract

Linux Foundation. (2025, December). *Linux Foundation announces the formation of the Agentic AI Foundation.* https://www.linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation

LlamaIndex. (2026). *OmniDocBench is saturated: What's next for OCR benchmarks.* https://www.llamaindex.ai/blog/omnidocbench-is-saturated-what-s-next-for-ocr-benchmarks

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). *Hallucination-free? Assessing the reliability of leading AI legal research tools.* Journal of Empirical Legal Studies. https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf

McKinsey & Company. (2025, March 12). *The state of AI: How organizations are rewiring to capture value* (survey fielded July 16–31, 2024, n=1,491; consultancy self-report). https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

METR. (2025, July 10). *Measuring the impact of early-2025 AI on experienced open-source developer productivity.* https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/

METR. (2026, January 29). *Time Horizon 1.1.* https://metr.org/blog/2026-1-29-time-horizon-1-1/

METR. (2026, February 24). *Developer uplift update.* https://metr.org/blog/2026-02-24-uplift-update/

Microsoft. (2025). *Form 10-Q (fiscal Q1 2026), reflecting Microsoft's equity-method share of OpenAI's losses* [basis for the inferred roughly US$12 billion OpenAI quarterly loss]. https://www.microsoft.com/en-us/investor/

MIT Media Lab (Project NANDA). (2025, August). *The GenAI divide: State of AI in business 2025* [origin of the "95% of pilots" figure; methodologically critiqued]. https://nanda.media.mit.edu/

Modern Pathology. (2021). *Independent (Yale) assessment of Paige Prostate cancer-detection accuracy* [near 98% sensitivity]. https://www.nature.com/articles/s41379-021-00794-x

Music Business Worldwide. (2025–2026). *Coverage of UMG/WMG settlements with Udio and Suno and related litigation.* https://www.musicbusinessworldwide.com/

Mustahsan, et al. (2025). *Stochasticity in agentic evaluations: Quantifying inconsistency with intraclass correlation* (arXiv:2512.06710). https://arxiv.org/abs/2512.06710

National AI Centre (Australia). (2026). *AI adoption insights: December 2025 – February 2026.* https://www.ai.gov.au/news-and-insights/blog/ai-adoption-insights-december-2025-february-2026

NEJM AI. (2025). *Evaluation of ambient AI documentation (DAX) at Atrium Health* [concluded the leading product was unlikely to produce appreciable system-level productivity gains]. https://ai.nejm.org/

Nerq. (2026). *State of AI assets, Q1 2026* [self-published MCP-server census (17,468 servers); exceeds PulseMCP and the official MCP Registry; the Linux Foundation's independently reported "more than 10,000" remains the body's anchor]. https://dev.to/zarq-ai/

OmniAI. (2025–2026). *OCR benchmark: 1,000 real-world documents.* https://getomni.ai/blog/ocr-benchmark

OpenAI. (2026, February 23). *Why SWE-bench Verified no longer measures frontier coding capabilities.* https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/

OpenAI. (2026, April). *Introducing GPT-5.5* [one-million-token context; $5/$30 per million tokens]. https://openai.com/index/introducing-gpt-5-5/

OpenAI. (2026). *What to know about the Sora discontinuation* [the 26 Apr 2026 app close and 24 Sep 2026 API sunset are corroborated by The Decoder]. https://help.openai.com/en/articles/20001152-what-to-know-about-the-sora-discontinuation

Ord, T. (2025). *Is there a half-life for the success rates of AI agents?* (arXiv:2505.05115). https://arxiv.org/pdf/2505.05115

PwC. (2025). *AI agent survey* [79% of organisations had implemented AI agents "at some level"; respondents projected an average ROI of 171% (192% for US enterprises); consultancy self-report, projected not audited]. https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html

RACGP. (2025–2026). *Artificial intelligence scribes: Position and fact sheet.* https://www.racgp.org.au/running-a-practice/technology/artificial-intelligence-ai/artificial-intelligence-ai-scribes

RICS. (2026, March 9). *Responsible use of artificial intelligence* [mandatory professional standard for regulated surveyors; flags AVMs and defect detection as high-risk]. https://www.rics.org/

Sacra. (2026). *Lovable revenue and growth research; product and engineering leader interview on Replit churn and retention.* https://sacra.com/c/lovable/

Salesforce. (2025). *State of service report 2025* [vendor survey; supports consumer preference for AI on simple tasks, humans for complex matters]. https://www.salesforce.com/news/stories/state-of-service-report-announcement-2025/

Salesforce Ben. (2025, April 7). *Why Agentforce adoption is slower than expected, and what Salesforce needs to do* (independent ecosystem publication; supports ~3,000 paid customers). https://www.salesforceben.com/why-agentforce-adoption-is-slower-than-expected-and-what-salesforce-needs-to-do/

Scale AI. (2026). *SWE-bench Pro leaderboard* [contamination-resistant coding benchmark; best models near 80%]. https://scale.com/leaderboard

Security.org. (2025). *Tea app data breach analysis.* https://www.security.org/identity-theft/breach/tea-app/

SemiAnalysis. (2026, February 5). *Claude Code is the inflection point* [commit-share estimate; later cited by METR]. https://newsletter.semianalysis.com/p/claude-code-is-the-inflection-point

