

Buying an AI agent team: the manager isn't the expensive part

Boardrooms are being sold a picture: a manager agent directing worker agents across a company's tools and data. The picture is roughly right. Almost everything expensive about the way it's being bought is not, and you can audit every number behind that sentence.

AUG 22, 2026 · REVISION 6 · LAB.WORTHINGTON.CLOUD

Buying an AI agent team: the manager isn't the expensive part

Boardrooms are being sold a picture: a manager agent directing worker agents across a company's tools and data. The picture is roughly right. Almost everything expensive about the way it's being bought is not, and you can audit every number behind that sentence.

002 Field Report Aug 22, 2026 Revision 6

Instrument and questions: Rick Worthington

Analysis and prose: Agent

The business problem, first

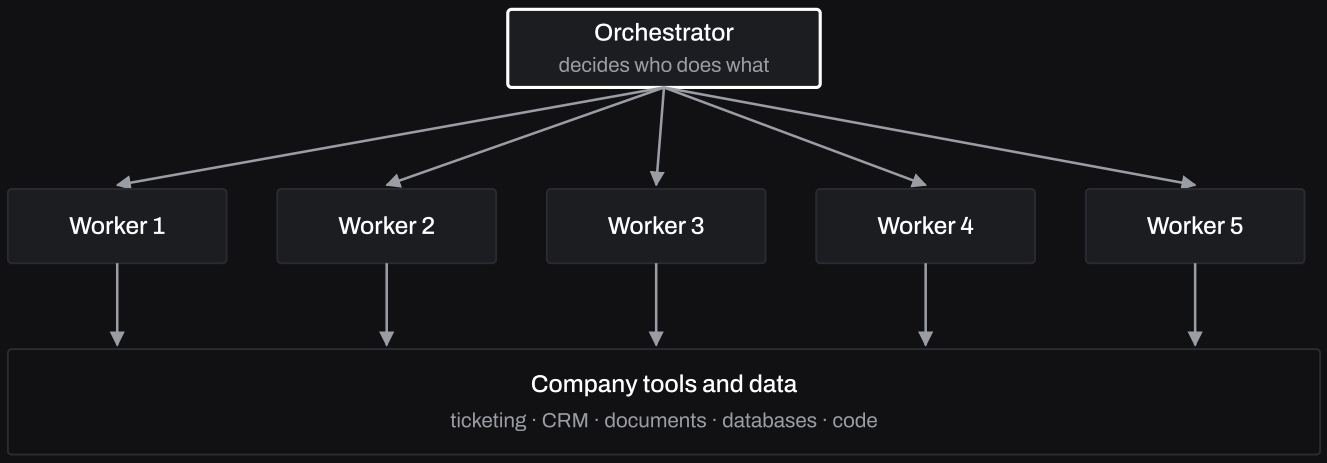
AI is not a strategy. It is an instrument for raising productivity, and raising productivity is how a company gets to whatever its actual strategy is.

That distinction sounds academic until it starts costing money — because a company that believes AI *is* the strategy will buy the most impressive version of it, and the most impressive version is not the one that pays.

The vision currently being sold to executives is a team of agents: one "orchestrator" directing several "workers," each sitting on top of company tools and data, doing the routine work of a department. An **agent**, plainly, is a program that uses a language model to do a multi-step job — read a ticket, look something up, make a change, check its own work — rather than answering one question and stopping. The **orchestrator** is the agent that decides which other agent or tool handles each step, and when the job is finished: the shift supervisor, not the smartest person on the floor.

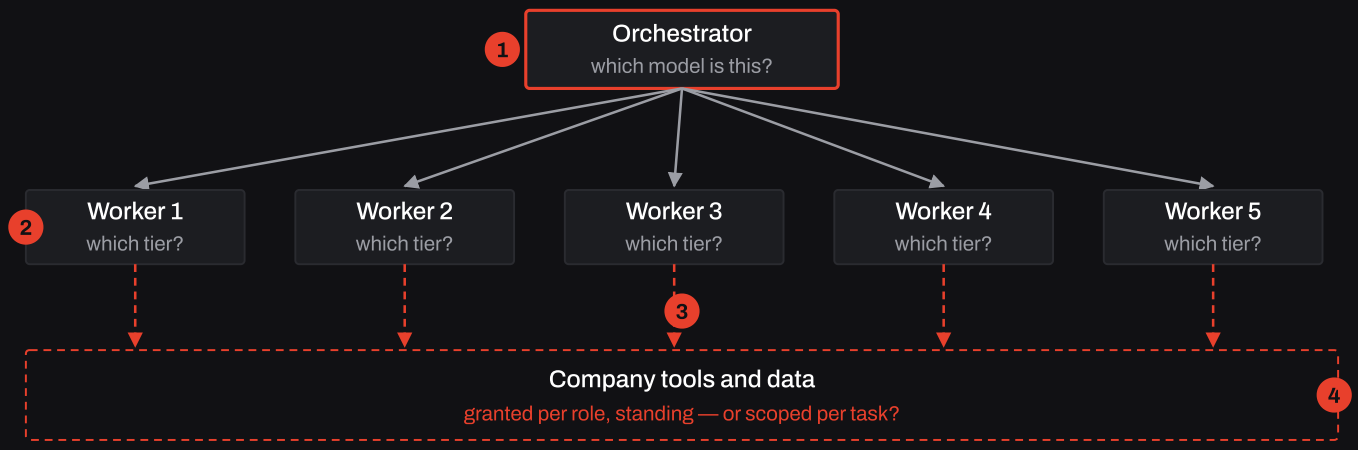
This report does not argue against that picture. The evidence broadly supports it. The question it asks is narrower and more useful: **if teams of agents really are the coming shape of work, what is the least expensive way to get there, and how would a company know whether it was working?**

Most executives have been shown a version of this slide, or drawn one:



The common picture — and it is not wrong. An orchestrator delegates to specialized workers, each with access to the systems it needs. Every diagram that follows keeps this skeleton.

Nothing about it is wrong. The difficulty is that the picture is silent on four decisions, and those four decisions are where essentially all of the cost, all of the compounding, and all of the exposure are determined. Same diagram, with the silences marked:



1. The supervisor's model — the most expensive default in the diagram *(F7)*
2. Each worker's model tier — where routing savings live, or do not *(F5)*
3. Where the work gets checked — nothing compounds without it *(F10)*
4. What each agent may reach, and when *(F15)*

Identical skeleton. The four marked decisions are the ones the picture leaves open — and each is where the cost, the compounding, and the exposure are actually determined.

A company can implement that first diagram four different ways and see a twenty-fold difference in running cost, with one version improving every month and another staying flat forever. The rest of this report is about those four decisions — the architecture the evidence points to is drawn in Conclusions. Four positions come out of the evidence.

1 **The supervisor premium is real, small, and priced out of all proportion to it.** When twelve models were placed in charge of the same workers, the most expensive one did win — and bought six points over a model costing a quarter as much, while three expensive entries were beaten outright by cheaper ones. Orchestration is a different skill from reasoning, it can be trained directly and cheaply, and the default assumption that the coordinator must be the most capable model available is the single most expensive unexamined line in the architecture.

Justified by: twelve supervisors run over one fixed worker pool, where cost spanned more than a hundredfold and management quality spanned under four (Finding 24); and

by orchestration being trainable into an 8B model for 6.5 points and two-thirds of the bill (Finding 7).

2 **The famous cost savings are real but come from a lever most buyers are not pulling.** Routing and owning hardware are two different levers with two different price tags. Pull the routing lever first — it is a software change against hosted models. The purchase decision and the savings decision are separable, and only one of them is urgent.

Justified by: the thousandfold figure being a marginal electricity cost that excludes ~$45,000 of hardware (Finding 4), while routing alone delivers a measured third off at matched quality with no capital at all (5, 6).

3 **The thing that makes a team of agents improve itself is a gate, not a model.** Before planning any self-improving loop, find the automatic pass/fail test for the work in question. Where no such test exists, the flywheel has no mechanism and the plan is a hope.

Justified by: the cleanest controlled experiment in the corpus — swap only the verification gate for a lenient one and the entire self-improvement gain vanishes; training on unfiltered output measured worse than not training at all (Finding 10).

4 **Nobody has measured the productivity gain, including this report.** Any vendor quoting such a figure is quoting their own marketing. The only credible number a buyer will ever have is a baseline captured before deployment — the step that is almost always skipped and cannot be recovered afterward.

Justified by: a corpus rich on cost, accuracy and architecture, and empty on organizational productivity (Finding 16).

The rest of this document justifies those four positions, one marked decision at a time. Every finding carries an evidence grade and a citation; what failed re-analysis is named in Removed & limits.

WHAT THIS REPORT IS NOT

It is not a build-versus-buy verdict, and it is not a claim that money is flowing anywhere. Corporate filings show a sharp rise in how companies *talk* about this technology; that is language, not capital, and the two are measured differently. Where the evidence runs out, the report says so rather than rounding up.

HOW THIS WAS MADE, AND WHAT FAILED

Twenty-two claims were written down before testing, then worked against 290,240 research papers, a purpose-built collection of practitioner posts, and public filings — and adversarially re-analyzed twice, with the second pass published inside this report. Of the 22 claims, one came back refuted and fourteen mixed; the second re-analysis replaced one finding's numbers outright and reframed four more. **Revision 4 then corrected this report's own headline reading** — not because anyone challenged it, but because re-reading a source already cited showed the claim had generalized from one model to all of them. Each is named in Removed & limits, and the full account is in Provenance.

Why you can trust this

Every claim lives on a ledger that keeps refuted claims instead of deleting them — this report was attacked by its own red team before it was published, twice, and it has since corrected its own headline without being asked to.

Hypotheses were written down *before* they were checked, together with the method that would settle them; derived claims are stored with the exact query behind them, so any number can be re-run by anyone with the corpus.[1] Of the 22 claims on the ledger, 11 were the author's own hypotheses — and most did not survive intact: 14 claims settled as *mixed*, one was outright refuted and became this report's headline finding, one proved untestable and says so.

290,240
research paper abstracts searched, January 2023 to August 2026

68
sources read at full length for this report, stored with content hashes — the curated shelf, counted apart from the bulk archive around it

29,287
corporate filing passages, 2023–2026, on a balanced company panel

18,620
social posts collected for one finding, because the existing corpus contained zero

22
claims on the ledger, methods recorded with them

15
claims that did not survive intact — mixed or refuted, all still on the board

Where a language model sits in the pipeline — exactly one place. The social-corpus composition (Finding 3) is produced by a model classifying posts. That step was calibrated, not trusted: the reference judge was measured against itself first (98% self-agreement on topic, 97% on speaker), candidates were scored against that ceiling, the cheap winner's weakest field was caught running 6–10 points hot and replaced with the reference model's stratified-sample levels, and an

> adversarial pass re-judged the same sample with a third model family and got the same answer.[5][6] Every other number is a deterministic query over stored data, or is quoted from a named source with its own methodology.

How it was tested

Listed whether they passed or not. A test that changed nothing is worth less than a test that killed something — and the two worst defects found were on this very tab.

ADVERSARIALLY RE-ANALYZED, SECOND PASS — A FRESH SESSION SENT TO DESTROY THIS REPORT Before publication, a red team with no stake in the draft attacked every finding: all 18 cited papers re-read at the methods level, every quoted number re-derived or traced to its source, the classifier re-judged with a third model family, the collection censoring tested for time correlation. **No finding died. Five were reframed:** the orchestrator comparison recut around the same-toolbox baseline (F7), the social-corpus levels re-anchored to the reference judge (F3), the decision-path cost ground narrowed to the frontier tier (F19), the bandit finding scoped to offline experiments (F20), and one finding's numbers replaced because their original query had never been persisted (F18).[6]

CAUGHT BY THAT RED TEAM: THE TRUST TILE ITSELF WAS WRONG The "sources read at full length" count was rendering the instrument's whole library — including 3,154 video transcripts belonging to a different sensor — instead of the ~65 documents actually read for this report. Provenance inflation of roughly fifty-fold, on the page whose only job is trust, live until the red team read the query behind the tile. Fixed by scoping the count; kept here because a trust page that has been wrong should say so.[6]

CAUGHT BY THAT RED TEAM: THE AI DISCLOSURE WAS FALSE An earlier revision of this tab said "no model scores, ranks or classifies any evidence." Untrue: the social-corpus composition is produced by a model classifying posts. The disclosure below now states what actually happens and how the classifier's error was measured, instead of denying it exists.[6]

KILLED: THE FLAGSHIP EXPERIMENT The plan was to rent a GPU and test whether filtering training data through a verification gate beats training on everything. A corpus search found the decisive version already published, *with controls that would not have been designed here* — including one that swapped only the gate and watched the entire gain vanish. The experiment was canceled rather than run, and the finding now cites someone else.[14]

KILLED · REV 4: A SECOND EXPERIMENT, FOR THE SAME REASON AS THE FIRST A bake-off was designed to settle this report's own headline — swap only the supervisor across four price tiers over a fixed worker pool, on a battery of knowledge-work tasks, with a gate per task and cost measured per passing result. It was budgeted and about to be built. A corpus and literature sweep run before spending found the controlled version already published by an academic group, with twelve models instead of four, execution-based scoring instead of a model judge, and a design detail nobody here had thought of: crippling the supervisor's own perception so that delegating is forced rather than optional.[30] The experiment was canceled and F24 cites someone else. **This is the second time on this report that searching first replaced building** — and the first time the thing it replaced was an experiment intended to test the report's own conclusion.

CAUGHT · REV 4: THIS REPORT'S OWN READING OF A SOURCE IT ALREADY CITED The orchestration finding had been carried through three revisions on a summary of its source rather than the source's results table. Re-mining the full text for this revision showed the central sentence generalized from one model to all frontier models, when the table shows two others moving the opposite way. No red team caught this; no external party challenged it. It was found by re-reading something already on the shelf, which is the argument for re-deriving rather than re-checking — and it is the correction described in Removed & limits.

