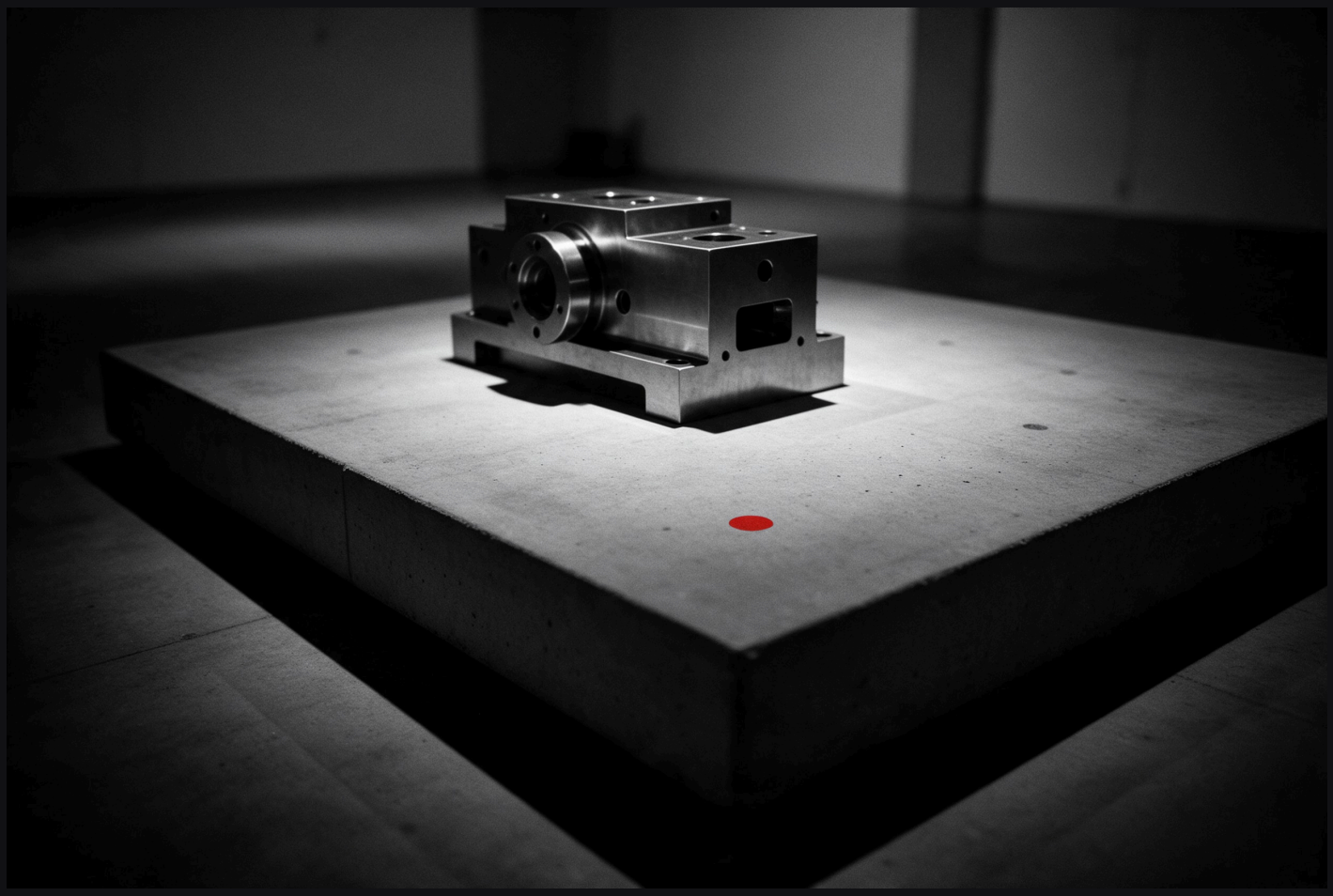


What AI customer decisions should stand on

AI is going to make more of the decisions companies make about their customers. So what should those decisions stand on? This report measures the question and shows its work, so you can audit the conclusion instead of taking it on faith.

AUG 12, 2026 · REVISION 10 · LAB.WORTHINGTON.CLOUD

What AI customer decisions should stand on

AI is going to make more of the decisions companies make about their customers. So what should those decisions stand on? This report measures the question and shows its work, so you can audit the conclusion instead of taking it on faith.