Stack Overflow. (2025). *2025 developer survey*. https://stackoverflow.co/company/press/archive/stack-overflow-2025-developer-survey/

Stanford Digital Economy Lab (Brynjolfsson, E., Chandar, B., et al.). (2025, August). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence.* https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/

Stanford HAI. (2026). *AI Index report 2026.* https://hai.stanford.edu/ai-index/2026-ai-index-report

STAT News. (2026, April 1). *AI ambient scribes deliver modest time savings in clinical documentation.* https://www.statnews.com/2026/04/01/ai-ambient-scribes-modest-time-savings-clinical-documentation/

Supreme Court of New South Wales. (2025). *Practice Note SC Gen 23: Use of generative artificial intelligence.* https://supremecourt.nsw.gov.au/content/dcj/ctsd/supreme-court/supreme-court-home/practice-procedure/generative-artificial-intelligence.html

TechCrunch. (2026, June 3). *Still facing copyright lawsuits, AI music generator Suno raises another $400M.* https://techcrunch.com/2026/06/03/still-facing-copyright-lawsuits-ai-music-generator-suno-raises-another-400m/

TechCrunch. (2026, June 9). *Anthropic released Claude Fable 5, its most powerful public model, days after warning AI is getting too dangerous.* https://techcrunch.com/2026/06/09/anthropic-released-claude-fable-5-its-most-powerful-model-publicly-days-after-warning-ai-is-getting-too-dangerous/

The Decoder. (2026, April). *OpenAI unveils GPT-5.5, claims a "new class of intelligence" at double the API price.* https://the-decoder.com/openai-unveils-gpt-5-5-claims-a-new-class-of-intelligence-at-double-the-api-price/

The Decoder. (2026). *OpenAI sets two-stage Sora shutdown.* https://the-decoder.com/openai-sets-two-stage-sora-shutdown-with-app-closing-april-2026-and-api-following-in-september/

The Register. (2025, June 29). *AI agents fail a lot* [coverage of CMU TheAgentCompany]. https://www.theregister.com/2025/06/29/ai_agents_fail_a_lot

The Register. (2025, October 28). *AI browsers and prompt injection.* https://www.theregister.com/2025/10/28/ai_browsers_prompt_injection/

TheNextWeb. (2026). *Cursor in talks at ~$50 billion valuation; Lovable security crisis reporting.* https://thenextweb.com/news/cursor-anysphere-2-billion-funding-50-billion-valuation-ai-coding

Therapeutic Goods Administration (Australia). (2026, January 30). *Clarification: AI scribes that interpret clinical conversations are regulated medical devices requiring ARTG inclusion.* https://www.tga.gov.au/

VentureBeat. (2025–2026). *Xero's JAX shows why accuracy and user control matter more than flashy AI.* https://venturebeat.com/ai/xeros-jax-shows-why-accuracy-and-user-control-matter-more-than-flashy-ai

VentureBeat. (2026). *Anthropic hits $30 billion revenue run rate; Claude Code at $2.5B annualized.* https://venturebeat.com/technology/anthropic-says-it-hit-a-30-billion-revenu

e-run-rate-after-crazy-80x-growth

Visa. (2026, April). *Visa accelerates stablecoin momentum* [investor relations; ~US$7 billion annualised stablecoin-settlement run-rate across nine blockchains]. https://investor.visa.com/

Wang, Yu, Song, Lin, & Shen. (2025). *When retrieval succeeds and fails: Rethinking retrieval-augmented generation for LLMs* [review/survey preprint, arXiv:2510.09106]. https://arxiv.org/abs/2510.09106

Willison, S. (2026, May 28). *Claude Opus 4.8* [independent assessment]. https://simonwillison.net/2026/May/28/claude-opus-4-8/

Xero. (2025–2026). *About auto bank reconciliation powered by JAX* (open beta; Xero Grow in AU/NZ/UK, Growing in the US, Standard and above globally; matches or creates transactions only at high confidence; vendor source, confidence-gating design independently corroborated by VentureBeat). https://central.xero.com/s/article/About-auto-bank-reconciliation-powered-by-JAX

XLANG Lab. (2026). *OSWorld and OSWorld-Verified leaderboards* [computer-use agent benchmark; frontier ~70% standard, ~78–80% Verified, June 2026]. https://xlang.ai/blog/osworld-verified

Yahoo Finance. (2026). *Goldman Sachs' David Solomon on whether AI capital will produce returns* [Economic Club of Washington, D.C. remarks]. https://finance.yahoo.com/news/goldman-sachs-david-solomon-whether-164633542.html

YouTube. (2025, July 15). *Channel monetization policies: Inauthentic content update* [renamed the "repetitious content" policy; corroborated by Social Media Today]. https://support.google.com/youtube/answer/1311392

Zhang, Z. (2025). *Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice* [Comment]. npj Digital Medicine. (Commentary; the 86%-omissions and accent-disparity figures originate in the primary validation studies it cites.) https://www.nature.com/articles/s41746-025-01895-6

A further set of secondary and aggregator sources consulted during research, and the basis for treating them cautiously, is described in Appendices A and B.

Appendix A: Methodology and Verification Approach

Research process. This review was produced through structured web research conducted on 13 June 2026, organised as six parallel research streams covering the ten capability categories plus the temporal and macroeconomic frame. Each stream ran 12 to 22 distinct searches and, where a claim was load-bearing, fetched the primary page directly (for example, Higgsfield's MCP documentation and METR's uplift update were fetched in full rather than relied on through summaries).