RECOUNTED · REV 4: THE SOURCE TILE, AGAIN The full-text shelf that this report's "sources read" tile counts has since been bulk-loaded with roughly 1,300 further documents belonging to other work. Counting the table would have restated 65 as a four-figure number — the same failure mode the red team caught at 50× in Revision 3, arriving by a different route. The tile now counts the curated shelf for this report (68: 43 papers, 25 web documents)

and says so in its own label. **A provenance number is not safe once it has been audited; it is safe while the thing underneath it holds still.**

FOUND AND REPAIRED · REV 4: THE SHELF RECORDED WHERE ITS SOURCES WERE STORED, AND IT WAS WRONG

Sourcing the new citations turned up a defect in the instrument's own provenance. Of the arXiv documents on the full-text shelf, 1,244 recorded a storage location that **does not exist** — a collector running inside a container had written the container's private view of the path into the database. The files themselves were never at risk: every one was found, and a hashed sample matched its stored checksum exactly. But "stored as an immutable original with a content hash" is a trust claim this report makes on its citation page, and for those rows the pointer behind it led nowhere. The same collector was also storing each paper's title as the raw query string used to fetch it, because a pattern written to read the paper's title was reading the search feed's title instead. Both were traced to one collector, fixed at the source so the next run cannot repeat them, and backfilled: 1,244 titles and 1,244 paths corrected, no file missing, a ten-row sample re-verified byte-for-byte against its recorded hash. **A provenance chain is only as good as the weakest pointer in it, and nothing about this one looked broken** — the rows returned data, the hashes were right, and the titles were merely ugly rather than absent.

KILLED: A NOVELTY CLAIM "Provisioning an agent as a scoped bundle of capability and permission" was held as the novel half of the argument. A sweep found a July 2026 paper opening with the same problem statement, with a released dataset and a stronger principle. The novelty claim was dropped.[19]

REVERSED: A DESIGN RULE THE AUTHOR BELIEVED "Weights hold behavior, files hold facts" was proposed as a clean split. A controlled comparison found the opposite on accuracy — tuning beat retrieval by nearly seven points while retrieval added nothing. The rule survives as a risk argument, not a capability one.[18]

REVERSED: THE ARCHITECTURAL ASSUMPTION That a loop should be supervised by a frontier model was contradicted by a measured result in the other direction. Recorded as refuted on the ledger, and it became the headline of this report.[10]

NULL CHECK: IS ROUTING RESEARCH JUST TRACKING THE ARCHIVE'S GROWTH? Routing paper counts rose steeply — but the archive itself nearly tripled over the same window. Normalizing to share of corpus was built so that "no real trend" was an available answer. The share rose anyway, roughly fifteen-fold. The keyword remains uncleaned, so the *shape* is published and the level is not.[2]

NULL CHECK: IS THE FILINGS TREND JUST A CHANGE IN WHO FILES? The rise in control-layer language was re-run on a balanced panel — only companies filing in all four years — so that composition change could produce a flat line if it were the true cause. The trend got *stronger*, not weaker.[3]

FAILED, THEN REBUILT: THE PRACTITIONER-SENTIMENT SENSOR A social corpus of 4,328 posts was expected to show what developers say about these tools. It contained *zero* matching posts — it had been collected scoped to named projects, so the silence was an instrument gap, not a result. A topic-scoped collector was written and run: 18,620 posts, 15 terms, no errors.[4]

CAUGHT: KEYWORD PRECISION WAS NOT FLAT The new corpus was audited by month before any finding rested on it, on the assumption that flat precision would license the trend. Precision was not flat — it rose from 67.5% to 75.5% as the vocabulary settled, which would have manufactured part of any trend measured on raw counts. Every composition figure was recomputed within confirmed on-topic posts.[5]

FAILED: A TREND READ OFF A SAMPLE TOO SMALL TO CARRY IT An early pass called advocacy "flat" from the 60-per-month reference sample. At that size the noise band is about ±6 points — wider than every trend in the finding. Re-measured at 400–800 posts per month, advocacy rises modestly. The error was in reading a trend off a sample sized for accuracy, not for power.[5]

MEASURED THE JUDGE BEFORE TRUSTING IT — THEN GOT CAUGHT QUOTING THE WRONG JUDGE ANYWAY The reference model was run twice over the same 60 posts at temperature zero to establish a ceiling — 98% self-agreement on topic, 97% on who was speaking. Candidates were scored against that ceiling, and the cheap winner was *weak* on identifying the speaker (68% against a 97% ceiling). Revision 2 then published the cheap model's speaker

levels anyway, while its provenance text claimed the reference model's. The red team caught it: the cheap model runs 6–10 points hot on the practitioner share in both years. This revision publishes the reference model's stratified-sample levels, which is what this paragraph should have described the first time.[5][6]

FAILED: THE ASSUMPTION THAT A PRIOR BAKE-OFF TRANSFERS The model that won an earlier classification job on corporate filings came *fourth* here, and the winner on this task was a different one entirely. Neither price tier nor past performance predicted accuracy. Whole-corpus classification cost $0.31.[5]

DEFEATED BY ITS OWN COST CONTROL: THE SHARE MEASUREMENT Five control terms were collected specifically so topic volume could be read as a *share* of conversation rather than a raw count. Collection was then capped at four queries per month to fit a time budget — and that cap pins the busiest terms at a ceiling of roughly 160 posts per month. Every control term and the three busiest topic terms sit at that ceiling, so both sides of the ratio are clipped and the share cannot be computed. The controls were collected and are unusable for their stated purpose. Composition measures, being ratios within a month, are unaffected — and the red team verified the 22 censored windows split 12/10 across the two years, too small and too balanced to manufacture the composition trend.[4][6]

KILLED BY THE PRIOR RED TEAM — AND NEARLY MISSED TWICE A within-company panel finding (F22) was carried in from earlier work, re-run, and confirmed to reproduce numerically. It was already dead: an adversarial re-analysis had shown the qualifying companies are mostly sellers of the software in question, that two are outright false positives, and that a placebo panel puts the persistence rate between on-premise and blockchain. Re-running the query verified the arithmetic and missed the critique entirely. The finding is kept in its broken state rather than deleted — and the second red team then found the same dead reading reused, un-caveated, in a different claim's evidence on the ledger. It is annotated there now.[27][6]

EXCLUDED: A WIDELY QUOTED INDUSTRY STATISTIC A frequently cited figure on vendor purchases outperforming internal builds was ruled inadmissible by the prior red team — a small interview-and-survey base, publicly criticized methodology, and unpublished underlying

data. It is not cited anywhere in this report, and F23 says the evidence base is absent rather than quoting it.[27]

REVERSED: THE CLAIM THIS SENSOR WAS BUILT TO TEST "Practitioner sentiment is turning toward local-first" did not survive its own data. Posts arguing against the practice sit at 1–3% all window, so there was no opposition to turn from — the corpus is self-selected and structurally cannot answer the question asked. It answers a better one: who is doing the talking.[4]

FOUND IN THE COLLECTOR: SILENT INCOMPLETENESS The first version of the new collector resumed per search term rather than per term-and-month. A run stopped part-way left a term looking finished, and the missing months would have read downstream as "nobody posted about this after November." Caught by checking window coverage rather than row counts, and fixed before any finding rested on it.[4]

How it was made, and what it took

The platforms, the models, and the effort — counted from session logs and a spend ledger, not estimated.

Data lives in a column-store database on a home server and is never pruned; history is the point of the instrument. Sources are fetched by purpose-built collectors that record every fetch, *including the ones that failed*, so a gap is visible as a gap. Full-text sources are stored as immutable originals with a hash, and the extracted text is regenerable from them.[1]

425

human prompts across 30 sessions, Aug 3–15, 2026, counted from session transcripts — the project through Revision 3, including

20

days from the instrument's first table to the current revision

4

revisions — one that reframed, one that was attacked and corrected, one that corrected itself

2

adversarial re-analyses: the prior report's, whose kills this report inherits, and a second pass aimed at this one

the red-team session that attacked this report

$7.42

model spend for this report's evidence work and its red team — $5.52 classification and calibration, $1.90 adversarial re-judge

$19.10

total model spend across the instrument's entire life, all projects, from the gateway's spend log — against a $50-a-month ceiling

Effort is counted across the whole project, not the drafting. It covers choosing the questions, standing up the collectors, the validation passes, the analyses that were killed, and the adversarial reviews — quoting only the writing sessions would describe a research program as though it were a blog post. Counts are programmatic from session transcripts, human prompts only, tool output and sub-agent turns excluded; transcript retention covers the instrument's whole life, so the count is complete rather than sampled.[29]

What Revision 4 is *not* counted in, and why. The prompt tile above stops at Revision 3. Revision 4's session is deliberately excluded rather than estimated into it: reproducing the original count needs the same script against the same definition, and a figure produced two ways is worse than a figure with a stated boundary. Its model spend is missing from the dollar tiles for a harder reason — that work ran through a subscription harness rather than the metered gateway those figures come from, so no per-request cost exists for it at any price. **The honest statement is that Revision 4's effort and spend are uninstrumented, not small.** The instrument can measure what passes through its gateway and cannot measure what does not, and a provenance tab that quietly absorbed the gap would be claiming a completeness it does not have.

What it cost, precisely. The classifier ceiling runs, the model bake-off, the whole-corpus classification and the stratified reference sample: **$5.52**. The adversarial re-judge of 672 posts by a third model family: **$1.90**. The corporate-filings classification behind F2 and F22, run for the prior report and reused here: $10.61. Every figure is from the gateway's

spend log, not estimated.[29] Hardware is a home server already owned; its marginal cost is the electricity — which, per Finding 4, is the cheap and misleading number.

Platforms. Python collectors feeding a ClickHouse analytical database, queried in SQL. Drafting and both adversarial passes ran as Claude Code (https://claude.com/claude-code) sessions against that database. The report is hand-authored MDX.

Models, by role. *Claude Opus 5* and *Claude Fable 5* (Anthropic) did research direction, drafting, and the red-team passes, under human instruction. *Claude Opus 5* also served as the classification reference judge — measured against itself before use. *deepseek-v4-flash* (open-weight, via API) classified the 18,620-post corpus for $0.31, with its published levels replaced by the reference judge's where its error ran hot. *gemini-3.5-flash* re-judged the calibration sample during the adversarial pass, blind to the other two. No model generated any number cited as a finding without a measured error bar attached.

Where AI was used — the full statement

Stated as method, not confessed as a caveat — and corrected once, in public, on this page.

The prose of this report was written by an AI (Claude, Anthropic), working from the ledger and the sources, under human direction. The human owns the questions, the instrument, and every editorial judgment, and ordered both adversarial passes.

The numbers come from three places, and the report says which everywhere it matters. Most are deterministic database queries over a stored corpus. Some are quoted from named source documents with their own methodology. One finding's numbers (F3) are produced by a model classifying text — disclosed as method: the classifier was measured against a reference judge before use, its published levels are the reference model's, and an adversarial pass re-judged the same sample with a third model

family and got the same answer. No model summarized evidence into a finding or graded its own work.

What that means for the reader: the framing and the sentences are machine-drafted and human-directed, while the quantities are auditable and re-runnable — and where a number could not be produced honestly, the report says it is missing instead of supplying one. An earlier revision of this disclosure denied the classifier existed; the red team caught it, and the correction is listed above rather than papered over.

Version history

This page renders Revision 6, unabridged.

REV 1 · AUG 15, 2026 Organized around the instrument's own question ("is local-first cheaper?"), presented as a sequence of hypotheses checked. Staged privately, never published.

REV 2 · AUG 15, 2026 Reframed to start from the business problem, with every term explained before use and one narrative through the evidence; the claim-by-claim accounting moved to Provenance. Two claims settled since Revision 1, both as *mixed*, and a third reframed from a sentiment turn — which its own data refuted — to a shift in who is talking. Staged privately, never published.

REV 3 · AUG 15, 2026 — THE ADVERSARIAL PASS, AND PUBLICATION A fresh red-team session attacked every finding before release. No finding died; five were reframed (F3's levels re-anchored to the reference judge, F7 recut around the same-toolbox comparison, F18's numbers replaced with a persisted query, F19's cost ground narrowed to the frontier tier, F20 scoped to offline experiments). Two provenance defects on this tab — an inflated source count and a false AI-disclosure sentence — were found, fixed, and documented above. Effort and cost figures were instrumented and added. This is the first published revision.

REV 4 · AUG 22, 2026 — THE CONTROLLED STUDY THIS REPORT ASKED FOR, PUBLISHED BY SOMEONE ELSE A bake-off designed to test this report's own headline was canceled when a sweep found the controlled version already in the literature. Three findings were added, all in Part C: a twelve-model supervisor benchmark over one fixed worker pool (F24), the permission-scoping failure that runs flat across every price tier (F25), and a large study where better delegation decisions produced no better work (F26). Two citations added. **Two things were corrected rather than added.** F7's central sentence had generalized from one model to all frontier models across three revisions, on a summary rather than the source's own results table — the numbers stand, the generalization is withdrawn, and the mechanism is published in its place. Position 1 in the Overview was sharpened accordingly: the controlled study finds the most expensive supervisor genuinely does win, so the position is now about the price of that premium rather than its existence. The source-count tile was restated from 65 to 68 on a stated definition, after the shelf beneath it was bulk-loaded with unrelated work. Neither correction came from a challenge; both came from re-reading sources already cited.

REV 5 · AUG 25, 2026 A framing pass only. The report was retitled from "The orchestrator was supposed to be the expensive part" to the present title so the subject is named before the punchline, four evidence parts were re-headlined to state each part's takeaway rather than pose its question, the standfirst moved to the site's plain register, two residual first-person-plural phrasings were moved to the report's own register, and the index page now leads with the ambient claim this report measures. No finding, number, grade, confounder or citation changed — the evidence is identical to Revision 4.

REV 6 · OCT 15, 2026 One sentence added to the direction under Finding 23: a link to Field Report No. 6, which has a calculator for making the build-or-buy call on an organization's own numbers. No finding, number, grade, confounder or citation changed.