001 Field Report Aug 12, 2026 Revision 10

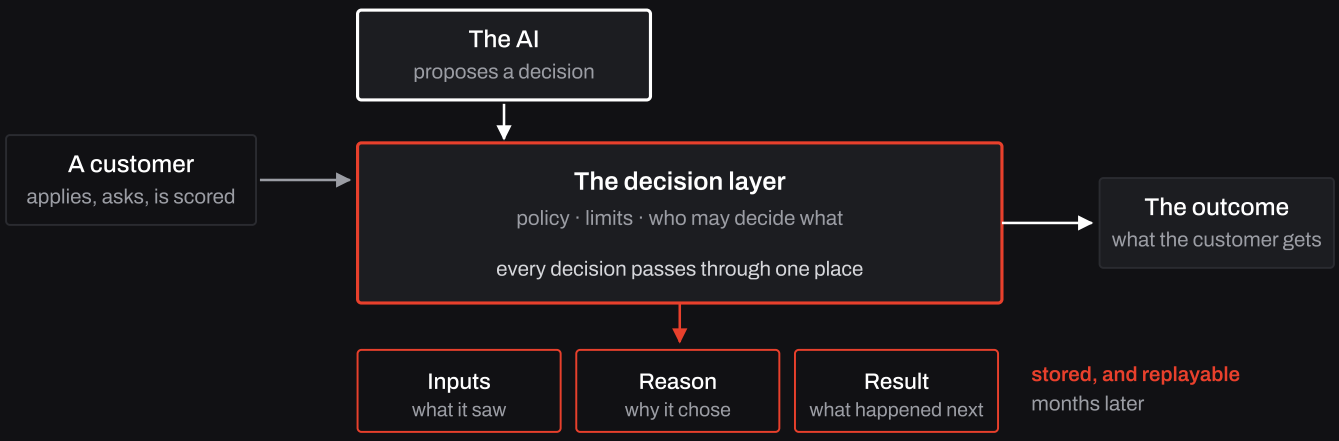
Instrument and questions: Rick Worthington

Analysis and prose: Agent

The question underneath the AI question

More of the decisions a company makes about its customers are going to be made by AI. The argument worth having is not whether that happens — it is what those decisions get made *on top of*.

A **decision layer** is the thing between the model and the customer: the place where a proposed decision is checked against policy and limits, applied, and — the part most architectures leave out — written down. Every position in this report is about that object, so it is worth seeing before the evidence starts.



Without the layer: the arrow at the top reaches the customer directly, and nothing is written down.

The record is the part that is usually missing. A model that cannot be replayed months later cannot be audited, corrected, or defended — which is what makes it the requirement rather than a feature.

Nothing here says who builds it. That is the question the evidence answers, and it answers it *rent* — see the positions below.

The object the three positions are about. Without the layer, the model's choice reaches the customer directly and leaves nothing behind but its output. With it, every decision is made through one place that stores what it saw, why it chose, and what happened next.

Companies are buying the model and assuming the layer. That assumption is where the cost, the exposure, and the ability to answer *why did it do that* are all decided, and it is almost never the thing being evaluated.

Three positions come out of fourteen graded findings. Each is stated with what justifies it; the rest of the report is the evidence.

1 **Keep a decision layer under any AI making customer decisions, and rent it rather than build it** — on a contract that continues only while measured value clears the cost of the alternative.

Justified by: the rising accountability test (Finding 1), operations dominating lifetime cost (12), and the absence of any documented success for the build path (11, 14).

2 **Make an auditable decision record the non-negotiable requirement** of every candidate architecture — every decision stored with its inputs, reason, and result, replayable months later.

Justified by: what companies actually disclose as the accountability test — outcomes and regulation, near-zero mechanism (Finding 2) — and by agents winning the vocabulary while writing to nothing (5).

3 **Measure the existing baseline before anything is signed** — current decisioning performance, deployment cadence, and hands-on versus elapsed time.

Justified by: stated reliance on AI running ahead of stated evidence at every level of disclosure (Finding 3), and the one decisive metric being measurable only internally (4).

The rest of this document justifies those three positions. Every finding carries an evidence grade, a citation to its dataset or source, and — where a number failed adversarial re-analysis — a note in Removed & limits saying exactly what died and why. Read the findings in any order; the navigation follows you.

HOW THIS WAS MADE, AND WHAT FAILED

Fourteen graded findings worked from four evidence bases and twenty-four cited sources, every number traced to a stored query, then adversarially re-analyzed before publishing. Three findings from earlier revisions did not survive that pass — including a persistence test whose result proved statistically indistinguishable from blockchain's. All of them are kept in Removed & limits rather than deleted. The full account is in Provenance.

Why you can trust this

This is not a chatbot's research summary. It is the output of a standing measurement instrument — and everything in it can be audited.

A fair first reaction to any AI-era research document is that someone typed a clever prompt into a model and formatted the answer. That is not what this is, and the difference is checkable. **The numbers here are query results, not model opinions.** Raw primary documents — SEC filings, software-contribution events, academic papers, price catalogs — were collected from their original sources into a permanent analytical database, and every figure attributed to this analysis is produced by versioned query code that can be re-run against that stored data, by anyone, for any audit.[1]

18,634
SEC filings from 530 public companies, each traceable to its EDGAR accession number

29,287
passages where those companies discuss AI under legal liability, 2023–present

141M
public software-development events behind the practitioner panel

66,704
practitioners tracked month-by-month by what they contribute code to

290,240
academic papers mapped to a shared concept vocabulary

406
AI models price-snapshotted daily — a series that exists nowhere else

Where a language model was used, and where it was not. An LLM plays exactly one role in this pipeline: classifying filing passages into categories (does this passage state a dependency? quantify an outcome?). That step was treated as an instrument to be calibrated, not trusted: the classifier is an open-weight model pinned to a fixed prompt version, chosen by measuring candidate models against a frontier reference, and its noisiest term family was hand-audited on a 300-passage sample — **79% precision, published, and corrected for wherever**

> **it is used**.[2] Every other number is deterministic SQL over stored data. No finding in this document is a model's summary of the internet.