Source hierarchy. Sources were weighted in the following order: (1) peer-reviewed studies and randomised trials; (2) independent evaluation organisations with published methodology (METR, Epoch AI, Stanford RegLab and HAI, DORA, Arc Prize Foundation); (3) government and regulator publications (UK DBT trial, AHPRA, ASIC, TGA, court practice notes, federal registers); (4) on-chain or otherwise externally auditable data (Dune Analytics, DefiLlama); (5) reputable financial and technology press (Bloomberg, CNBC, Reuters, TechCrunch, Fortune, The Register); (6) vendor documentation and announcements, used for what a product is and claims, never for how well it works; (7) aggregator and SEO content, used only to locate primary sources, and flagged wherever a figure could not be traced past it.

Vendor-claim labelling. Any number originating with the party selling the product (benchmark scores, ARR, resolution rates, time savings, accuracy percentages) is identified as a vendor claim in the text. Where independent evidence contradicts a vendor claim, both are reported with the discrepancy stated (for example, Intercom Fin's claimed 65 to 70 per cent resolution against 42 to 53 per cent in its own case studies; Bittensor's claimed quarterly revenue against an independent estimate more than ten times lower).

Date discipline. Every capability claim carries a date, either of the release, the study, or the observation. The paper should be read as a photograph of June 2026, not a standing description.

Known limitations. (1) The research could not access paywalled analyst reports or private deployment data. (2) Independent evaluation lags product release by six to twelve months, so the newest products (Claude Fable 5, GPT-5.5, Gemini 3.1 Pro) are assessed substantially on vendor evidence, with that stated. (3) Search results in 2026 are heavily polluted by machine-generated SEO content; several specific "facts" encountered during research exist only on such sites and were excluded (see Appendix B). (4) The authors used AI systems in the research and drafting of this paper, with human direction of scope, structure, and judgement; this is disclosed in the spirit of the disclosure standards the

paper itself discusses. (5) Some search-derived figures rest on result summaries rather than full primary documents; where this is true of a load-bearing figure it appears in Appendix B.

Verification approach. Load-bearing claims were checked across at least two independent source families where possible. Claims that survived on only one source family are either attributed explicitly to that source in the text or listed in Appendix B. No claim sourced solely from an SEO aggregator appears in the body of this paper as fact.

Appendix B: What Could Not Be Verified at Writing Time

The following claims were encountered during research, are relevant to the paper's subject matter, and could not be verified to this paper's standard as of 13 June 2026. They are listed with the specific reason. Their presence here is not an assertion that they are false.

This appendix was re-swept during the v1.0-to-v1.1 (Pass 2) fact-check. Twelve items whose primary sources were located in Pass 2 were promoted into the body and removed from this list; three items were refuted; the remainder were retained with sharpened reasons. The renumbering below reflects those removals. The Pass 2 corrections log records every move.

Frontier models

1. *Parameter counts for any current Claude, GPT-5.x, or Gemini 3.x model.* None disclosed by any vendor; no credible leak located. Circulating "5T/10T" figures trace only to SEO content farms and do not correspond to any real release. All "trillion-parameter" language in this paper refers to open-weight models with published architectures.
2. *The existence of "GPT-5.6" and "Gemini 3.2 Pro".* Split on re-check: "Gemini 3.2 Pro" does not exist (Google's line runs Gemini 3 Pro to 3.1 Pro to the 3.5 series), while "GPT-5.6" is unannounced but anticipated as of 13 June 2026 (a Codex-log mention, no official announcement), not an SEO fabrication. Both reinforce that no lab had formally announced its next flagship by mid-June 2026.
3. *Whether Meta's Llama 4 Behemoth ever shipped publicly.* Resolved in Pass 2: not shipped. Independent 2026 retrospectives agree it remained an internal teacher model; Meta's first proprietary frontier release was the closed-weight Muse Spark (around 8 April 2026). Body §3.1 is correct.
4. *Anthropic's claimed cyber capabilities for Mythos 5 / Fable 5.* Vendor claims with significant security-press coverage (The Hacker News, CSO Online), but all capability evidence is vendor-internal; no independent replication exists. Body §3.1 already applies the vendor-claim caveat.
5. *"Claude Code writes 4% of public GitHub commits."* Traces cleanly to its primary source (SemiAnalysis, who frame 4% as a floor) via the *Co-authored-by: Claude* signature, public repositories only. As of 13 June 2026 no independent replication has appeared; all downstream mentions re-report SemiAnalysis. Used in the paper with that

caveat.

Agents and infrastructure

6. *"97 million monthly MCP SDK downloads."* Untraceable to any primary registry export; official sources cite "more than 10,000 active public servers" (December 2025), not a download figure. The body uses the server count rather than this number.
7. *Nerq "17,468 public MCP servers, Q1 2026".* Now traced to a self-published Nerq census republished via dev.to, not an independent index; it exceeds PulseMCP (~5,500) and the official MCP Registry (~2,000). The body retains the Linux Foundation's "more than 10,000" (December 2025) as its anchor.
8. *A 2026 rerun of CMU's TheAgentCompany with current frontier models.* Confirmed not to exist; 2026 mentions restate the early-2025 Gemini 2.5 Pro result (~30%), and METR's task-horizon work is adjacent rather than a rerun. The absence is reported in §3.2 as an evidence gap.
9. *OSWorld scores for current models.* Resolved in Pass 2 as a variant/date artefact: frontier results are roughly 70% on standard OSWorld and 78 to 80% on OSWorld-Verified (XLANG, June 2026), now cited with that caveat in §3.2; "high 30s" is outdated early-2025 data.