PART A

It is where things are going: attention, adoption and the conversation all moved

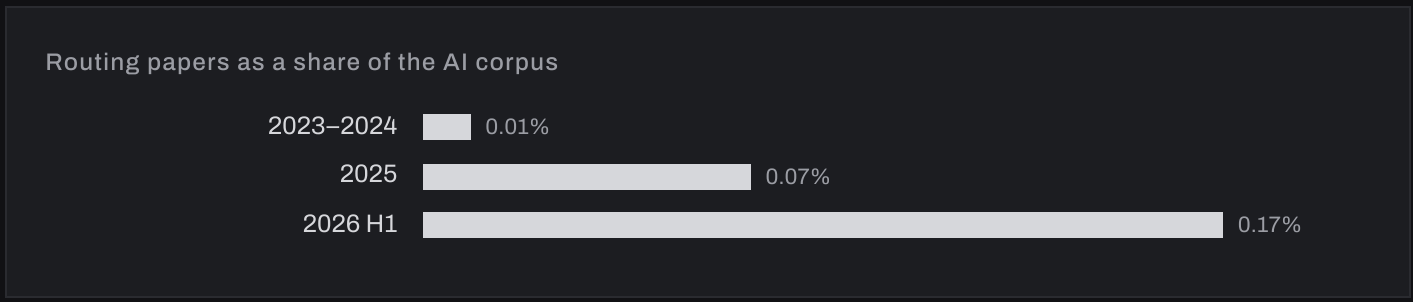
Before spending anything: is the direction real, or a sales cycle? Three independent populations were checked — what researchers publish, what public companies tell investors, and what practitioners say out loud.

FINDING 1 STRONG · OWN CORPUS

Research attention to routing is accelerating, faster than AI research overall

> **Takeaway: the question of which model should handle which task went from a rounding error to one of the fastest-growing topics in the field, inflecting in early 2026.**

Routing, plainly: automatically sending each request to the cheapest model that can handle it, instead of sending everything to the most capable one. Counted as a share of the whole archive rather than as raw totals — the archive itself grew from 11,303 to 32,206 papers per quarter over the window, so raw counts would have risen even if nothing changed.[2]



Roughly fifteenfold, with the share doubling in a single quarter at the start of 2026 and holding. The keyword is uncleaned — it still catches network routers — so the shape is the finding and the level is not published.

DIRECTION

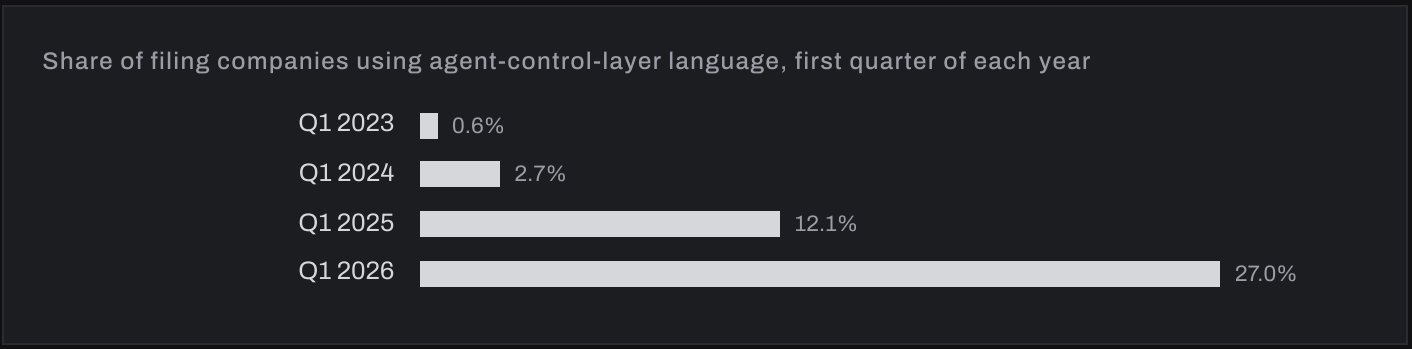
Treat model selection as a live engineering discipline with a moving state of the art, not a settled configuration choice. A decision made on today's landscape should be expected to need revisiting within a year.

FINDING 2 MODERATE · LANGUAGE, NOT CAPITAL

Public companies started describing this layer to investors, and the curve is steep

> **Takeaway: among companies filing in every year since 2023, the share whose filings discuss routing, orchestration, model governance or inference cost went from under one percent to roughly a quarter.**

Measured on a **balanced panel** — only companies that filed in all four years, so the trend cannot be an artifact of new companies entering the corpus. Restricting the panel made the trend stronger, not weaker.[3]



Like quarters are compared because first quarters carry the annual-report wave at roughly double the volume of other quarters — comparing adjacent quarters would read a filing-calendar effect as a decline. **The objection, stated plainly:** this measures what companies wrote, not what they spent. Filings are strategically written, and vocabulary spreads through them by imitation as much as by action. The keyword list is also uncleaned — "AI agents" will catch a company describing a customer-service chatbot. The shape is robust; the exact level is a function of a particular word list.

DIRECTION

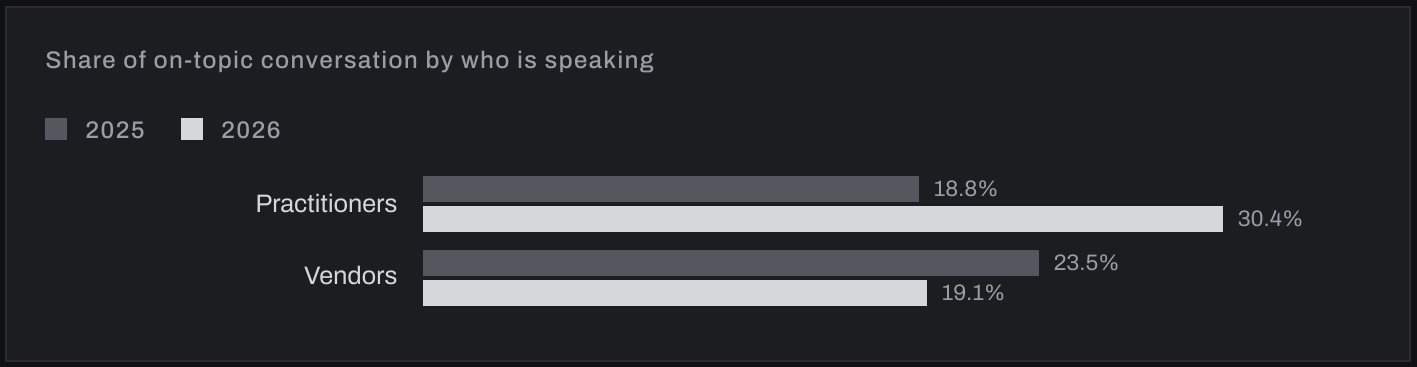
Expect this vocabulary in competitors' disclosures and vendor pitches within the year. Treat its presence as evidence of attention, not of spending — and ask any vendor citing "market adoption" which of the two they measured.

FINDING 3 MODERATE · PURPOSE-BUILT CORPUS

The conversation shifted from vendors to practitioners — which is not the same as opinion changing

> **Takeaway: among posts genuinely about this subject, the practitioner share rose from roughly 19% to 30% over the year while the vendor share drifted down — measured by the reference judge, cross-checked by a third model family.**

This required building an instrument rather than running a query. The social corpus held before this revision — 4,328 posts — had been collected scoped to named software projects and contained *zero* posts on this subject. That silence was an instrument gap, and it looks identical to a genuine absence. A topic-scoped collector was built and run: 18,620 posts across 15 search terms.[4]



Measured by the reference judge on a 672-post stratified sample (±8 points on the practitioner change at 95%), within on-topic posts only. The cheap whole-corpus classifier shows the same trend steeper (29% → 38%) but runs 6–10 points hot on the practitioner level; the reference-anchored figures are the ones published.[5]

The audit changed the analysis, which is why it was run first. Keyword precision is *not* flat across the window: it rises from 67.5% to 75.5% as the vocabulary settles. Left uncorrected, that drift alone would have manufactured part of the trend; every figure above is therefore computed within posts confirmed on-topic, which removes the drift by construction. **And the trend survived a hostile re-measurement:** an adversarial pass re-judged the same sample with a third model family, blind to the first two — practitioner 17.7% → 33.4%, vendor 24.9% → 16.3%, within one to three points of the reference on every cell.[6]

The original framing does not survive. The claim on the ledger was that sentiment is *turning* toward local-first. It is not, because there is nothing to turn from: posts arguing against the practice run at 1–3% for the entire window. A self-selected corpus cannot measure a change of mind. It can measure who is talking, and that did change.

> DIRECTION
>
> When a vendor cites "developer momentum," ask who is doing the talking. That is measurable and it is the objection that matters — but note that no keyword corpus can tell you whether anyone changed their mind, only who showed up.

PART B

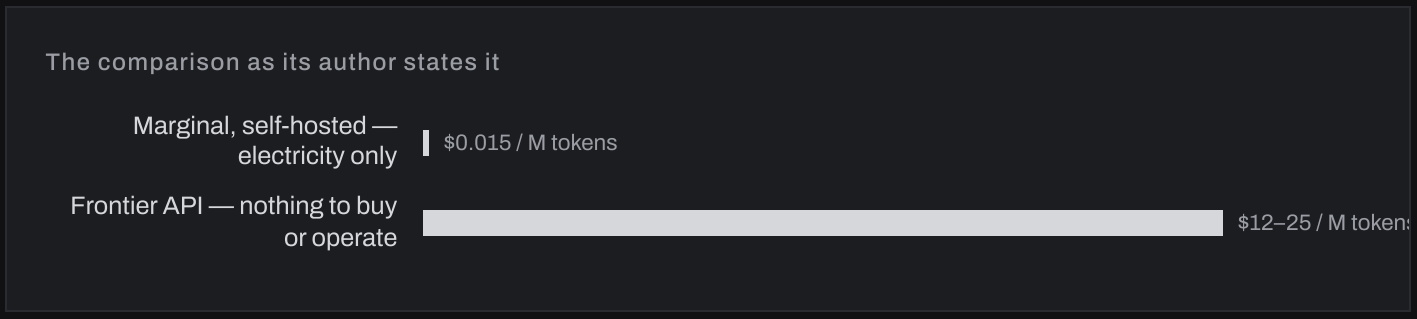
The famous savings number is real, and it is the wrong one to decide on

The cost claims in circulation mix three different things: the price of electricity, the price of a subscription, and the price of the hardware. They have to be separated before any of them can be acted on.

FINDING 4 MODERATE · VERIFIED BUT NARROW

The thousandfold saving is real, correctly labeled, and the wrong number to decide on

> **Takeaway: it is a marginal electricity cost that assumes the hardware is already bought and stays busy — the author says so himself; the figure travels without the caveat.**



What it excludes, quantified. Roughly $45,000 of built hardware — two professional GPUs at about $15,000 each plus memory, before processor and storage. At the same frontier prices, that capital alone buys on the order of **3.5 billion tokens** of frontier-class output with nothing to operate and nothing to depreciate.[8] The marginal figure also only holds while the machine stays busy; idle hardware depreciates on an asset that is obsolete in about three years.

DIRECTION

When a cost claim is quoted, ask what it excludes. A marginal figure answers "what does one more unit cost" — the right question only after the capital is already spent and committed.

FINDING 5 STRONG · TWO LEVERS, SEPARATED

Routing saves about a third. Owning the hardware is where the large multiples live.

> **Takeaway: these are independent levers with wildly different costs, and presenting them as one number hides which one a company would be pulling.**

Measured across a routing benchmark: up to **31.7% cost reduction** while matching the best single model's performance, and about a 4% accuracy gain over the best single model when optimizing for accuracy instead.[9] Real, bankable, and not an order of magnitude. Not universal, either: several published routers fail to beat simply always using the best single model, and two named approaches struggle to trade cost for savings without losing accuracy. Pool size is not the lever — measured against a perfect router, adding more candidate models shows clear diminishing returns. *Which* small set is offered matters far more than how many.

DIRECTION

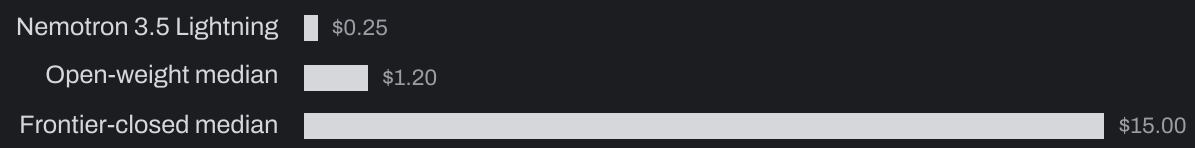
Pull the routing lever first. It is a software change against hosted models, needs no capital, and delivers a measured third off at matched quality — while the large multiple requires $45,000 and an engineer.

FINDING 6 STRONG · LIVE CATALOG

The cheap tier is now genuinely cheap, and it is a hosted product

> **Takeaway: open models with no hardware to buy are priced roughly twelve times below frontier models at the median, and about sixty times below for the specific small model this topic grew up around.**

Output price per million tokens, live catalog snapshot of 413 models



An **open-weight model** is one whose trained parameters are published, so anyone can run it — on their own machines or a hosting provider's. It is a licensing distinction, not a quality one. Frontier-closed spans $10–37.50 across its middle half. Price is not

capability — nothing here measures whether the tiers are substitutable for a given task.[7] One caution: several vendors publish free endpoints for these models. A price of zero exists at the vendor's discretion and should never be modeled as durable.

DIRECTION

A company can act on the cost argument this quarter with no capital expenditure at all, by moving routine work to hosted open models. The purchase decision and the savings decision are separable, and only one of them is urgent.

PART C

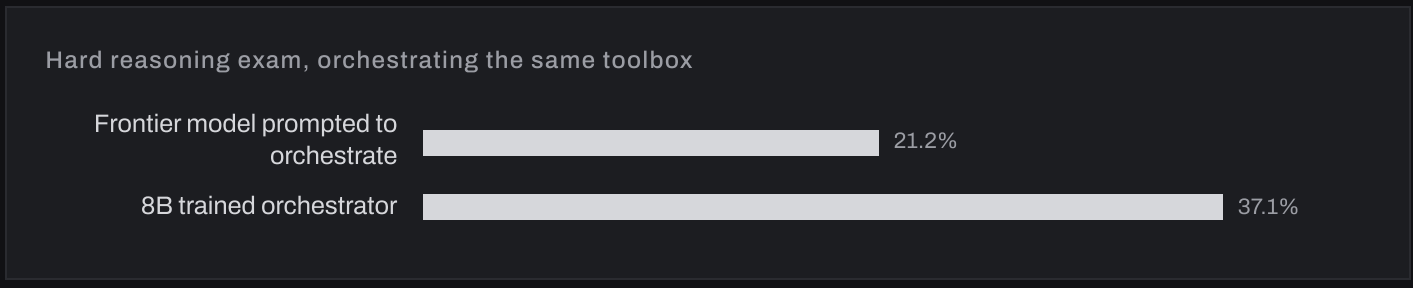
Coordination is cheap to buy, and paying more for it buys almost nothing

If the direction is real and the cheap tier is cheap, the question becomes structural: who supervises, how much that seat is worth paying for, what the supervisor is reliably bad at, and how many specialists one machine can hold.