Three disciplines stand behind every published number. **Trends are re-run within each stratum**, because a trend across a mixed population can be pure composition shift. **Every "companies that do X also do Y" claim is re-tested within bands of how much each company writes**, because verbose filers mention everything more. And on August 12, **an adversarial re-analysis re-ran every load-bearing number from raw data with instructions to destroy it**.[4] Three findings from the previous revision failed and were removed — they are named, with cause of death, in Removed & limits. What remains survived the attempt.

How the instrument was tested

No sensor was trusted because it returned data. Each one was audited first — and the audits changed what could be claimed, killed one analysis outright, and refuted this project's own first headline.

Any measurement instrument is a claim about the world before it is a source of numbers, and the failure mode is specific: a broken sensor rarely returns nothing. It returns something plausible. So each source below was tested against something outside itself before any finding was allowed to rest on it, and the tests are listed here whether they passed or not.[19]

3	0.96×	1.03×	18/18
independent confirmations required before accepting that a data source had been withdrawn upstream	a control that had to come out near 1.0, and did — the check that validates the arithmetic around it	what this project's own first headline measured out-of-sample, against 8.7× in-sample. Published as a refutation	filing sections split correctly in a hand-audited pilot before the corpus was sectioned at scale

300

passages hand-audited to measure the classifier's noisiest term family — 79% precision, corrected for

3

findings killed by the adversarial re-analysis, each named in this report with its cause of death

THE UPSTREAM STREAM WAS AUDITED BEFORE IT WAS USED Signal-bearing event types in the public software-development stream fell about **85%** between October 2025 and June 2026 — while the monthly total held flat near 110M, because push events grew 56% and absorbed the difference. Every day was present and the decay was smooth, so no coverage check trips on it, and a naive trend computed across that boundary would have reported that every project on earth was dying. Confirmed against the archive upstream of this project's own collection, establishing it as a real collection artifact rather than a loading error here. The permanent consequence: the usable window ends around March 2026, and every score is a **ranking within a period**, never a level compared across periods.[19]

IDENTITY WAS TESTED, NOT ASSUMED A project's name is not its identity. **124 of 1,840** repositories above 700 stars turned out to be aliases, and the rename rate rises with success — 10.3% above 10,000 stars against 5.9% in the 700–2,000 band. One project renamed twice in four days *during* a viral takeoff; keyed on its raw name it reads as three unremarkable mid-tier repositories instead of the largest single story in the window. Detecting renames structurally — one project's activity dying as another's is born — was tried first and rejected as far too noisy. The rule kept: structure may propose a candidate, the API decides.[20]

A CHECK THAT HAD TO COME OUT BORING, AND DID The crowd measure was run against the very project that defined the crowd — a number that *must* land near 1.0, because a crowd cannot discover what it has already discovered. It scored **0.96×**. A pipeline with an error in it would have had no particular reason to land there. That the null case came out null is the strongest single reason to trust the arithmetic that surrounds it.[20]

A SOURCE WAS WITHDRAWN, AND IT TOOK THREE INDEPENDENT CHECKS TO BELIEVE IT

The original sensor was public star data. It stopped existing mid-2026 with no announcement, and "my pipeline broke" looks identical to "the source is gone" from inside a pipeline. Three checks, chosen to fail differently: the live event stream (star events fell from ~2M a month to ~20k while total events held steady), the historical archive sampled at the same hour across months (thousands per hour in March, roughly fifty by August, pushes flat), and the API endpoints themselves (star listings return not-found while the neighboring fork endpoint still answers). Only then was the instrument re-sensored.[21]

THE REPLACEMENT SENSOR WAS SCORED AGAINST GROUND TRUTH — AND THE FIRST SCORE WAS WRONG

Validated against the one month for which full firehose data was held, the new contribution sensor initially scored **74%** recall. Investigating the apparent misses rather than accepting them: **62 of the 91** were projects held under a stale name. The sensor was not failing; the labels were. Resolved first, recall is **90%**, and the new sensor surfaced 226 person-project pairs the old firehose never had. A validation number is itself a measurement, and it can be wrong in the direction that makes you abandon something that works.[12]

THE JUDGE WAS GRADED BEFORE ANYTHING IT JUDGED

Classification uses a cheap open-weight model checked against a frontier reference. Before that, the reference was run twice on the same sample to see how often it agreed with *itself*: **92%** on the load-bearing field — a real ceiling, now known — and **60%** on another. That second field was not given a better prompt or a bigger model. It was **dropped from the analysis**, because a question the best available judge cannot answer consistently is not a measurement.[2]