Vision and media

10. *The University of Florida claim that AI plan review compresses permit approvals from 30–60 days to 3–5 days.* Refuted in Pass 2: the UF primary source does not contain that figure; it says a plan taking "a few weeks" can be reviewed by AI "in as little as 30 minutes." The "30–60 to 3–5 days" framing is a commercial-blog embellishment.
11. *"AVM 2.0" achieving 94.2% of valuations within 5% of sale price.* No primary publication or named product located; the figure remains orphaned on a single secondary source. The body uses the verifiable comparator instead (CoreLogic, roughly 90% within 15%).
12. *Post-January-2026 Veo pricing and "Veo 3.1 Lite" details.* Google first-party sources checked; no confirmation of a "Veo 3.1 Lite" SKU or post-January-2026 per-second pricing (SEO blogs only).
13. *Sora 2's maximum clip length.* Resolved in Pass 2 as a timeline rather than a conflict: roughly 10 seconds at launch, raised to 15 seconds for all users and up to 25 seconds for Pro by mid-October 2025. The body's "4 to 15 seconds for most models" remains consistent.

14. *Economic claims attached to Higgsfield MCP pipelines* ("replaces a $5K/month retainer"). Confirmed to be Higgsfield's own marketing copy, an unaudited vendor claim; the MCP integration itself is first-party verified.

Voice and audio

15. *Nearly all aggregate voice-fraud statistics* (US$1.1B 2025 losses; 1,600% vishing growth; $680K average loss). Checked against FBI IC3 2025 and FTC data; the specific trio could not be sourced (independent vishing-growth figures are nearer 442%). The Arup Hong Kong case is separately verified (CNN: US$25.6M / HK$200M, May 2024) and is the only aggregate-adjacent figure the body treats as established.
16. *AI receptionist ROI statistics* ("62% of SMB calls unanswered"; "500–3,700% ROI"). Lineage confirmed as vendor-commissioned (Aira and other receptionist vendors); no primary research underpins either figure. §3.8 makes its case without the numbers.
17. *Reported 2025 ChatGPT memory-wipe incidents.* Partially corrected: OpenAI's status page does acknowledge a memory incident (6–7 November 2025), so "no OpenAI acknowledgment" was too strong; but the sweeping "silent February 2025 wipe / 83% MIT failure study" narrative remains SEO-only. Not load-bearing in the body.

Knowledge and enterprise

18. *The Pinecone 12-deployment enterprise RAG study* (31% out-of-distribution queries). Pinecone's site checked; no primary publication. The figure appears only on an SEO blog citing an uncorroborated event. §3.6 does not rely on it.
19. *Salesforce 2025 consumer-preference survey* (consumers favour AI for simple tasks, humans for complex or emotional matters). Substance corroborated by Salesforce's own State of Service Report 2025 and an agentic-AI personas piece; a single canonical "preference inverts" report is spread across multiple Salesforce publications. Vendor source.
20. *APQC "≈85% of organisations have not operationalised AI knowledge tools at scale".* Traced to a specific APQC study (survey of 1,000 professionals, sponsored by eGain, a knowledge-management vendor); the exact wording is "have not operationalized AI" (the paper's "AI knowledge tools at scale" is a slightly tighter gloss). The reference is pinned to the press release; the instrument was not independently examined.

Crypto-AI

21. *NEAR Agent Market transaction volumes, active agent counts, or GMV.* The marketplace is verified to exist (near.ai, market.near.ai; launched 9 May 2026), but no transaction, GMV, or active-agent data is published or independently indexed.

Consistent with §3.9.

22. *Tempo / Machine Payments Protocol transaction throughput.* Tempo mainnet (18 March 2026) and the Machine Payments Protocol are confirmed; the only throughput number ("100,000+ TPS") is a vendor capacity claim, not measured throughput, and no neutral dashboard was located.
23. *Bittensor's "$43M quarterly AI revenue".* Both sides now verified: the $43M Q1-2026 figure (largely the Chutes subnet) circulates in crypto press, and an independent March 2026 analysis estimates genuine external revenue at $3–15M annually (a 22:1–40:1 emission-subsidy ratio). The dispute is real and §3.9 frames it correctly.