FINDING 7 STRONG · MEASURED; ITS GENERALIZATION CORRECTED AT REV 4

Coordination is a separate skill, and it can be trained into a small model cheaply

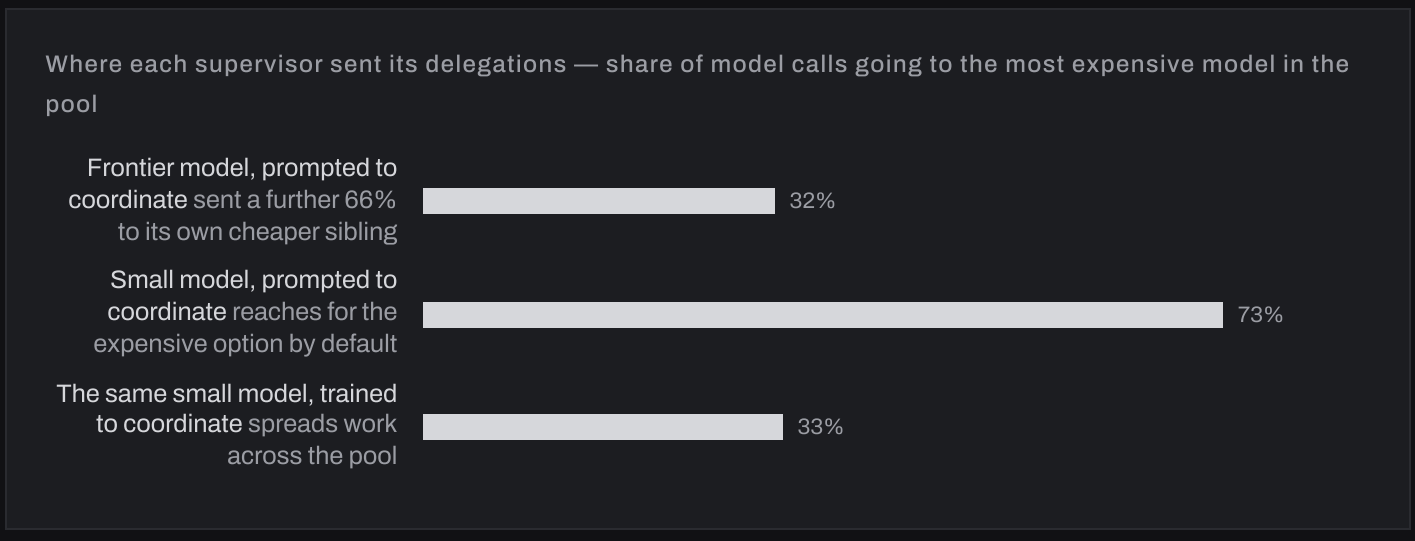
Takeaway: an 8B model trained for the coordinating job outscored every model prompted to do it, including frontier ones — but the reason is trained habit, not size, and the earlier reading that frontier models simply fail at coordination does not survive the results table.



An 8-billion-parameter model trained with reinforcement learning against outcome, efficiency and preference rewards, coordinating other models and tools — including the frontier model itself as a callable tool. The frontier model given only basic search-and-code tools scored 35.1%; given the full toolbox and asked to orchestrate it, 21.2%. On two further benchmarks the trained orchestrator finished 2.3 and 2.5 points ahead at roughly 30% of the cost, and generalized to tools it had not been trained on.[10]

Read the comparison carefully — the red team did. The trained orchestrator *can call the frontier model as a tool*, so this is not "small model beats big model." It is: the coordinating skill can be trained into a small model, and it is trainable cheaply. The single cleanest number in the study is the one that changes nothing but the training: the orchestrator's own base model, prompted to coordinate the identical pool, scores 30.6% at three times the cost. Training the role into that same backbone is worth 6.5 points and two-thirds of the bill.

Revision 4 corrects an over-general reading of this finding. Earlier revisions said that prompting a frontier model to coordinate "measurably fails." The full results table does not support that in general — it supports it about *one model*. Of the six models tested in the coordinating seat, only that frontier model got worse when handed the larger toolbox. A second frontier model improved sharply under orchestration, from 19.8% to 34.6%, and a large open-weight model went from 14.0% to 32.8%. The authors attribute the outlier's collapse to a specific documented habit rather than to coordination being hard:



The frontier model routed 98% of its delegations inside its own vendor family — a third to itself, two-thirds to its own cheaper sibling — and essentially never used the open models on the bench. The untrained small model has the opposite reflex, sending nearly three-quarters of everything to the most expensive option. Training replaces both reflexes with a distribution. **So the finding is about habits, not horsepower**, and a buyer should read it as evidence that coordination is a trainable behavior rather than as evidence that expensive models cannot coordinate.

Four honesty notes. The margins on the two further benchmarks are small (2.3 and 2.5 points, against the paper's own "wide margin" phrasing). The frontier vendor's self-reported score on one of them is 84.2 against the 77.7 these authors could reproduce — **had that reproduction succeeded, the trained orchestrator's 80.2 would be a loss rather than a win on that benchmark.** When the whole pool is swapped for tools the orchestrator never trained on, every score collapses — the orchestrator's own headline falls from 37.1% to 22.0% — and while its relative lead widens, the paper does not discuss the drop. And the study is authored by the same vendor research group as F8's position paper, so the two corroborate each other exactly once.

This was recorded on the ledger as a **refuted** claim: the author had written down the opposite — that the supervisor should be the frontier model — before testing it. It is the most useful row on the board, and F24 is what happens when someone runs the controlled version.

> DIRECTION
>
> Do not assume the coordinating role requires the most capable model available — but do not read this as the reverse claim either. What is measured is that coordination can be trained into a small model cheaply, and that at least one expensive model coordinates badly for a reason peculiar to itself.

FINDING 8 ARGUED POSITION, VENDOR-AUTHORED

Most of the work in an agent job is small work

> **Takeaway: tool calls, extraction, formatting and validation make up the bulk of the steps, and small models are argued to be sufficient for them — but the case is a position paper from a vendor that sells small models.**

The source is explicit that it is laying out a position rather than reporting a benchmark, and its authors are at the company selling the models in question. Graded accordingly: admissible as a framing, not as proof. It predates the discussion that surfaced it by nearly a year — and it comes from the same research group as F7's orchestrator study, so those two findings share an author list and must not be read as independent confirmation of each other.[11]

DIRECTION

Treat "small models are enough" as a plausible default worth testing on a real task mix, not as an established result. The claim is directionally supported by the cost evidence and has not been independently demonstrated.

FINDING 9 STRONG · MEASURED CONFIGURATION

One machine can hold many specialists at once

> **Takeaway: a single loaded base model can serve many lightweight specializations simultaneously, with per-request selection — which is what makes "a specialized agent per team" affordable rather than absurd.**

An **adapter** is a small file of extra parameters that specializes a general model for one domain, without retraining or duplicating the whole model — a skill the base model can pick up and put down. Measured: eight adapters loaded in parallel on a 20-billion-parameter base, producing 144 output tokens per second at 135 milliseconds to first token, improving to 171 and 124 with tuned settings — at 1,600 input and 600 output tokens, on a named software version.[12] The infrastructure keeps adapters separate from

the base model and pulls the right one into a cache per request; the same mechanism is documented by a second serving stack.[13]

The scar is worth more than the confirmation. The first working implementation was ten times *worse* on time-to-first-token than the plain base model, because a compiler treated a length-dependent value as fixed and rebuilt the adapter code for every new input length. One compiler hint fixed it. This is not work an ordinary IT team walks into.

DIRECTION

The economics of specialization depend on this working. Verify the software version before planning around it — the measured configuration required a specific release, and an earlier attempt was ten times slower than doing nothing.

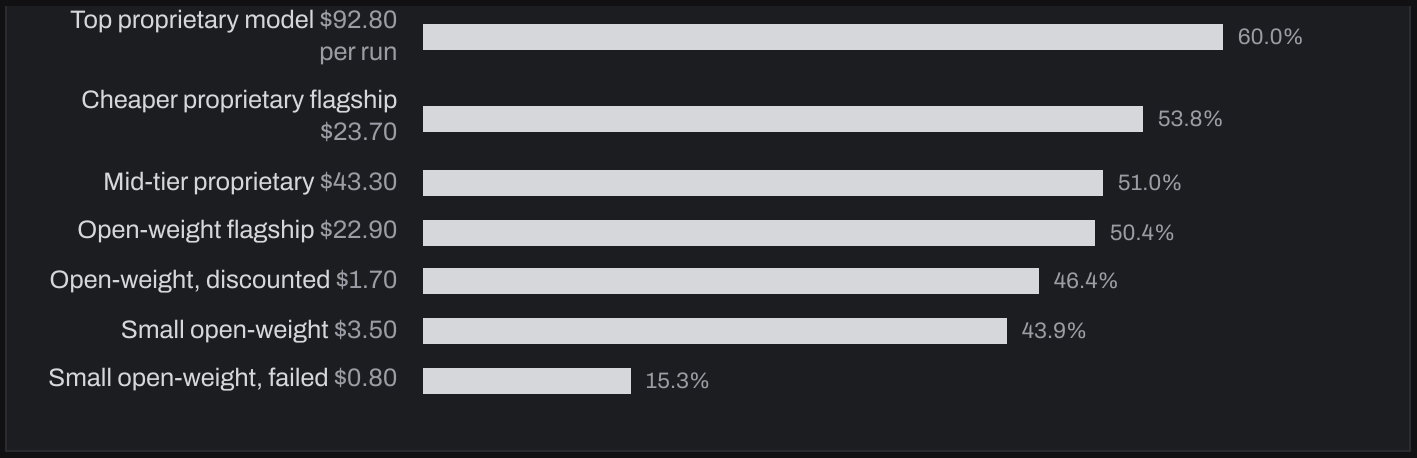
FINDING 24 STRONG · TWELVE MODELS, ONE FIXED WORKER POOL

Twelve supervisors were measured over the same workers — and price bought almost nothing

> **Takeaway: with the worker pool held fixed and only the supervisor swapped, running cost varied by more than a hundredfold while management quality varied by less than four — and the best supervisor was the most expensive one.**

This is the experiment F7 implies but does not run. An academic benchmark published in July 2026 constrains a main agent to text-only perception and partial access to its workspace, so that delegating is the only way to finish, then hands it a fixed pool of three locally-served small workers. Twelve different models were placed in the supervisor seat over identical scenarios — 41 multi-step scenarios across 258 scored rounds, every round checked by executing a command with a declared exit code rather than by asking a model to grade it.[30]

Same workers, twelve supervisors — management score, with running cost



The authors state the spread plainly: cost runs from $0.80 to $93 per run while the score spans under four times. The top model costs about four times the next one down and buys six points for it. Three of the expensive proprietary entries are *dominated* — another model scores higher and costs less. One model at $1.70 per run lands within fourteen points of the leader.

This does not say cheap supervisors are as good. The most expensive model genuinely won, and one cheap model collapsed outright at 15.3% — supervision is a real capability with a real floor, and the floor is not far below the prices most buyers are paying. What the data refutes is the *pricing* of the premium, not its existence.

Four caveats, stated. The top score is a composite: that model ships with a vendor-recommended fallback to a second model on refusal, so its number is the pair, not the model alone. Costs for self-hosted models are imputed from a public rate card, and the authors call them approximate. Every model was run once — there is no variance estimate anywhere in the table. And the fixed worker pool is what isolates management skill, but it also ties every number to one level of worker capability.

DIRECTION

Ask what the supervisor premium actually buys on the work in question, and price it. The evidence says the premium is real, small, and charged at a multiple wildly out of proportion to it — which makes this the cheapest place in the architecture to run an experiment before committing.

FINDING 25 STRONG · MEASURED, AND IT FAILS FOR EVERY MODEL

Supervisors are bad at granting permissions, and none of them are good at it

> **Takeaway: across all twelve supervisors, the thing they were worst at was not breaking work into pieces — it was handing each worker only the access that piece needed.**

In the same twelve-model run, the components of the score separate cleanly. Choosing the right kind of worker: above 92% for every capable model. Respecting read-only boundaries: above 92% for every capable model. Granting the right files: **never above 50% for any model tested**, with the best at 49.2%. The authors report that workers are routinely handed roughly twice the files they touch, and more tools than they use.[30]

The behavioral spread underneath the scores is far wider than the scores themselves. Below the leader, ten models cluster inside a 9.9-point band — but the rate at which their workers hit access they should never have had ranges across roughly an order of magnitude. **Models that look interchangeable on a leaderboard are not behaving alike**, and the leaderboard is not measuring the difference.

This is the same exposure F15 describes, arriving from a different direction and with a number attached. It is also a cost problem rather than only a safety one: handing a worker twice the files it needs inflates what it has to read.

DIRECTION

Do not expect any model to scope permissions correctly on its own, at any price tier. This is the argument for granting access per task in the harness rather than trusting the supervisor to be careful — and a reason to log what each worker was granted against what it actually opened, since that gap is measurable from day one.

FINDING 26 MODERATE · LARGE STUDY, SMALL DELEGATION COUNTS

Teaching a supervisor who its workers are made it choose better and changed nothing

> **Takeaway: a study that doubled how often a supervisor picked the genuinely best worker measured no improvement in the finished work — while separately showing that perfect delegation would be worth 15 to 31 points.**

A May 2026 benchmark fixed the tasks, fixed a pool of eleven models from seven vendors, fixed the delegation interface, and varied only how much the supervisor was told about its workers. Across 23,375 task instances, the rate at which the supervisor picked the best available worker rose from 14.2% to 29.5%. End-task quality across those same conditions moved by at most a hundredth of a point on a zero-to-one scale, with every condition statistically indistinguishable from knowing nothing at all.[31]

Meanwhile the same study measures what perfect delegation would be worth by giving every task to the demonstrably best worker for it: between 15 and 31 points above what any configuration achieved. The headroom is real and nobody is reaching it.