THIS PROJECT'S OWN FIRST HEADLINE WAS REFUTED BY ITS OWN TEST

An earlier finding said a crowd of practitioners spotted breakout projects **8.7×** more often than chance before takeoff. It survived every in-sample check. Then the method was frozen and pointed at a later window it had never seen, with a control matched on *topic* as well as activity — and the earliness it claimed to measure came out at **1.03×**, indistinguishable from nothing. It was published as a refutation rather than quietly dropped. What survived is the part that replicated: the crowd is a 2.57× population filter, and it is not early.[22]

AN ANALYSIS WAS KILLED FOR BEING TOO THIN TO PUBLISH A planned cross-reference between the practitioner panel and social discussion found **13 of 2,990** post authors were panel members — 0.4%, a headcount rather than a trend, which more projects would not improve. The remaining collection work, roughly 35 hours, was canceled rather than spent producing a weak signal that would have looked like a finding.[20]

AN AUDIT PACK WAS WRITTEN FOR SOMEONE WITH NO CONTEXT On August 5 the whole instrument was documented for an outside second opinion: every figure queried live from the running system rather than recalled, a stated bias disclosure (it was written by the same agent that built the thing), and a section naming what its own author thought was weakest and most deserving of an outsider's skepticism.[20]

How it was made, and what it took

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

228
human prompts across the 15 sessions that built and ran the instrument, Aug 3–13, 2026, counted from session transcripts

11
days from an empty server to this report — database, four collectors, the analyses, and the writing

6
revisions — two that removed or downgraded findings, two that corrected numbers, one that added navigation, one that changed only framing

1
adversarial re-analysis: every load-bearing number re-derived from raw data with instructions to destroy it

$1.48
total classification cost for the filing corpus, open-weight model

$8.93
total spend on the instrument to date, against a self-imposed ceiling of $50 a month

The prompt count is the whole project, not the writing. It covers deciding which questions were worth asking, standing up the database, building and re-building the collectors, the validation passes listed above, the analyses that were killed, and the drafting — because quoting only the drafting sessions would describe a research document as though it were a blog post.[23]

Platforms. Collection and analysis run on a self-hosted stack: Python collectors feeding a ClickHouse analytical database, queried in SQL. Drafting, charting, and the adversarial re-analysis ran as Claude Code (https://claude.com/claude-code) sessions against that database. The report itself is hand-authored HTML/MDX — no generator, no template engine.

Models, by role. Three models touched this report, in three different jobs. *Claude Opus 5* and *Claude Fable 5* (Anthropic) did the research direction, drafting, chart construction, and the red-team pass — under human instruction at every step. *gpt-oss-120b* (open-weight, run via API) did exactly one mechanical job: classifying 18,200 filing passages against a pinned prompt, selected for that job by measuring candidates against a frontier reference and costing $1.48 where the frontier model would have cost $212. No model generated any number cited as a finding — every figure is a SQL query result.

Where AI was used — the full statement

Stated as method, not confessed as a caveat.

The prose of this report was written by an AI (Claude, Anthropic), working from query results, under human direction, across 22 prompts. The human — who owns the questions, the instrument, and every editorial judgment — reviewed each revision and ordered the adversarial pass that removed three findings the AI had drafted and initially defended.

The numbers were not produced by an AI. Every figure attributed to this analysis is the output of deterministic SQL over stored primary data. The single place a language model sits inside the measurement pipeline — passage classification — is treated as an instrument: pinned prompt, measured precision (79% on its noisiest family, hand-audited on 300 passages), error corrected for where it is used, and any field the reference model could not answer consistently (below 92% self-agreement) dropped entirely.

What that means for the reader: you are trusting the stored data, the published query code, and the stated calibration — not a model's recollection of the internet. All three can be audited; the report's citations say where.

Version history

This page renders Revision 10, unabridged. Earlier revisions are superseded, not hidden — what each one lost is listed in Removed & limits.

REV 1 · AUG 11, 2026 First full draft: 17 candidate findings from the four evidence bases. Two findings removed the same day under standard controls (verbosity, composition).

REV 2 · AUG 12, 2026 The adversarial re-analysis — every load-bearing number re-derived from raw data by a fresh session instructed to destroy the report. Three findings failed and were removed; two more were re-scoped. The surviving content is what this page shows.

REV 3 · AUG 12, 2026 Content unchanged from Revision 2 except where noted in the text. Adds navigation, per-claim citations, charts, and the provenance summary.

REV 4 · AUG 13, 2026 A framing pass only. Earlier revisions addressed a single organization in the first person; this one addresses any reader facing the same question. No finding, number, grade, confounder or citation changed — the evidence is identical to Revision 3.