Verticals and macro

24. *Heidi Health Series C, and its ~$21.9M ARR estimate.* No Series C as of June 2026 (latest is the October 2025 Series B; confirmed by Sacra and PitchBook); third-party commentary cites only a roughly US$25M 2026 ARR target, not the "$21.9M" figure, which could not be located. Body §3.10 Series B statement is correct.
25. *Independent accuracy benchmarks for Xero JAX, Intuit Assist, Dext, or ServiceM8 AI features.* Confirmed absence: no independent, methodologically transparent benchmark exists; all accuracy figures are vendor or SEO claims. Supports §3.10.
26. *Spotlight Reporting's 2026 AI feature set.* No datable 2026 AI feature set located; correctly excluded from the body.
27. *Specific 2026 generative-AI consumer features from REA Group or Domain.* Trade press reports REA has shipped agent-facing AI features (imprecisely dated), but no specifically dated 2026 consumer-facing generative feature, and none for Domain, was verified.
28. *OpenAI's reported ~$12B quarterly loss and ~$143B projected cumulative cash burn.* Refined: the ~$12B is inferred from Microsoft's SEC filings (its share of OpenAI's losses, Q3 2025) and the ~$143B is a Deutsche Bank cumulative-free-cash-flow projection (2024–2029), both independent of OpenAI and neither OpenAI-audited. §4 now carries that attribution.
29. *Australian National AI Centre SMB adoption figures (40% to 69% regular use; 9% to 28% daily use).* Figures confirmed against the NAIC's December 2025 to February 2026 summary; the remaining caveat is narrow: the underlying survey instrument and sample design were not published or examined.
30. *Deloitte State of AI in the Enterprise 2026 "15% with significant measurable ROI" figure.* The sample (3,235 senior leaders, August to September 2025, 24 countries) is confirmed and now stated in §2.3; the specific 15% ROI figure could not be located in

any accessible source and remains publisher-reported. Consultancy self-report.

Corrections Log (v0.9 to v1.2 fact-check phase)

The pre-fact-check draft of this paper is preserved at *archive/state-of-ai-mid-2026-v0.9-pre-fact-check.md*. The corrections below trace every change applied during the v1.0 fact-check pass.

Pass 1 corrections

Total findings processed: 78 (Findings #1–77, one per References entry, plus Finding #78, the body-wide inline-citation audit). Of the 77 reference findings, 36 were VERIFIED and required no change; the remaining 41, plus the inline-citation audit, drove the corrections below.

Major corrections (4):

1. *Finding #33 (UK Copilot)*: the cited PDF is the Government Digital Service cross-government report (~20,000 employees, self-reported +26 min/day, high satisfaction, not a randomised controlled trial), not a ~1,000-person DBT trial showing no productivity gain. The claim was rewritten to match the source across the Abstract, §3.8, and Conclusion, and the "only randomised government trial" framing was removed.

2. *Finding #64 (McKinsey)*: the cited URL is the March 2025 report (July 2024 survey: 78% use AI, 71% regular generative-AI use, >80% no enterprise EBIT impact, 1% mature); it does not contain the body's "88% regularly / ~6% with 5%+ earnings impact." §2.3 and §4 were revised to the figures the source supports.

3. *Finding #43 (RAG arXiv)*: arXiv:2510.09106 is a RAG review/survey, not an empirical study, and does not contain the "47–67% of queries" ignore rate. The specific figure was removed and the citation reclassified as a survey.

4. *Findings #8/#68 (CMU TheAgentCompany)*: the body's 30.3% is the Gemini 2.5 Pro result reported by The Register (29 June 2025); the CMU news page reports 24% for an earlier run. §3.2 was re-sourced to The Register and the 24% figure noted.

Material citation corrections: 8. npj Digital Medicine (commentary, not the named validation study), Stanford "Canaries" (URL repointed), Innolitics (reports 295 in 2025, not the cumulative ~1,250), Salesforce Ben (April 2025, not 2026), Lovable/Sacra (estimate ~US$400M), Lovable RLS conflation reworded, the Fortune prompt-injection quote re-quoted verbatim, and the IrisAgent containment figure tagged as vendor-authored.

Mechanical fixes: 5. EU AI Act date corrected from 7 May to 6 May; GPT-5.5 context corrected from "1.05-million-token" to "one-million-token"; SemiAnalysis re-dated February 2026; Adviser Ratings re-dated May 2025; Instant Checkout sunset softened to "late March / early April 2026."

Unverifiable items handled (12): kept in the body with added caveats (9); flagged into Appendix B with 10 new items; removed outright: 0.

Items requiring further user judgement (TBDs): 0.

Detailed entries: major corrections

Correction 1: Finding #33 (UK Copilot), Abstract / §3.8 / Conclusion.
Original (Abstract): "the only randomised government trial of Microsoft 365 Copilot finding no measurable productivity improvement."
Original (§3.8): "it is the only government-grade trial in the category: the UK Department for Business and Trade gave roughly 1,000 civil servants Copilot for three months (late 2024, published 2025) and found no conclusive evidence of measurable productivity improvement…"
Changed to: the claim now describes the Government Digital Service cross-government evaluation (~20,000 civil servants, three months, self-reported time savings of around 26 minutes per day, high satisfaction, not a randomised controlled trial, no productivity gain established under controlled measurement).

Correction 2: Finding #64 (McKinsey), §2.3 and §4.
Original (§2.3): "McKinsey's 2025–26 State of AI found 88 per cent of organisations using AI regularly but only around 6 per cent achieving enterprise-wide earnings impact of 5 per cent or more."
Changed to: §2.3 now reports 78% using AI in at least one function, 71% regularly using generative AI, more than 80% reporting no tangible enterprise-level earnings impact, and 1% describing rollouts as mature, labelled a consultancy self-report.

Correction 3: Finding #43 (RAG), §3.6.
Original: "studies find generators ignoring the retrieved documents in 47 to 67 per cent of queries (arXiv 2510.09106, 2025)."
Changed to: the specific figure was removed; the citation was reclassified as a review/survey. Status: misattributed figure removed; qualitative claim retained and correctly sourced.