The honest reading is that this is a null result with a soft denominator. The 29.5% figure rests on 122 delegation decisions, because telling a supervisor about its workers made it delegate *less* — and the response is not orderly, since one of the information formats scored 7.5%, below telling it nothing. No confidence interval is published on those rates. The flat-quality result is the robust half, resting on the full 23,375 instances.

One useful side result: the study checked whether supervisors favor workers from their own vendor. Several do, by 1.5 to 3.7 times chance — though the largest ratios rest on very few decisions — and one major vendor's models were neutral.

DIRECTION

Do not assume better routing decisions convert into better finished work; on this evidence they did not, and the mechanism connecting them is unestablished. The measurable thing is the gap between what was chosen and what would have been best, which needs the per-step decision log F11 already argues for.

PART D

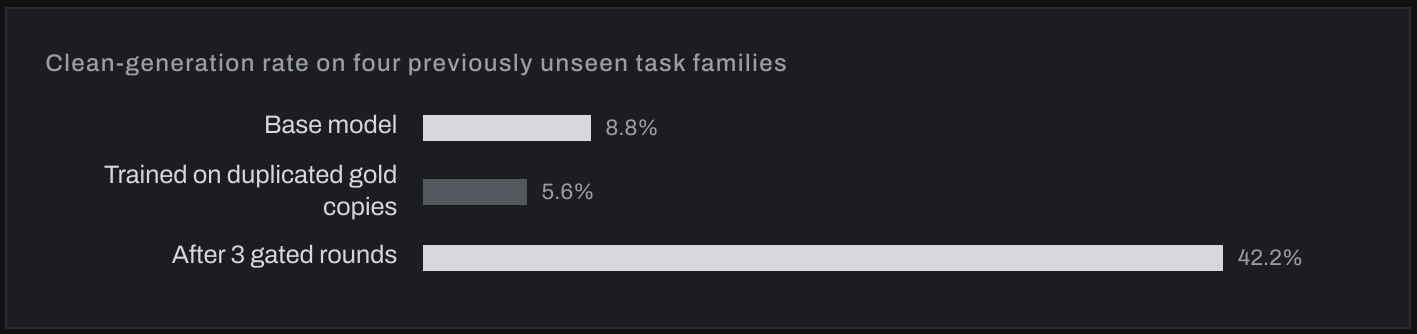
The verification gate is the whole gain

The promise that makes a team of agents strategically interesting is not that it works on day one. It is that it improves with use. That promise has a mechanism, and the mechanism is narrower and stricter than it sounds.

FINDING 10 STRONG · CONTROLLED, WITH THE DECISIVE CONTROL

The gate is the whole gain

> **Takeaway: training a model on its own work improves it dramatically — but only when the work is filtered by an automatic check of whether it actually functioned; swap in a lenient check and the entire improvement disappears.**



The middle bar is the control that matters: simply duplicating good examples regressed below the untrained base model. More data made it worse. Coverage across repeated attempts went from 18 of 25 to 25 of 25 — the ceiling — and all three training rounds were statistically significant. Then the decisive test: the authors reran the identical loop changing *only* the gate, to a lenient check that passes 99.9% of attempts. The entire gain vanished back to baseline.[14]

What the numbers are numbers about. The domain is generating small video-game projects — the gate is "does the game actually launch" — with one 14-billion-parameter model and twenty-five held-out tasks, from academic authors with no vendor stake. The 42.2% is a fact about game scaffolds, not about ticket queues; what travels is the design rule (the gate is the mechanism), not the effect size. That is why this is the load-bearing

finding of the whole report: the value is not in the training, the model, or the data volume. It is in having a test that can say no.

> DIRECTION
>
> Before planning any self-improving loop, find the automatic pass/fail test for the work in question. If the output cannot be checked mechanically, the flywheel has no mechanism and the plan is a hope.

FINDING 11 MODERATE · PUBLISHED ARCHITECTURE

The routing decisions are themselves the training data

> **Takeaway: deciding which model handles each step, inside the running job rather than at the front door, produces a structured record of every decision and its outcome — which trains the next router.**

The published framing is worth borrowing: an agent is base models *plus a harness*, not one model in a wrapper. Models are specializing along different axes — code editing, long-context recall, tool use, latency — so no single model is best at every step, which makes model selection a systems problem rather than a serving trick.[15]

> DIRECTION
>
> If a company wants a system that compounds, the logging design matters more than the model choice. What gets recorded at each step is what can be learned from later; a system that logs only inputs and final answers has thrown the useful part away.

FINDING 12 OPEN · OBSERVED, CAUSE UNESTABLISHED

The raw material for that flywheel is getting harder to export

> **Takeaway: detailed reasoning traces from frontier providers are becoming less available, with at least one instance where they are returned encrypted by design — a supply risk to any plan that depends on capturing them.**

Eighteen open issues on one major provider's command-line tool concern reasoning summaries alone; one is a mechanism rather than a bug — reasoning content returned encrypted, a product decision, not a regression.[16] **The causal claim is not established, and is recorded as refuting the author's own thesis.** The stated industry rationale for withholding raw reasoning is safety and monitorability, not protecting commercial value. Both explanations predict the same observation, and this evidence cannot separate them.

DIRECTION

Do not build a strategy whose critical input is another company's optional output. Check today whether the traces a plan depends on are actually retrievable, and design assuming they may stop.

PART E

What cannot be bought around

Three limits appear repeatedly in the literature and will not yield to budget. They are the ones most likely to be discovered late and expensively.

FINDING 13 STRONG · 21 METHODS, 5 BENCHMARKS

Routing has a ceiling, and it is not close to perfect

> **Takeaway: twenty-one routing methods across five benchmarks converge into a narrow band far below a perfect router, and the best of them trails that ideal by ten to thirty points everywhere.**

On one benchmark the top fifteen routers differ by **0.23 percentage points**. A plain nearest-neighbor lookup with no learning at all ranks in the top two on all five. The plateau persists under cost-aware objectives, not just accuracy-maximizing ones.[17] **The cause matters more than the number:** routers learn coarse general patterns of which models tend to be good, not instance-level judgments about which model gets *this specific request* right. The authors' own remedy — far more training data, a bigger encoder, end-to-end tuning — bought 2.13 points, closing under 15% of the gap.

DIRECTION

Never accept a promise of frontier-quality answers from cheap models via clever dispatch. Routing is a cost instrument at acceptable quality — never a parity instrument.

FINDING 14 MODERATE · SOURCES DISAGREE

Teaching a model your facts by training is a risk, not a shortcut

> **Takeaway: fine-tuning new facts into a model increases confident wrong answers about what it already knew — though on raw accuracy in one controlled test, tuning beat retrieval outright.**

Retrieval, plainly: looking the answer up in an organization's own documents at the moment of the question, instead of trying to bake it into the model beforehand. Sources genuinely disagree here, which is the finding. One study establishes the mechanism — teaching new facts causes interference among overlapping representations, and freezing the parameter groups that hold facts preserves performance while reducing fabrication. The freezing remedy has a price the paper states plainly: it works by suppressing new-

fact learning almost entirely, so it fits only where the model is not supposed to acquire facts in the first place. A controlled comparison at small scale found the opposite on accuracy: domain tuning gained nearly seven points while retrieval added nothing. A third found a small tuned model with retrieval matching a model nearly twice its size.[18]

The practical read: they are complementary, there is a known-good configuration rather than a shrug, and the failure mode of getting it wrong is fluent, confident, invented specifics.

DIRECTION

Use training to shape behavior, vocabulary and procedure. Use retrieval over live documents for anything that must be exactly right — and expect the two to be complementary rather than competing.

FINDING 15 MODERATE · PRIOR ART, STRONGER THAN EXPECTED

Permissions are the hard part, and someone has already framed it better

> **Takeaway: enterprise agents are typically granted every credential their role might ever need, for every task — and that standing over-privilege, not the model's reasoning, is the attack surface.**

Published work opens with exactly this problem statement and takes it further: role-based ceilings, a task-context classifier and policy-derived prohibited combinations, with a released dataset over a fifteen-permission taxonomy validated at high inter-rater agreement on a reviewed sample. Iterating the dataset against policy cut ceiling violations from 46 to 3.[19] One scope note: the dataset is the contribution — the architecture around it is proposed, not yet implemented or benchmarked, so this is prior art on the problem framing rather than a system to buy.

DIRECTION

Scope an agent's access to the task in front of it rather than to the role behind it. A credential that is not present cannot be misused however the model behaves — prevention rather than detection.

PART F

What nobody has measured

The honest center of the report: the quantities the evidence base does not contain.

FINDING 16 OPEN · NO SOURCE MEASURES IT

There is no credible published measure of the productivity gain

> **Takeaway: the corpus is rich on cost, latency, accuracy and architecture, and empty on whether any of this makes an organization measurably more productive.**

The entire business case rests on productivity, and productivity is the one quantity the evidence base does not contain. What *can* be measured cheaply, and would stand in usefully: cost per completed task, the automatic pass rate from F10's gate, and how often a job escalates to an expensive model.

DIRECTION

Any vendor quoting a productivity multiple is quoting their own marketing. A buyer who wants that number will have to measure it — which means capturing a baseline before deploying anything, the step that is almost always skipped and cannot be recovered afterward.

FINDING 17 OPEN · INSTRUMENT TOO YOUNG

How fast this depreciates is not yet measurable

> **Takeaway: the argument against buying hardware rests on models turning over faster than the hardware amortizes — and the price history needed to measure that turnover is days old.**

Daily pricing snapshots of a 413-model hosted catalog began mid-August 2026. Vendors overwrite yesterday's prices, so missed days are unrecoverable at any cost — which is why the collection runs whether or not a question is waiting on it. A usable decay rate needs roughly a quarter of history.[7]

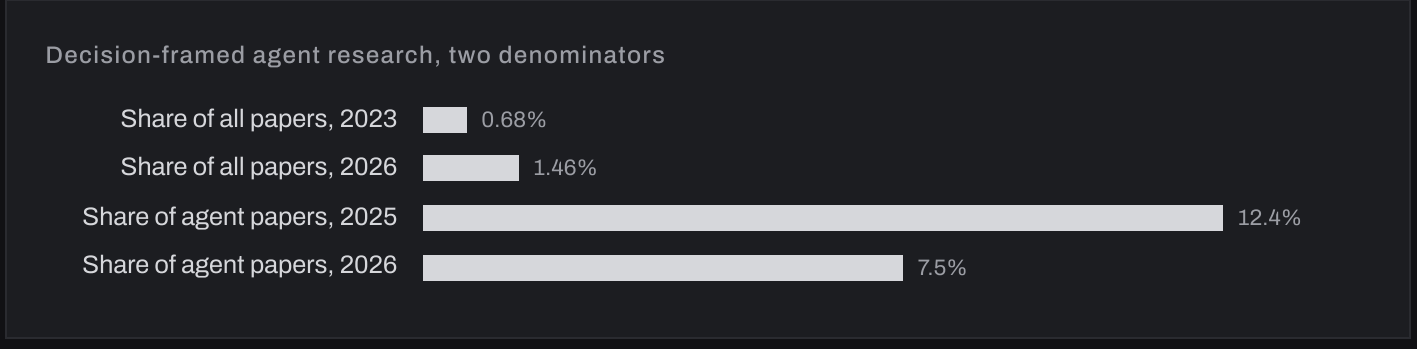
DIRECTION

Treat any claimed depreciation rate as invented until someone shows the price series behind it. This is measurable, but only by someone who started collecting before they needed the answer.

FINDING 18 MODERATE · COUNTS ONE WAY, SHARE THE OTHER

Decision-support framing is growing in volume and shrinking in share

> **Takeaway: agent research framed around decision-making roughly doubled as a portion of all published work, while losing about a third of its portion of agent research specifically — the field is growing faster somewhere else.**



Two denominators, two different stories — which is why both are shown. Against the whole archive the framing is growing strongly; against agent research alone it is losing ground. Both keyword lists are uncleaned, so the shape is the finding and the level is not — and the query behind these bars is stored on the ledger. An earlier revision showed different levels from a drafting-session query that was never persisted; the red team flagged it, and this revision replaced the numbers with re-runnable ones.[20]

DIRECTION

Do not assume the research frontier is working on the decision-support framing an enterprise buyer cares about. The gap between where research concentrates and where a business problem sits is itself a planning input.

PART G

The question an executive will actually ask

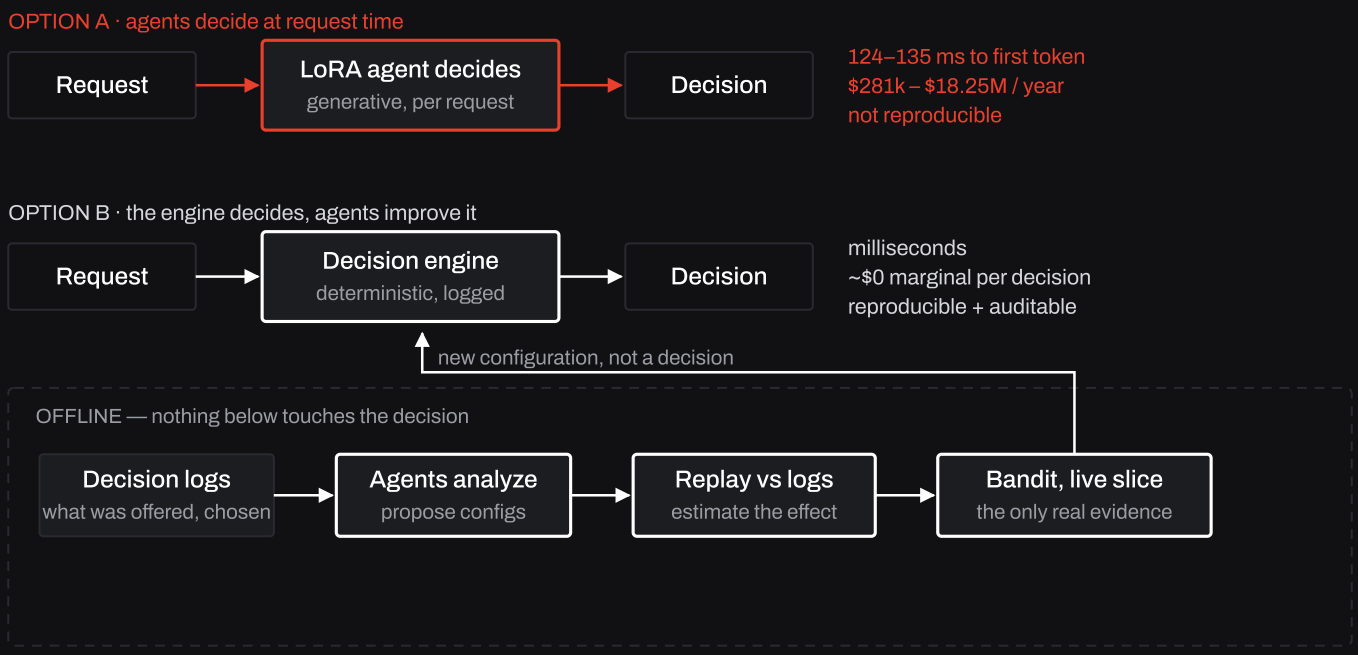
If small models are cheap and good enough, why keep the decision system at all — why not let the agents make the decisions? And when the invoice for a decisioning platform arrives, is this research an argument for replacing it? Both questions have answers in the evidence, and they point the same way.