REV 5 · AUG 18, 2026 A correction. Revisions 1–4 sized the practitioner panel at 66,704, which is the number of people its rolling recruitment has admitted — not the number whose contribution activity was ever collected. The collected cohort is 29,499. The finding, its direction and its grade are unchanged, because shares were always normalized by the active panel rather than by the headline count; what changes is the size of the population the reader was told stood behind it, and the confounder now states that 37,205 recruited members remain uncollected.

REV 6 · AUG 18, 2026 The panel is now whole, and a scoping error is corrected. The 37,205 members that rolling recruitment had admitted but never collected were collected — closing the gap Revision 5 disclosed rather than living with it — and Finding 8 was re-derived from the full panel rather than relabeled. Its three series moved: the two single-agent tools fell 10.9%→1.2% and 5.2%→1.6% on the old cohort and fall 9.1%→1.3% and 4.2%→1.5% on the whole one, while fleet infrastructure rose 0.03%→2.45% before and 0.05%→2.7% now. Direction, grade and conclusion are unchanged; the falls are slightly shallower and the rise slightly larger, measured on 2.3× the people. The aging-cohort confounder is retired because the cohort is no longer frozen, not because it became inconvenient. Separately, Finding 9's project counts were being taken over the whole label table rather than over the repositories the panel actually touched — a scope error present since Revision 1 and inherited unnoticed through four revisions. Corrected here, which moves those counts DOWN even though the panel grew.

REV 7 · AUG 25, 2026 A framing pass only. The report was retitled from "The decision layer" to the present title, the five evidence parts were re-headlined to state each part's takeaway rather than name its topic, the standfirst moved to the site's plain register, the index page now leads with the ambient claim this report measures, and the prose was swept to the house voice — first-person plural replaced with the report's own register throughout. No finding, number, grade, confounder or citation changed — the evidence is identical to Revision 6.

REV 8 · OCT 11, 2026 One sentence added to the platform direction in the close: a link to Field Report No. 5, which tested OpenAI's new decision model against Jev and found the

wording of the question moved scores more than the vendor did. No finding, number, grade, confounder or citation changed.

REV 9 · OCT 14, 2026 A correction. Finding 12 attributed its "60–80%" lifecycle figure to a 2026 academic cost model. That paper restates a widely cited estimate (40–80%) from studies and textbooks published between 1980 and 2002; it does not estimate the figure itself. The finding now says so and adds the closest current measurement, a 2025 federal audit that found 79% of the largest agencies' planned IT spending going to operations and maintenance. Recommendation 2 drops the range. Direction, grade and conclusion are unchanged.

REV 10 · OCT 15, 2026 One sentence added to Finding 14: a link to Field Report No. 6, whose later search found the same missing record for business software and one exception in infrastructure. No finding, number, grade, confounder or citation changed.

Every revision has removed or downgraded something. None added reach. What each one lost is named, with cause of death, in Removed & limits.

PART A

Companies now answer for AI decisions they cannot yet measure

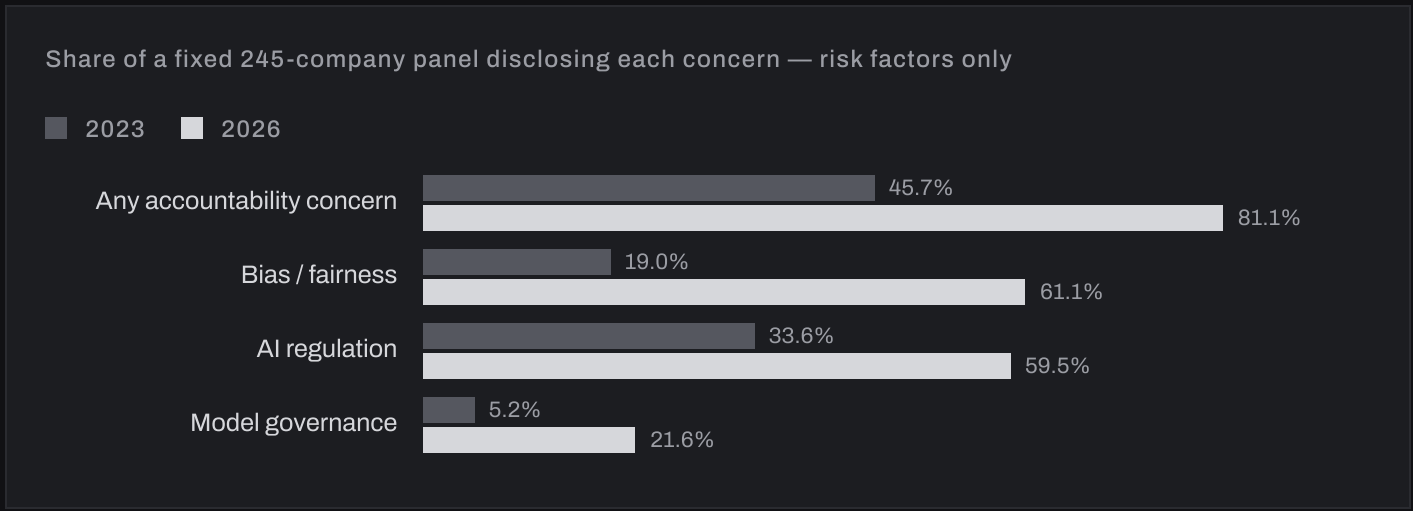
SEC filings are the one venue where a company describes its AI position under legal liability. What is rising there — and what is absent there — defines the test any architecture must pass.