Correction 4: Findings #8/#68 (CMU TheAgentCompany), §3.2.
Original: "the best agent completed 30.3 per cent."
Changed to: "in the run reported by The Register (29 June 2025), the best agent (Gemini 2.5 Pro) completed 30.3 per cent of tasks autonomously… (The CMU news write-up of an earlier run reported 24 per cent for the best agent then tested.)"

Detailed entries: material citation corrections

- §3.10 / §3.5, npj Digital Medicine (#45): original cited "Validation of ambient AI clinical documentation: Error taxonomy and demographic disparities." Changed to "Zhang, Z. (2025). Beyond human ears… [Comment], npj Digital Medicine," with the body re-worded to attribute the 86%-omissions and accent-disparity figures to the primary studies the commentary cites.

- §4, Stanford “Canaries” (#39): reference URL repointed from the October 2025 commentary to the August 2025 publication page. Status: now verified (figure was already accurate).
- §3.3, Innolitics (#58): original attributed “About 1,250 AI-enabled devices… by August 2025” to Innolitics. Changed to attribute the cumulative >1,250 figure (by ~July 2025) to the FDA running list (corroborated by the Bipartisan Policy Center), noting Innolitics recorded 295 clearances during 2025.
- §3.8, Salesforce Ben (#67): reference re-dated to April 2025; body rewritten to “slow adoption, CFO guidance to ‘modest’ sales, ~3,000 paid customers.”
- §3.7, Lovable/Sacra (#65): “Lovable reported around US$400 to 500 million” changed to “estimated at around US$400 million as of February 2026 (Sacra estimate).”
- §3.7, Lovable RLS (#70): reworded to “~10% (170/1,645) exposed personal data and ~70% had row-level security disabled (May 2025 study)” to avoid conflating the two findings.
- §3.2, Fortune prompt-injection quote (#25): “remains a frontier, unsolved security problem” re-quoted to the verbatim “unlikely to ever be fully ‘solved’.”
- §3.5, IrisAgent (#56): the ~50% containment contrast now tags IrisAgent as itself a voice-AI vendor.

Detailed entries: date and metadata corrections

- §2.4: EU AI Act provisional agreement date corrected from “7 May 2026” to “6 May 2026 (Council confirmation 13 May)” (#18).
- §3.1: GPT-5.5 context window corrected from “1.05-million-token” to “one-million-token” (#48).
- §3.7 + reference: SemiAnalysis re-dated from “May 2026” to “February 2026 (later cited by METR)” (#41).
- §3.10 + reference: Adviser Ratings re-dated from “October 2025” to “May 2025”; “74% using” qualified as “using or planning to use”; the unsourced “39% compliant-use policies” figure removed (#59).
- §3.9: Instant Checkout sunset softened from “24 March 2026 with only around 30 merchants” to “late March / early April 2026, with very few merchants,” pending confirmation. (Superseded by the Pass 2 promotion, which restored the exact date and the roughly 30-merchant count.)

Detailed entries: orphan inline-citation resolutions (Finding #78)

References entries were created for the following previously orphaned inline citations: Demand Sage (§2.2), Futuriom (§2.3), arXiv:2512.06710 (§3.2), Scale AI / SWE-bench Pro (§3.1), Dermatology Times (§3.3), British Journal of Dermatology (§3.3), Modern Pathology / Paige Prostate (§3.3), Futurum Group (§3.4), the Salesforce 2025 consumer survey (§3.5), APQC (§3.6), and DefiLlama (§3.9). An Epoch AI MMLU-Pro entry and an OpenAI February 2026 SWE-bench contamination entry were added for §3.1. The Anthropic Economic Index (March 2026) entry was retained as a methodology-list source. Where a precise URL could not be pinned (Salesforce survey, APQC, Brave, Nerq, Escape.tech), the entry carries an explicit caveat and a cross-reference to Appendix B.

Pass 2 corrections

Total Pass 2 findings processed: 48 (Findings #79–#126): four mechanical "to be pinned" reference items plus a substantive re-sweep of all 44 Appendix B entries. Distribution: 18 VERIFIED (12 of them promotions into the body), 15 PARTIALLY VERIFIED, 3 REFUTED, 10 UNVERIFIABLE, and 2 split items.

Items promoted from Appendix B into the body (12): Amazon v. Perplexity injunction into §3.2; NotebookLM "Cinematic Video Overviews" into §3.6; Visa ~US$7B stablecoin-settlement run-rate into §3.9; the Solomon "will not produce returns" quote into §4; the AI Index 2026 "60% to near 100% SWE-bench" figure into §3.7; the Escape.tech primary report into §3.7; the Instant Checkout exact date, ~30-merchant count, and sales-tax detail restored in §3.9; Xero JAX specifics de-hedged in §3.10; NEAR Intents volumes de-hedged in §3.9; the YouTube "inauthentic content" policy pinned for §3.4; the PwC 79%/171% survey into §3.8; and the Brave Comet prompt-injection post pinned into the References for §3.2.