A **decision engine**, plainly: the system that picks what to show, offer or approve for a specific customer in a specific moment — which options were eligible, which were withheld and why, what was recommended, in what order. It runs inside a page load or an ad auction, and it logs every decision.

FINDING 19 STRONG · THREE GROUNDS, SEPARATELY EVIDENCED

Putting a large generative model on the decision path fails on latency and reproducibility — and, for the frontier tier, on cost

> **Takeaway: a production advertising system states plainly that a few-hundred-millisecond auction budget rules out running a large model at request time — and resolves it with exactly the split this report recommends. Small models on the path are normal; large ones do not fit.**

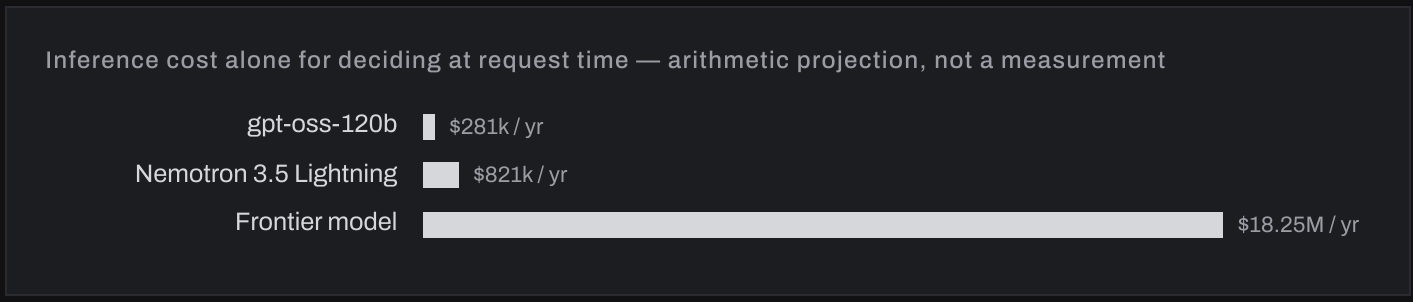


The offline band is the whole argument: everything in it improves the engine, and none of it is in the request path. Exploration is decided by the bandit, not by the model — published work found that letting the model choose its own directions produces unstable search *(F20)*.

The same request, two architectures, drawn at the same scale. Latency and cost figures are carried from the findings; the cost line is an arithmetic projection from published prices at 10 million decisions a day, not a measurement.

Latency. The constraint is quoted from a production system, not inferred: ads must be scored within a few hundred milliseconds to enter the auction, which rules out a large sequence encoder at request time.[21] Their fix is a heavy offline encoder writing a cached representation, with a lightweight model at serving time — the same boundary, reached

independently. It recovers 72–80% of a full-history runtime model's quality, though that ideal cannot be served inside the budget anyway.



Projected from published prices at 2,000 input and 100 output tokens per decision and 10 million decisions a day; the token assumption does real work, and cost scales linearly with both tokens and volume.[7] Two honest adjustments a buyer should make: request-path prompt caching roughly halves the frontier bar (batch pricing does not apply — batch processing is asynchronous, which the latency budget already rules out), and a deterministic engine's near-zero marginal cost sits on top of a platform license this report does not price. At low decision volume the cost comparison inverts; latency and reproducibility do not.

Reproducibility. Ask a model the same question twice and the answer moves. On a fully crossed corpus of 12,933 responses, that variance separates into at least four distinct sources — resampling, prompt phrasing, model identity, and language.[22] Graded honestly: that study measures brand-sentiment scoring, not decisioning, so it is cited for the *structure* of the variance — four separate doors it walks in through — not for the magnitudes. A decision that cannot be reproduced cannot be explained afterward, and a regulated decision that cannot be explained is a liability with a latency budget.

The strongest counter-case reinforces the boundary. A unified generative recommender that replaces a whole ranking cascade still distills a *training-only* teacher model into the serving module, explicitly to avoid putting an expensive second model on the serving path.[23] Even the end-to-end generative camp keeps the big model offline. One disclosure the red team insisted on: this counter-case and the latency paper above come from the same organization and share an author — one production culture's design

philosophy observed twice, not two independent confirmations. The cost and reproducibility grounds do not depend on either.

DIRECTION

Refuse the "cut out the middleman" proposal on arithmetic rather than on principle. Latency and reproducibility each rule it out on their own at any tier; the cost ground rules out the frontier tier specifically.

FINDING 20 STRONG · MEASURED, AND THE REASON IS STATED

Let a bandit decide what gets tried, not the model

> **Takeaway: a published system pairs an agent with an explore/exploit router precisely because letting the model choose its own directions produces unstable search on a limited experiment budget.**

Explore/exploit, plainly: spending most of a limited budget on what already works while spending a little on alternatives, so you keep learning. The budget can be experiments in a lab or live traffic in production — this finding is about the first; the logging evidence under F21 is about the second.

The published system separates the loop into two roles: a bandit router picks the next modification direction from historical validation feedback, while the model generates the concrete hypothesis and code edit *within* that direction. Across multiple tasks, datasets and model backbones it produces more stable improvement and uses a limited trial budget more effectively — and it is not an assertion: the ablation that lets the model choose its own directions hits an improving trial 22% of the time against the bandit's 48%.[24] **The reason they split the roles is the finding:** allowing the model to both select directions and generate hypotheses "often leads to unstable search under limited experiment budgets." The bandit is not decoration on top of the agent; it is there because the agent explores badly.

Scope stated plainly: the bandit routes *offline experiments* — which change gets tried against held-out data next. Its one live element tested a single already-chosen candidate. Live-traffic exploration is a different budget with its own evidence, and this paper is not evidence about it. That evidence exists: replaying candidate configurations against logged decision data is a mature technique — roughly 950 papers since 2023 on evaluating a policy from data collected under a different one.[25] It is also emphatically *not* the same thing as simulating customers with a language model, which failed badly (F21).

DIRECTION

If an organization builds a self-improving optimization loop, put a classical explore/exploit mechanism in charge of which experiment runs next and let the agent generate candidates within the chosen direction. That division is not bureaucratic caution — it is the measured configuration.

FINDING 21 MODERATE · REFUTED FOR ONE USE, SUPPORTED FOR ANOTHER

Replay the configuration through the real stack — do not ask a model to imitate the customers

> **Takeaway: estimating a change by replaying it against real logged decisions is well founded; simulating customer responses with a language model is not, and four independent studies say so.**

These two get conflated constantly, and the evidence pulls in opposite directions. Replaying a proposed configuration against logged decisions uses *real recorded human behavior*. Asking a model to role-play customers generates synthetic behavior, and the synthetic version fails. One precision the red team demanded: the four studies below measure survey answers and belief updates, not purchases — this is refutation by strong analogy plus mechanism, not a direct offer-simulation experiment. The analogy is unusually tight: the failure they measure is individual-level and subgroup-level infidelity, which is exactly the level an offer simulation must get right; the strongest of the four

names offer personalization as the decision at risk; and one contains a small direct choice probe (a booking task) showing the same collapse.[26]

Across two independent domains of real survey data, four models and two model families, no model beat even the strongest non-model baseline at the individual level.
Independent model agents grounded on 2,414 real respondents collapse onto a modal answer rather than reproducing the population — an 85% collapse rate.
Against 843 real respondents, only one of eight models met a prespecified equivalence criterion.
Against 391 real participants, all six models failed to produce faithful belief updates from their own generated starting points.

And the logs are not a free byproduct. The accuracy of any replay estimate depends heavily on how the data was collected, which creates a reward–coverage tradeoff: concentrating on the best-known action reduces variance but stops producing evidence about the alternatives.[25] An engine tuned to always present its single best offer is quietly destroying the data its own future optimization depends on. Deliberate exploration is a precondition, not a refinement.

DIRECTION

Treat model-simulated customers as a way to generate hypotheses, never as a way to test them. The replay path uses real recorded behavior and is the one that carries weight.

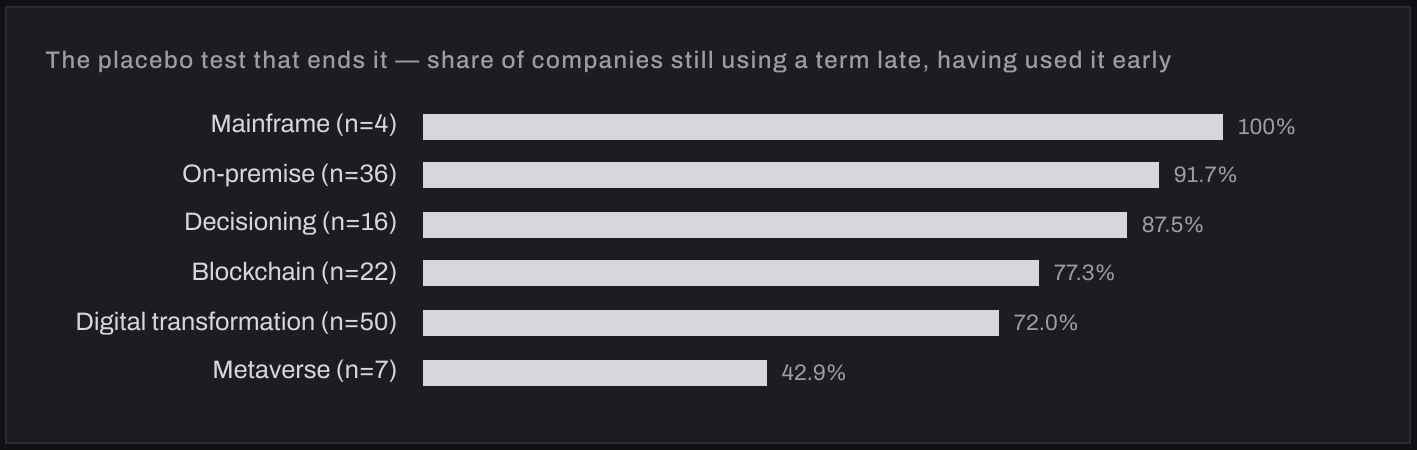
FINDING 22 OPEN · KILLED BY ITS OWN RED TEAM

The market-behavior reading of this data did not survive, and the placebo test is why

> **Takeaway: companies using decisioning language before 2025 overwhelmingly still use it — but they are mostly the companies that sell it, and dead technologies persist in filings at the same rate.**

This finding was carried into an earlier report and then destroyed by an adversarial re-analysis of that report. It is kept here, in its broken state, because a report that shows its dead is harder to ambush.[27]

Who the 16 companies are. Most sell decisioning or decision-adjacent software — so their continued use of the vocabulary is product marketing, not buyer behavior. Two were false positives on inspection: an oilfield-services firm matching on drilling "real-time decision" and a storage vendor matching on marketing copy. Genuine buy-side usage is roughly one company. Both firms that supposedly "dropped the layer" are the noise, not a signal.



Decisioning sits between on-premise and blockchain. Language persists in filings whether or not the technology is winning, so persistence measures the stickiness of vocabulary rather than the fate of a category. What survives is narrow and worth keeping: among companies that talk about this area at all, agent language is being *added* far more often than decisioning language is dropped. That is a statement about how vendors describe themselves. It is not evidence about what buyers run.

DIRECTION

Distrust persistence-in-filings as evidence that a technology is holding its ground. Ask who the persisting companies are, and ask what the same measurement says about a technology everyone agrees is finished.

FINDING 23 OPEN · NO OUTCOME RECORD EXISTS

Build versus buy has no evidence base in this category, in either direction

> **Takeaway: the current published work on the decision is a reasoning framework explicitly designed for cold-start situations where historical data is unavailable — which concedes that the evidence base is missing.**

The available structured approach offers an ontology of decision factors with rule-based reasoning and reference-level matching, built to function "in cold-start scenarios where historical data is unavailable." That is a way to reason without evidence, and it is honest about being one.[28]

What this research does and does not bear on. A decisioning contract buys at least two distinct things: an auditable, low-latency runtime, and the tooling to optimize it. Everything in this report bears on the *second*. Agentic analysis has become genuinely cheap — this report's own classification work cost $0.31 — so the optimization half of such a contract now has a credible alternative, while the runtime half does not. That is a renegotiation lens, not a cancellation argument, and it is the honest limit of what the evidence supports.