FINDING 1 STRONG EVIDENCE

AI accountability disclosure near-doubled in four years on a fixed panel

> **Takeaway: the compliance surface under AI decisions is widening every year. Build for the test tightening, not settling.**

On a fixed panel of 245 companies, within risk factors alone, the share disclosing an AI accountability concern rose from **45.7% to 81.1%** in four years; bias and fairness disclosure roughly tripled.[3] The panel is balanced (only companies present in all four years) and measured within a single filing section, eliminating both sample drift and section-mix effects. The adversarial pass rebuilt the trend from scratch with independently written term lists on a stricter panel and got the same shape and steepness.[4]



Governance is now the most frequently raised AI topic among these companies — more common than any capability topic. Regulated sectors lead (bias disclosure 72.6% against 53.3% elsewhere), but the unregulated line rises at a similar slope: the gap buys timing, not a moat.[3] One interpretive caution, stated rather than hidden: risk-factor language spreads partly through counsel copying other filers, so this measures what companies will formally state as exposure under liability — not board attention directly.

DIRECTION

Treat AI accountability as a disclosed, board-level exposure. Being early to a control framework everyone will eventually need is cheaper than retrofitting one after an examination.

FINDING 2 CORRECTS A COMMON ASSUMPTION

The disclosed test is outcomes and regulation — almost never model mechanism

Takeaway: the durable requirement is an auditable decision record — it satisfies the outcome test the market discloses and the reason-giving test an examiner applies.

58 / 29,287 passages in four years of corporate AI disclosure use any explainability word. Discrimination, fairness and regulation language appears in the thousands. [5]

Companies frame AI accountability as an *outcome* question — did the system produce a discriminatory result, is a regulator about to act — not as a mechanism question about explaining the model. The adversarial pass probed the adjacent vocabulary the term list might have missed ("adverse action": 13 passages; "model risk": 19; "audit trail": 19; human-in-the-loop: 79): all equally negligible. The absence is real, not a word-choice artifact.[5]

The claim is scoped to disclosure, deliberately. In regulated credit, mechanism-level obligations exist — adverse-action notices demand a stateable reason for an individual decision — and filings would not use that vocabulary either way. Both readings of the requirement are satisfied by the same asset: every decision stored with its inputs, its reason, and its result. That satisfies the examiner, and it is also the raw material for improving the next decision.

DIRECTION

Make the auditable decision record a hard requirement of any architecture under consideration — bought, built, or agent-based — and assess each candidate on whether it produces that record natively or leaves the buyer to construct it.

PART B

Dependency is announced faster than it is measured

Whichever architecture wins, somebody has to prove it worked. Almost nobody's disclosure does.

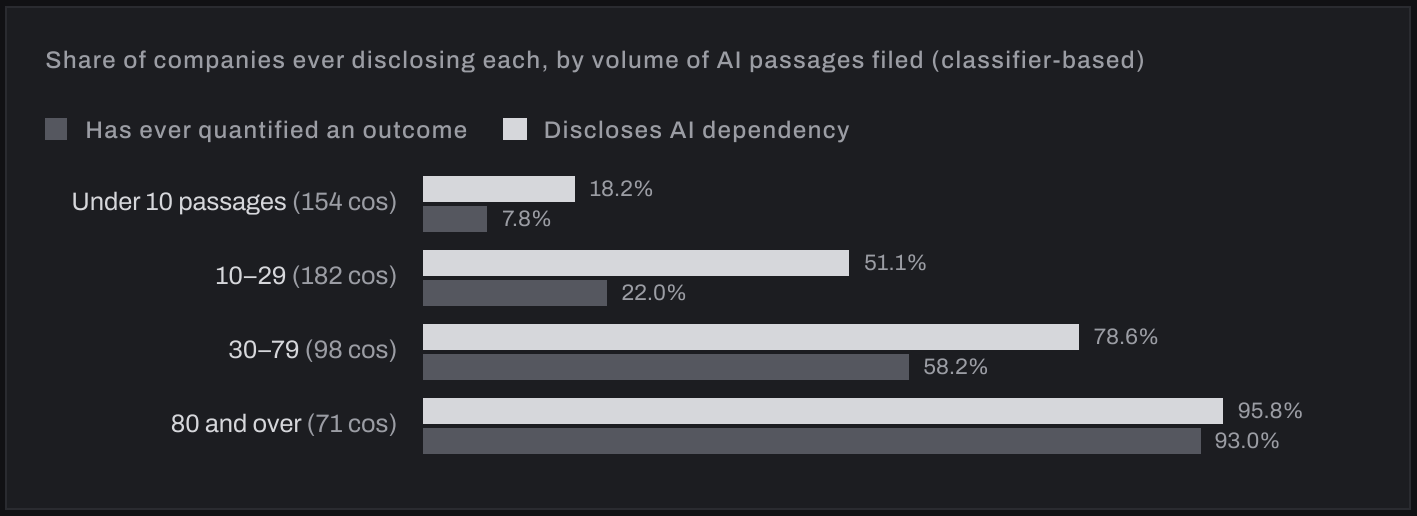
FINDING 3 STRONG EVIDENCE

Dependency on AI is disclosed faster than it is quantified — at every level of disclosure volume