Material citation corrections from Pass 2 (5): the §3.2 Mustahsan paraphrase softened; the OpenAI loss figures in §4 re-attributed to Microsoft's SEC filings (~US$12B quarter) and a Deutsche Bank projection (~US$143B cumulative); the §2.3 McKinsey/Deloitte block updated with Deloitte's confirmed sample (n=3,235); a variant-specific OSWorld range added to §3.2 (~70% standard, ~78–80% Verified, XLANG, June 2026); and the Escape.tech figure re-sourced to the primary report with the 175 personal-data exposures added.

Refutations (3): the "2026 Forrester Wave scoring Notion AI 94/100" (Forrester uses a 0–5 scale and publishes no Notion-AI Wave; a single-SEO-source fabrication); "Whisper-V4" (no such product; the model is OpenAI's GPT-Realtime-Whisper); and the University of Florida "30–60 days to 3–5 days" plan-review figure.

Mechanical pins resolved (4): Brave reference pinned to brave.com/blog/comet-prompt-injection/; the OpenAI SWE-bench contamination note pinned; arXiv:2509.25498 corrected to "Not Wrong, But Untrue: LLM Overconfidence in Document-Based Queries"; and arXiv:2512.06710 confirmed as "Stochasticity in Agentic Evaluations."

§3.2 paraphrase softening: the claim that "single-run benchmark scores have, in effect, stopped carrying information" overstated its source (Mustahsan et al.); the sentence was softened to "single-run benchmark scores cannot reliably distinguish genuine capability gains from sampling noise."

Detailed entries: promotions

- §3.2, Amazon v. Perplexity (item 7 / Finding #89): replaced “case details could not be fully verified” with the verified specifics, Senior US District Judge Maxine M. Chesney (N.D. Cal.), preliminary injunction 10 March 2026 on Computer Fraud and Abuse Act grounds with a destruction order, and a Ninth Circuit temporary administrative stay on 16 March 2026.
- §3.6, NotebookLM Cinematic Video Overviews (item 19 / Finding #101): added as a verified first-party Google feature (March 2026; Gemini 3, Nano Banana Pro, Veo 3; AI Ultra; up to 20 per day).
- §3.9, Visa stablecoin run-rate (item 25 / Finding #107): added the roughly US$7B annualised stablecoin-settlement run-rate across nine blockchains (Visa IR; CoinDesk, 29 April 2026), explicitly distinguished from agentic-payment volume.
- §4, Solomon quote (item 32 / Finding #114): added the on-the-record Economic Club of Washington, D.C. quote (Banking Dive; Yahoo Finance).
- §3.7, AI Index SWE-bench figure (item 33 / Finding #115): “roughly 40 per cent to the mid-90s” replaced with the AI Index 2026 verbatim “60 per cent to near 100 per cent in a single year,” paired with the existing contamination caveat.
- §3.7, Escape.tech (item 36 / Finding #118): re-sourced to the primary report “The State of Security of Vibe Coded Apps” with the 175 personal-data exposures added.
- §3.9, Instant Checkout (item 41 / Finding #123): restored “24 March 2026,” “~30 Shopify merchants,” and the February 2026 US-sales-tax gap.
- §3.10, Xero JAX (item 42 / Finding #124): hedge removed; plan names added and the high-confidence-only / manual-review-of-low-confidence behaviour stated.
- §3.9, NEAR Intents (item 43 / Finding #125): “could not be machine-verified” hedge relaxed; the ~US$10B and ~US$20B milestones are corroborated by independent crypto press.
- §3.4 / References, YouTube policy (item 44 / Finding #126): reference repointed to the YouTube Help Center monetization-policies page (15 July 2025).
- §3.8, PwC survey (item 22 / Finding #104): “trace to surveys with unpublished methodology and should be ignored” replaced with the PwC 2025 AI Agent Survey figures (79% implemented “at some level,” 171% projected ROI, 192% US), framed as self-reported projections.

Detailed entries: refutations and splits

- Appendix B, Forrester Wave (old item 21 / Finding #103): removed. Forrester Waves use a 0–5 scale and tier classifications, never a “/100,” and no Forrester Wave covers Notion AI; the “94/100” is a single-SEO-source fabrication.

- Appendix B, “Whisper-V4” (old item 15 / Finding #97): removed. No such product exists; the model is OpenAI’s GPT-Realtime-Whisper (May 2026). The companion GPT-Realtime-2 was confirmed first-party.
- Appendix B, UF plan-review figure (old item 10 / Finding #92): the specific “30–60 days to 3–5 days” figure was refuted; the UF source says “weeks to 30 minutes” per plan. The UF/AutoReview.AI program is real.
- Appendix B, voice-fraud and Arup (old item 16 / Finding #98): split. The Arup case is verified (US$25.6M / HK$200M, May 2024, CNN); the aggregate trio ($1.1B / 1,600% / $680K) remained unverifiable.