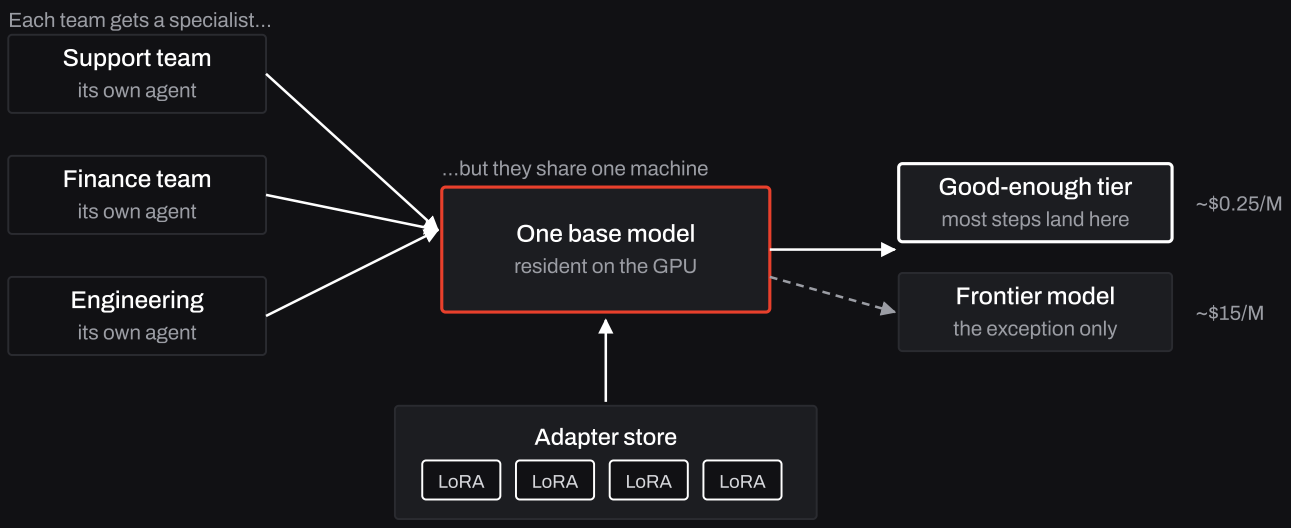
DIRECTION

Distrust any confident claim that building beats buying here, or the reverse, including one delivered with a case study attached. Nobody has published the outcome record, so the decision has to be made on an organization's own numbers. I heard software was dead has a calculator for that.

What the evidence actually points to

The same skeleton every executive has been shown, with the four silent decisions filled in — then the steps, in order, each traced to findings.

One machine, many specialists. The instinct behind "ten agents" is ten deployments, and that instinct is what makes the idea sound unaffordable. It is not how the serving layer works. A single base model stays loaded, and each team's specialization is a small adapter file selected per request — measured at eight running in parallel on one base (F9). The frontier model stays in the picture as the exception rather than the default (F5, F7).

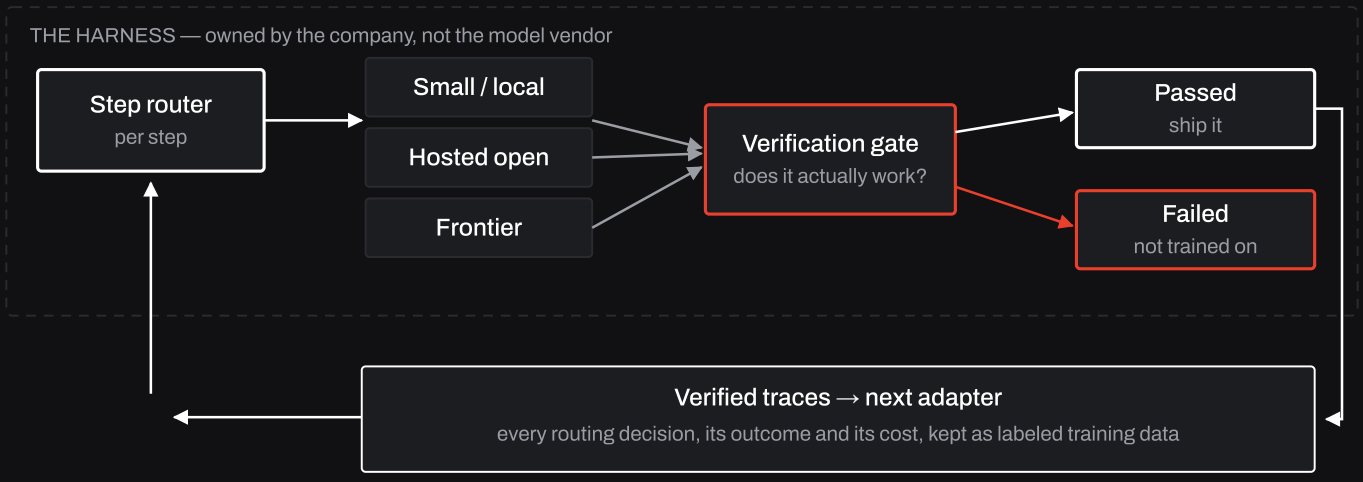


Ten specialized agents do not mean ten deployments. One base model stays resident and the per-team specialization is a small adapter file selected per request — measured at eight adapters in parallel on a 20B base (F9). This is what makes 'an agent per team' an affordable sentence.

This is available as a hosted product as well as something to run on owned hardware — the adapter-serving mechanism is the same either way, which is what makes Step 1 below possible without buying anything (F6).

The loop that makes it compound. The previous diagram is a cost structure. It does not, by itself, improve. What turns a deployment into an asset is the harness around it:

routing decided per step rather than per request, a mechanical check on whether the work functioned, and the verified results kept as training data for the next adapter.



The compounding asset is the trace log, not the model — and it only exists if the harness is owned *(F11, F12)*.

The difference between a deployment that stays flat and one that improves. The harness routes per step rather than per request, checks the result mechanically, and keeps only verified work as training data for the next adapter. Remove the gate and the loop does not merely weaken — training on unfiltered output measured WORSE than not training at all (F10).

Two things about this diagram carry the weight of the report. The **gate** is not a quality-assurance nicety — remove it and the loop does not merely weaken; training on unfiltered output measured *worse than not training at all* (F10). And the **trace log is the compounding asset**, not the model. It exists only if the harness belongs to the company, which is also why the shrinking exportability of frontier reasoning traces is a supply risk worth watching (F11, F12).

WHAT THESE DIAGRAMS DELIBERATELY DO NOT SHOW

No vendor product names, because the shape is what the evidence supports and the implementations move faster than any report. The serving pattern is

documented by more than one infrastructure stack, and both a commercial and an open-source path exist for every box drawn here. Choosing between them is a procurement question, not a research finding.

Step 1 · Take the free third before buying anything. Move routine, high-volume work to hosted open models and route the exceptions upward. This needs no hardware, no capital request and no data-center conversation. The measured saving is about a third at matched quality (F5), and the price gap between tiers is roughly twelvefold at the median (F6). A company that does only this step has captured the part of the argument that is real, bankable and reversible.

Step 2 · Find the gate before designing the loop. The single highest-leverage question in the whole program: *for the work in question, what is the automatic test that says it functioned?* Tests pass or fail. A ticket is resolved or reopened. A record validates or does not. Where such a test exists, a self-improving loop has a mechanism and the measured gains are large. Where it does not, the loop has nothing to filter on, and the evidence says training on unfiltered output makes things actively worse than doing nothing (F10). Choose the pilot domain by where the gate is strongest — not by where the enthusiasm is.

Step 3 · Run the pilot with one team, and instrument it before it starts. A workable first pilot is a single team with a repetitive, checkable workload, and four things captured from day one: **a baseline taken before anything is deployed** (productivity is unmeasured in the literature, F16 — the only credible number a company will ever have is its own, and it cannot be reconstructed after the fact); **cost per completed task**, not cost per million tokens; **the gate's pass rate over time** (F10); and **the escalation rate to expensive models** (F5). Record every step's decision and outcome, not just inputs and final answers — that record is the training data for everything that comes later (F11), and a system that did not log it has thrown the compounding away.

Step 4 · Price the supervisor on the pilot's own numbers — it is the cheapest experiment in the architecture. Once Steps 2 and 3 exist, this costs almost nothing: the gate already says whether the work functioned, the instrumentation already reports cost per completed task, and changing which model supervises is a configuration line. Run the same task mix with a cheaper model in charge and read the two numbers off the dashboard that is already running. The evidence says what to expect — measured over one fixed worker pool, running cost varied more than a hundredfold while management quality varied under four, and three expensive models were beaten outright by cheaper ones (F24). It also says not to expect the cheap seat to be free: one model collapsed entirely, so this is a test with a real failure mode and it should be run against the gate rather than against an impression. **Do not assume the answer transfers from anyone else's benchmark** — every published run of this kind is a single run on one task family, and the supervisor's job is specific to the work it supervises.

Step 5 · Scope permissions per task, from the first day. Grant an agent access to what the task in front of it needs, not to what its role might ever need (F15). This is far cheaper to build in at pilot scale than to retrofit across a company, and the standing over-privilege — not the model's behavior — is the exposure. Do not expect the supervisor to do this for you: across twelve models, granting the right file access was the one thing every one of them was bad at, and none cleared 50% (F25). Log what each worker was granted against what it actually opened — the gap is measurable from the first day and nothing else in the system reports it.

Step 6 · Only then consider owning hardware. The large cost multiples require roughly $45,000 of equipment kept genuinely busy, plus the engineering to run it (F4) — and the depreciation rate that would justify or sink that purchase is not yet measurable (F17). A company with a measured, high, steady volume from Steps 1–3 can evaluate this on its own numbers. A company without that evidence is buying a thesis.

Step 7 · Take this into a platform renewal as a narrower question. If an existing decisioning platform contract is up and the number causes sticker shock, the evidence here does *not* support treating agents as a replacement bid. Such a contract buys at least

two things — an auditable, low-latency runtime, and the tooling to optimize it — and everything in this report bears on the second (F19, F23). The useful question at renewal is narrower and more answerable: **how much of what is being paid for is optimization tooling, and does that half still earn its share now that agentic analysis costs cents?** If the engine is not being actively optimized today, that half of the contract's value is already unrealized, and the comparison is against zero rather than against the vendor.

Where to be skeptical

Nine claims a buyer will hear this year, each answered by a finding above.

Any promise of frontier quality from cheap models via smart routing. Twenty-one methods converge well below a perfect router and the gap does not close with effort (F13). Routing buys cost at acceptable quality, never parity.
Any productivity multiple. Nothing in the evidence base supports one (F16).
Any claim that the supervisor must be the most capable model. Measured over one fixed worker pool, the premium bought six points for four times the price, and three expensive models were beaten by cheaper ones (F24, F7).
Any claim that a supervisor can be trusted to scope its own permissions. No model tested cleared 50% at granting the right file access, and the failure is flat across price tiers (F25).
Any assumption that better delegation decisions produce better work. Doubling how often a supervisor picked the best worker moved finished quality by nothing measurable (F26).
Any plan to teach the model company policies by training it on them. That is the configuration that produces fluent, confident, invented specifics (F14).
Vendor-authored sufficiency arguments, including the one this report cites and grades as a position rather than a result (F8).
Any architecture whose critical input is another company's optional output (F12).
Any proposal to have agents make the decisions at request time. Latency and reproducibility each rule it out on their own; at volume, frontier-tier cost does too (F19).

Any claim that building beats buying here, or the reverse. No outcome record has been published for this category (F23).
Any plan to test changes on model-simulated customers. Use replay against real logged decisions instead (F21).

Removed & limits

What died, what was replaced, and the standing limits a reader should carry.

CORRECTED · REV 4 F7's central sentence. Revisions 1–3 read the orchestration result as "prompting a frontier model to coordinate measurably fails." Re-mining the source at full text showed that claim rests on one model out of six tested in that seat — two others *improved* sharply under orchestration, and the outlier's collapse has a documented cause specific to it. The finding stands and its numbers are unchanged; the generalization has been withdrawn and the mechanism published in its place. Nobody challenged this. It was found by re-reading a source already cited, which is the uncomfortable part.

SHARPENED · REV 4 Position 1 in the Overview. It read "the supervisor does not need to be the expensive model," justified by a single vendor study. A controlled twelve-model benchmark now measures the same question, and its answer is more specific and less flattering to the original phrasing: the most expensive supervisor genuinely won. The position is now about the *price* of the premium rather than its existence, and it is justified by the controlled study rather than the vendor one.

ADDED · REV 4 F24, F25 and F26, all in Part C, and two citations. They are numbered after the existing findings rather than inserted in sequence, because renumbering would break every cross-reference in the document; a reader meeting F24 immediately after F9 is meeting a later addition.

REPLACED · REV 3 F18's numbers. The original bars came from a drafting-session query whose term list was never persisted — a violation of this report's own provenance promise, caught by the red team. The finding survives with re-derived numbers whose query is stored on the ledger; the shape is unchanged, the levels moved.

REFRAMED · REV 3 F3's published levels (they were the cheap classifier's, 6–10 points hot — now the reference judge's), F7's headline comparison (now the same-toolbox baseline), F19's cost ground (now scoped to the frontier tier), F20's scope (now offline experiments, explicitly). No finding was withdrawn; each reframe is described inside its finding.

REMOVED · REV 2 The claim-by-claim structure of Revision 1. No finding was withdrawn; the accounting moved to Provenance, where the ledger belongs.

STANDING LIMIT · THE SOCIAL CORPUS IS SELF-SELECTED F3 measures who is talking, never whether anyone changed their mind. Collection capped bisection at four queries per month: 532 windows hit the cap and 22 remained censored at single-day granularity, which loses a window's oldest posts and therefore biases *against* finding an early rise. The red team verified the censored windows are balanced across years.

STANDING LIMIT · VOLUME TRENDS ARE NOT USABLE Two terms had their earliest months collected before the per-month cap was introduced, so those months are denser than later ones by collection method rather than by activity. Nothing in this report rests on a count of posts over time; F3 uses only ratios computed within a month, which that difference does not affect.

STANDING LIMIT · KEYWORD MEASUREMENTS The corpus trends (F1, F2, F18) rest on uncleaned keyword matching with no hand-labeled precision audit. Shapes are published; levels are not.

STANDING LIMIT · SINGLE-SOURCE FINDINGS F7, F9, F10, F24, F25 and F26 each rest on one published result. They are strong results with real controls; none has been independently replicated; and each finding names its domain, its authors' stake, and what does not travel beyond it. **F24 is not a replication of F7** and should not be read as one — it

measures models *prompted* to supervise, with no trained arm on its bench, so the two agree on the shape of the question and answer different halves of it.

STANDING LIMIT · THE TRAINED-SUPERVISOR RESULT IS STILL UNREPLICATED, AND IT IS THE LOAD-BEARING ONE

The claim that coordination can be *trained* into a small model rests entirely on one vendor-authored study (F7) whose orchestrator was trained on the same tool catalog it was evaluated over, and whose one genuinely held-out-tool experiment saw every score collapse. The controlled twelve-model benchmark that arrived since (F24) does not test a trained supervisor at all. **So the single most economically consequential claim in this report — that the coordinating role is cheap to buy because it can be trained rather than rented — has one source, one vendor, and no independent check.** It is reported as strong because its internal controls are real, and flagged here because a buyer betting on it is betting on one paper.