> **Takeaway: across the market, stated reliance on AI runs ahead of stated evidence for it. An organization that can put a number on its AI is making a claim most peers cannot.**

Band companies by how much they write about AI at all — the control that keeps verbose filers from driving the result — and in every band, more companies disclose a *dependency* on AI than have ever tied AI to a *number*.[6] An earlier, simpler version of this finding ("two thirds have never quantified AI") was removed in Revision 2: the raw share

tracks how much a company writes, not how it manages. This banded form is what survives that control.



Two honest notes. Among the heaviest disclosers the gap nearly closes — companies that talk most about AI do eventually attach numbers. And quantifying in a filing is not the same as measuring internally: counsel strips numbers from filings for liability reasons, so this measures disclosed proof, not private practice.

DIRECTION

Make measurement the first deliverable, not the last. Agree the metric, the baseline and the method for current decisioning performance before anything is signed or built.

FINDING 4 NOT MEASURABLE EXTERNALLY

Deployment velocity — the metric that matters most — cannot be measured from outside

> **Takeaway: the decisive number in this debate is one only the buyer can produce. It has to be instrumented internally, before the vendor conversations start.**

Across 29,287 corporate AI passages, only **16** mention deployment frequency or release cadence; "time to market" (205 hits) is dominated by semiconductor firms, where it means chip tape-out — a different construct entirely.[7] A proxy could have been built from

the 7,000+ passages of vague speed language. It would have produced a confident number that flattered this report's position while measuring marketing register. This report refused it, and reports the null instead.

The underlying question is still the right one: how many times a year can an organization change a customer decision, observe the result, and change it again? Weekly cadence is 52 chances to learn per year; quarterly is four. That difference compounds, and no external dataset sees it.

DIRECTION

Track, for decisioning changes specifically: elapsed time from idea to live, hands-on time within it, and changes deployed per quarter. Hands-on over elapsed exposes coordination cost, and it is unarguable in a way no market study can be — because it is measured in-house.

PART C

The market's language has moved to agents

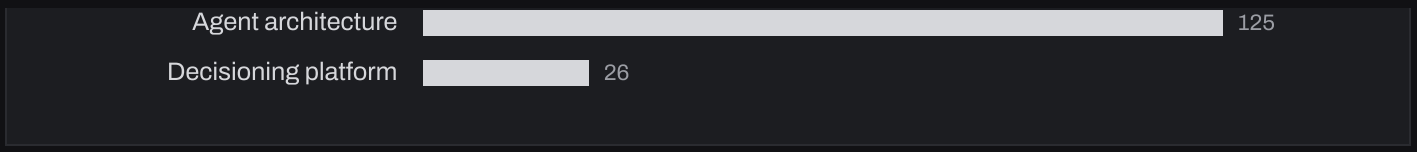
Filings describe language, not deployed architecture — this revision keeps that distinction strict. What the language shows is still worth having.

FINDING 5 STRONG EVIDENCE

Agents won the vocabulary — including among the decisioning vendors themselves

Takeaway: don't argue against the word "agent" — that argument is lost. Compete on the property: an agent that writes to a decision record is decisioning with a new interface; one that writes to nothing is unaccountable.

Public companies naming each architecture in filings, last twelve months



Among decisioning vendors, agent language now outweighs decisioning language two to one in their own filings — Pega, Appian, ServiceNow and Salesforce all sell "agentic" now.[8] The adversarial pass re-ran the split with a substantially widened decisioning vocabulary (recommendation engines, credit decisioning, underwriting models, risk scoring); the asymmetry barely moved.[4] Note what the vendors did *not* do: they repositioned toward agents while continuing to describe and sell decision infrastructure underneath. On the sell side, this is a layering story.

What this report can no longer claim, and says so. An earlier revision reported a within-company test suggesting enterprises kept their decision layers as they adopted agent language. On re-analysis, the companies describing decisioning in filings were overwhelmingly the companies that *sell* it, plus keyword accidents; genuine buy-side enterprises were too few to test. Whether enterprise buyers are layering or replacing is not answerable from filings — in either direction. Details in Removed & limits.

DIRECTION

Concede the terminology, and reframe the architecture debate as a requirement: any agent making customer decisions writes to an auditable decision record, or it does not ship.