Detailed entries: Appendix B reason refinements (UNVERIFIABLE and PARTIALLY VERIFIED retained)

The remaining Appendix B entries were retained with sharpened reasons reflecting the Pass 2 search: parameter counts (SEO “5T/10T” fabrications noted); GPT-5.6 / Gemini 3.2 Pro (split: Gemini 3.2 Pro non-existent, GPT-5.6 unannounced-but-anticipated); Llama 4 Behemoth (resolved: not shipped; Muse Spark was Meta’s first proprietary frontier release); Mythos/Fable 5 cyber claims (vendor, no replication); Claude Code 4% commits (primary-sourced to SemiAnalysis as a floor, no independent replication); the 97M MCP downloads and the Nerq 17,468 census (self-published, below the Linux Foundation anchor); TheAgentCompany rerun (confirmed absent); OSWorld (range now in §3.2); AVM 2.0 94.2% (untraceable); Veo 3.1 Lite (SEO-only); Sora clip length (resolved as a timeline); Higgsfield ROI copy (vendor marketing); AI receptionist ROI (Aira lineage); ChatGPT memory (a November 2025 incident is acknowledged first-party; the sweeping narrative is SEO-only); Pinecone RAG study (SEO-only); Salesforce and APQC surveys (substance corroborated, references pinned, sponsorship flagged for APQC/eGain); NEAR Agent Market (exists, no metrics); Tempo throughput (vendor capacity claim); Bittensor (both sides of the dispute verified); Heidi Series C (none; ~$25M target, not $21.9M); vertical-AI accuracy benchmarks (confirmed absent); Spotlight Reporting (none located); REA/Domain (REA agent-facing only); OpenAI losses (re-attributed to Microsoft filings and a Deutsche Bank projection); NAIC SMB figures (figures confirmed, instrument caveat narrowed); and Deloitte (n=3,235 confirmed, 15% ROI still publisher-reported).

Pass 3 corrections

> **Total Pass 3 findings processed:** 9 (Findings #127–#135), a final top-to-bottom read-through. Distribution: 8 PARTIALLY VERIFIED, 1 UNVERIFIABLE; no refutations and no load-bearing factual errors. All nine were consistency, drift, or citation-housekeeping items.
>
> **Drift corrections applied (3):** the Abstract's CMU attribution aligned with the corrected §3.2 sourcing; the §4 McKinsey/Deloitte parenthetical rephrased for consistent polarity; and §6/§4 framing tightened ("a reported legal injunction" to "a court injunction"; "UK Copilot trial" to "UK Copilot study").
>
> **Housekeeping fixes (3):** the References section re-alphabetised after the Pass 2 insertions; the Pass 1 Instant Checkout log entry annotated with a forward-reference to its Pass 2 restoration; and the Corrections Log heading updated to span v0.9 to v1.2.
>
> **New References entries created (9):** for the §4 OpenAI loss figures, Microsoft (Form 10-Q) and eMarketer (Deutsche Bank projection); and for previously in-text-only institutional sources, MIT Media Lab (Project NANDA), JAMA Network Open, NEJM AI, the TGA, RICS, the Civil Resolution Tribunal of British Columbia (Moffatt v. Air Canada), and the Florida Legislature (HB 683).
>
> **Anthropic Economic Index (#135):** the References entry had no inline anchor in the body and was removed as uncited rather than padded with a manufactured citation.
>
> **Body claims revised on the basis of Pass 3:** none. Every Pass 3 edit was a consistency, framing, or citation-list change; no factual claim was added, removed, or re-sourced.

Detailed entries

- Abstract (#127): "(Carnegie Mellon, 2025)" changed to "(Carnegie Mellon's TheAgentCompany; best-agent result reported by The Register, 2025)" so the abstract matches the body's careful sourcing of the 30.3 per cent figure to The Register (with CMU's own page at 24 per cent).
- §4 (#128): "(McKinsey's more than 80 per cent reporting no tangible enterprise-level impact; Deloitte's 15 per cent reporting significant ROI)" rephrased to "(McKinsey finding more than 80 per cent with no tangible enterprise-level impact, and Deloitte only 15 per cent reporting significant ROI)" to remove the mixed-polarity construction.
- §6 and §4 (#129): "a reported legal injunction against an agentic browser" changed to "a court injunction" to match the now-verified §3.2 detail; "government pilots like the UK Copilot trial" changed to "government evaluations like the UK Copilot study," consistent with §3.8's "not a randomised controlled trial."

- References (#130): re-alphabetised. The out-of-order Pass 2 insertions were sorted into place, with multi-entry authors ordered by date.
- Corrections Log (#131): a forward-reference was added to the Pass 1 Instant Checkout "softened" entry noting that Pass 2 restored the exact date and merchant count.
- Corrections Log (#132): heading updated from "(v0.9 to v1.0 fact-check pass)" to "(v0.9 to v1.2 fact-check phase)" to match the embedded Pass 2 and Pass 3 sections.
- §4 OpenAI losses (#133): References entries added for the Microsoft Form 10-Q (basis for the ~US$12B quarterly figure) and the eMarketer/Deutsche Bank coverage (the ~US$143B cumulative projection).
- In-text institutional citations (#134): formal References entries added for MIT NANDA, the JAMA and NEJM AI scribe studies, the TGA device-classification clarification, the RICS standard, Moffatt v. Air Canada (2024 BCCRT 149), and Florida HB 683 (2025).
- Anthropic Economic Index (#135): removed from References as an uncited orphan (no inline citation tied it to body text).

> *Perth AI Consulting publishes this review as a point-in-time resource. This is version 1.2, incorporating the Pass-1, Pass-2, and Pass-3 fact-check corrections logged above; the pre-fact-check draft is preserved at archive/state-of-ai-mid-2026-v0.9-pre-fact-check.md. Corrections, with sources, are welcome: the paper's method assumes some of it will need them.*