UNTESTABLE · THE SKILL BARRIER Whether this pipeline is accessible to ordinary IT staff cannot be settled from literature — it is a claim about how hard work feels to a non-specialist. The adjacent evidence cuts against easy optimism: the measured adapter configuration needed compiler-level debugging (F9).

What this report is not claiming. Not that capital is flowing to this layer — only that disclosure language is. Not that hosted models beat owned hardware — the depreciation rate that would settle it is unmeasured. Not that small models are sufficient for most work — that remains a vendor's argued position. Not that any of this raises productivity, which nobody has shown. And not that the cost figures here are capability figures; price is not quality, and no benchmark in this report says otherwise.

APPENDIX

Citations

Two kinds: **Data** — this project's own datasets and query code, stored, versioned, re-runnable for audit — and **Ext** — external sources, each graded where used. Where two sources share an organization or an author list, the entry says so — shared provenance is counted once.

DATA The claim ledger: 22 claims on the topic, each stored with its origin (external / the author's hypothesis / derived from reading), the method recorded with the claim, findings and conclusion kept as separate fields, refuted claims retained. The staging page renders its counts live from this ledger. Full-text reading shelf for this report: 68 sources (43 papers, 25 web documents) stored as immutable originals with content hashes, counted apart from the wider archive that shares the table. The stored location recorded against 1,244 of that archive's rows was found wrong during Revision 4 and repaired — see the audit note in How it was tested.

DATA Research corpus: 290,240 arXiv abstracts, January 2023 – August 2026, share-of-corpus routing measurement with the archive-growth null check (11,303 → 32,206 papers/quarter). Keyword uncleaned; shape published, level not.

DATA Corporate filings corpus: 29,287 extracted passages, 2023-Q1 – 2026-Q3, balanced-panel restriction (companies filing in all four years), like-quarter comparison. Q1 2023 0.6% → Q1 2026 27.0%.

DATA Social corpus: 18,620 posts from 12,608 accounts, Sep 2025 – Aug 2026, 15 search terms (10 topic + 5 control), 1,081 queries, zero errors, resumable per term-and-month. Cap-censoring accounting: 532 windows capped, 22 censored at day granularity.

DATA Classifier calibration: reference judge measured against itself first (98% on-topic / 97% speaker self-agreement, n=60, temperature 0); candidate bake-off scored against that ceiling; whole-corpus classification $0.31; monthly precision audit (67.5% → 75.5%, drift removed by conditioning on confirmed on-topic posts); published composition levels from the reference judge's 672-post stratified sample.

DATA Adversarial re-analysis, second pass, Aug 15, 2026: a fresh session with no stake in the draft. Methods-level re-read of all 18 cited papers; every quoted number re-derived or traced to source; third-model re-judge of the 672-post calibration sample ($1.90, results within 1–3 points of the reference); censoring-by-month test; two provenance defects found on the trust page itself. Verdict file and re-judge artifacts stored with the analysis (REDTEAM.md, rejudge_x.py, rejudge_gemini.json).

DATA Inference price sensor: daily snapshots of a 413-model hosted catalog (prices, capability flags), begun mid-August 2026. Vendors overwrite yesterday's prices, so the series exists only because collection started before the question needed it. Cost projections in F19 are arithmetic over this catalog's published prices.

EXT Practitioner cost write-up (public forum, 2026): the $0.015/M-token marginal electricity figure with the author's own statement that it excludes hardware; ~$45,000 build cost from the public discussion beneath it. Graded: honest primary source, narrow scope.

EXT LLMRouterBench (arXiv:2601.07206): 400K+ instances, 21 datasets, 33 models. Up to 31.7% cost reduction at matched best-single-model performance; ~4% accuracy gain; named binary routers fail to trade cost for savings; pool-size diminishing returns. Graded: strong for the cost figure. Verified at methods level by the red team.

EXT ToolOrchestra (arXiv:2511.21689, vendor research lab): 8B orchestrator on a Qwen3-8B base, GRPO-trained on outcome/efficiency/preference rewards, 16 H100s. 37.1% vs 21.2% (frontier model prompted to orchestrate the same expanded toolbox) and 35.1% (frontier model with basic tools only) on the text-only subset of a hard reasoning exam; +2.3/+2.5 points on two further benchmarks at ~30% cost. Same-backbone control: the untrained base model scores 30.6% in the identical full-pool configuration at 27.6¢ against the trained model's 9.2¢. Of six models tested in the orchestrating seat, only the one frontier model degrades under the larger toolbox; two others improve sharply (19.8→34.6, 14.0→32.8) — the paper attributes the outlier to vendor-family routing bias, measured at 98% of its delegations. Held-out-tool generalization collapses every score (its own 37.1→22.0) while widening the relative lead; the paper does not discuss the drop. ⚠ The authors could not reproduce the frontier vendor's self-reported 84.2 on one benchmark, reaching 77.7 — at 84.2 their 80.2 would be a loss. Limitations exist only on the project page, not in the paper: unproven above 8B, untested outside reasoning tasks. Graded: strong result, vendor-authored, judge circularity noted (the frontier model also scores the outcome reward and sits in the pool as a tool); same research group as c11 — counted as one voice with it. Re-mined at full text for Revision 4.

EXT Small-model sufficiency position paper (arXiv:2506.02153, vendor research, v2 Sep 2025): explicitly a position, not a benchmark. Graded: contested; used only as framing, never as evidence of sufficiency. Same research group as c10.

EXT Multi-adapter serving report (inference project with a cloud provider, Feb 2026): GPT-OSS 20B base, 8 adapters in parallel, LoRA rank 32, 1,600 input / 600 output tokens, vLLM 0.15.0 — 144 OTPS / 135 ms TTFT, 171 / 124 tuned; the 10× TTFT compiler-recompilation regression and its one-hint fix. Graded: strong for the configuration, vendor-published. All numbers verified verbatim by the red team.

EXT Adapter-store serving documentation (a second inference stack): the same per-request adapter selection mechanism, documented independently of c12.

EXT "The Verifier is the Curriculum" (arXiv:2607.09709, academic — Science Tokyo / Zhejiang / NUS): execution-gated self-distillation on video-game project generation, Qwen3-14B + LoRA. 8.8% → 42.2% over three gated rounds (all significant); gold-duplication control regresses to 5.6% (p=0.019); gate-swap control (lenient check passing 99.9%) erases the entire gain (p=1e-3); coverage 18/25 → 25/25. Graded: strong; the decisive control is what earns it; single model, single benchmark, N=25 — the design rule travels, the effect size does not.

EXT Harness-native routing architecture (arXiv:2607.11399, open-sourced Jul 2026): architectural argument, not a benchmark. Graded: moderate; used for framing.

EXT Public issue tracker of a frontier provider's tooling, counted Aug 15, 2026: 18 open issues on reasoning summaries, one documenting encrypted-by-design reasoning content. Graded: observational; establishes the pattern, not the cause; count not independently re-verified.

EXT "The Routing Plateau" (arXiv:2606.07587, university + industry): 21 methods, 5 benchmarks, one unified setup. Best router trails oracle by 10–30 points; top-15 spread 0.23 points on one benchmark; kNN top-2 on all five; plateau persists cost-aware; authors' remedy closes 14.6% of the gap. Graded: strong; the central negative result of this report. Verified at methods level.

EXT Fine-tuning and hallucination mechanism study (arXiv:2604.15574 — interference among overlapping representations; freezing mitigation, which suppresses new-fact learning and is scoped by its authors to settings where that is acceptable); controlled tuning-vs-retrieval comparison (arXiv:2604.23801 — +6.8 points at 4B scale, retrieval null); LoRA configuration ablation (arXiv:2605.28222). Graded: mixed — they disagree, and the disagreement is reported rather than resolved.

EXT Dynamic capability scoping (arXiv:2607.22445, independent researcher, workshop): 600-prompt dataset over a 15-permission taxonomy, κ = 0.917/0.967 on a reviewed sample; ceiling violations 46 → 3 by iterating dataset against policy. Graded: strong prior art on the framing; the architecture is proposed, not implemented.

DATA Decision-support framing share: re-runnable corpus query stored on the ledger (claim `decision-support-framing-share`) with its full term lists. 2023: 0.68% of all papers / 10.3% of agent papers; 2026: 1.46% / 7.5%. Replaces Revision 2's unpersisted drafting-session numbers.

EXT Long-history user transformers for real-time ad ranking (arXiv:2607.14331, production advertising system): the few-hundred-millisecond constraint quoted verbatim; offline encoder + cached representation + lightweight runtime model; 72–80% offline quality retention of an unservable full-history model; deployed via live A/B. Graded: strong, production. Shares an organization and an author with c23 — counted as one production culture with it.

EXT Variance-components decomposition of LLM non-determinism (arXiv:2607.13304): 12,933 responses, fully crossed; four sources — resampling, prompt phrasing, model identity, language.

Graded: moderate for this use — the task is brand-sentiment measurement, so it is cited for the structure of the variance, not the magnitudes.

EXT Gryphon-v2 generate-and-rank recommender (arXiv:2608.06213): replaces a 15-generator production cascade; distills a training-only teacher into the serving module explicitly to avoid a second model on the serving path; deployed via online A/B. Graded: strong, production. Shares an organization and an author with c21.

EXT RecHarness bandit-routed agentic harness (arXiv:2607.29241): Thompson-style routing over offline experiment directions; ablation vs model-chosen directions (improving-trial hit rate 47.9% vs 21.7%); one live element tested a single already-chosen candidate. Graded: strong for the offline split; not evidence about live-traffic exploration, and cited accordingly.

EXT Off-policy evaluation literature (964 papers, 2023–2026, from the research corpus); logging-policy design and the reward–coverage tradeoff (arXiv:2605.15108); replay-to-launch-readiness framework (arXiv:2605.12840 — offline improvement is evidence for a live test, not a substitute; benchmark-log study). Graded: mature method literature.

EXT Four studies refuting LLM-simulated survey respondents: cross-domain benchmark vs non-LLM baselines (arXiv:2607.26348 — no model beats the strongest baseline at the individual level; its own discussion names offer personalization as the decision at risk); distribution-first population simulation (arXiv:2607.18310 — 85% collapse on 2,414 respondents; a measured mitigation exists but over-disperses; includes a small booking-task choice probe showing the same collapse); urban-publics behavioral replication (arXiv:2607.27100 — 1 of 8 models met the authors' prespecified criterion against 843 respondents); belief-update simulation (arXiv:2607.28347 — all 6 models fail from self-generated starting points; 4 of 6 pass given true starting points, so the failure is the cold-start regime this use case requires). All four verified at methods level by the red team.

DATA Prior adversarial re-analysis (Aug 11, 2026, of the preceding report): seller-cohort and false-positive audit of the 16-company decisioning panel; placebo persistence panel (mainframe 100%, on-premise 91.7%, decisioning 87.5%, blockchain 77.3%, digital transformation 72.0%, metaverse 42.9%); exclusion of a widely quoted vendor-outcome statistic as inadmissible. This report inherits those kills; the second red team verified the inherited rewrite reproduces the prior verdict exactly and annotated the one place on the ledger where the dead reading had been reused without it.

EXT Build-vs-buy decision framework (arXiv:2606.29816): cited from its abstract, where the quoted cold-start phrase appears verbatim; not on this report's full-text shelf. Graded: honest framework, concedes the missing evidence base.

DATA Effort and cost accounting: human prompts counted programmatically from Claude Code session transcripts (tool results and sub-agent turns excluded), 425 prompts across 30 sessions, Aug 3–15, 2026, covering the project through Revision 3; transcript retention covers the instrument's entire

life, so that count is complete rather than sampled. **Revision 4's session is excluded from it** rather than estimated in, and Revision 4's model spend is absent from the dollar figures below because that work ran through a subscription harness rather than the metered gateway — uninstrumented, and stated as such rather than absorbed. Model spend from the LiteLLM gateway's per-request log: $5.52 report evidence work, $1.90 adversarial re-judge, $10.61 prior-report filings classification, $19.10 instrument lifetime.

EXT ClawArena-Team (arXiv:2606.31174v2, July 2, 2026, academic — UNC-Chapel Hill / UC Berkeley / UC Santa Cruz): twelve models placed in the supervisor seat over one fixed, locally-served worker pool. 41 multimodal multi-directory scenarios, 258 scored rounds, 72 staged mid-run updates; the main agent is constrained to text-only perception and partial workspace access so delegation is forced. Management score = task completion × the mean of four least-privilege and routing components. **All scoring is execution-based with declared exit codes — no model grades any outcome.** Cost $0.80–$93 per run (>100×) against a score span under 4×; workspace-permission precision never reaches 50% for any model (best 49.2); ten models cluster in a 9.9-point band while per-subagent forbidden-access rates spread roughly an order of magnitude. Graded: strong — the controlled version of F7's question, and the only source here that isolates the supervisor with an execution-checked score. Four author-stated limits carried into F24: the top score is a model-plus-refusal-fallback composite, self-hosted costs are imputed from a public rate card and called approximate, every model was run **once** with no variance estimate, and the fixed pool ties all results to one level of worker capability.

EXT DecisionBench (arXiv:2605.19099v1, May 18, 2026, academic — Penn / Columbia / MIT / Stanford / OpenMesh AI): fixed task suite (GAIA, τ-bench, BFCL v4 multi-turn), fixed 11-model / 7-vendor peer pool at a pinned price date, fixed delegation interface; only the supervisor's information about its workers varies across five conditions. n = 23,375 task instances. Best-worker selection rate 14.2% (blind) → 29.5% (on-demand profiles); end-task quality flat, mixed-effects $|\beta| \le 0.010$ with all $p \ge 0.21$ on the full sample. Perfect-delegation ceiling sits 15–31 points above every observed configuration and stays positive under a 70%-realization stress test. Vendor self-preference 1.5–3.7× chance for several models; one major vendor neutral. ⚠️ Graded: moderate, and the grade is about the denominators — the 29.5% rests on 122 delegation decisions, awareness *reduces* delegation frequency, one information format scores 7.5% (below blind), and no confidence interval is published on those rates. The flat-quality null is the robust half and is what F26 rests on. Authors' own limits include single-seed evaluation and near-zero delegation on one of the three suites.