FINDING 6 MIXED EVIDENCE

The decisioning category's own market signal is contested — in the same quarter

> **Takeaway: the market is re-pricing this exact question right now. That is negotiating room for a buyer with its own baseline — and a reason to trust no one's narrative, this report's included.**

One pure-play's cloud subscription line grew **19%** in 2025 and accelerated to **25%** and **23%** in the first two quarters of 2026 — its strongest in two years, with management crediting AI adoption for the wins.[9] The other's first-half revenue fell **1%**, its total contract value growth slowed to **7%**, its shares fell 19% on the report — and management attributed the stall to "unprecedented changes in the AI market" causing clients to **delay purchasing decisions**.[10] An earlier revision reported only the growing lines and called the weak half "license timing"; the vendor's own explanation is quoted instead, and it is not a timing story.

Read straight, this cuts both ways. The category is not being hollowed out — one of its two public pure-plays is accelerating through the third year of the AI boom. Nor is it serenely compounding — the other's own management says buyers are hesitating *because of* agent-era uncertainty. Two companies are not a category, and neither quarter settles the argument.

DIRECTION

Accept neither "legacy category" nor "safe category" as a premise. Use the sell-side uncertainty to obtain the value-contingent contract structure in Recommendation 5 — it is easiest to get from a vendor that needs to prove its category.

FINDING 7 PUBLIC RECORD

Independent analysts now rank decisioning-heritage vendors alongside the general-purpose AI platforms

> **Takeaway: "serious AI companies versus legacy workflow vendors" is an out-of-date framing, retirable with a public citation.**

In a major analyst firm's 2026 evaluation of AI platforms, a vendor whose heritage is enterprise decisioning was placed in the **Leaders** group — publicly reported as scoring second on strategy across the field, behind only one hyperscaler.[11] The same firm

characterizes that vendor's best fit as complex, *regulated* workflows executed *using AI agents* — the layered model, not an alternative to it. Two qualifications: several major participants declined to take part, so read the ranking as directional; and the same analysts warn against single-platform strategies — a caution against buying one platform for everything, not against buying the right platform for the decision layer.

DIRECTION

Scope any proposal deliberately: the decision layer for regulated customer decisioning, not a general-purpose AI platform for the enterprise.

PART D

Fleets are arriving faster than the tools to run them

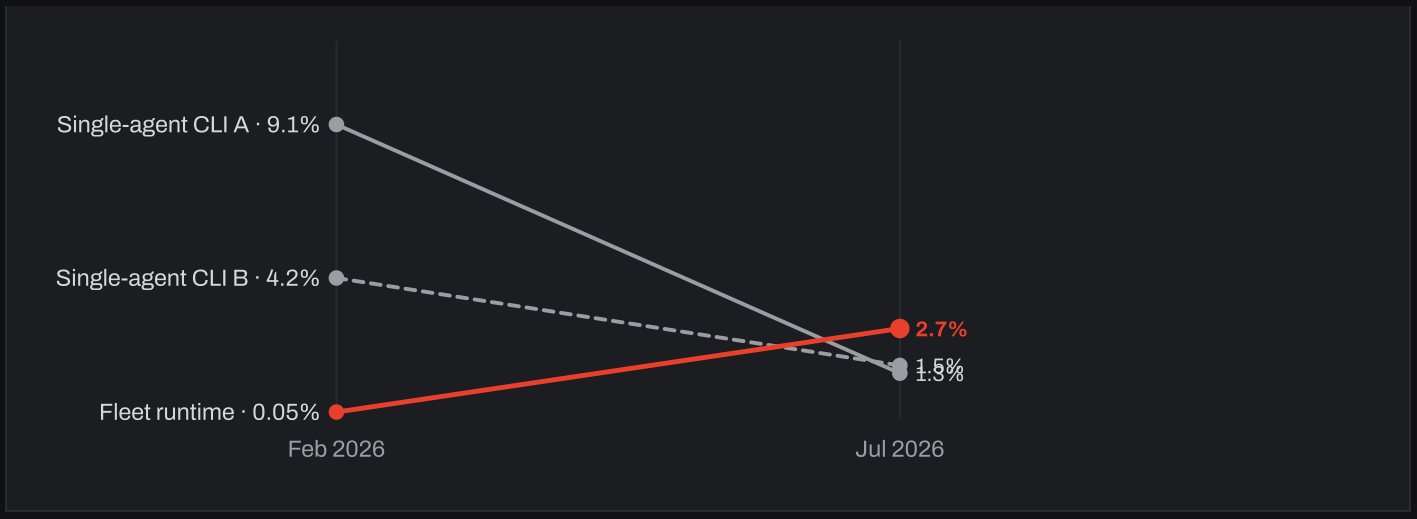
Practitioners run ahead of enterprises by roughly eighteen months. Their behavior shows where the open problems are — which is not the same as where the users are.

FINDING 8 STRONG DATA, TWO READINGS

Builder contribution moved from single-agent tools to fleet infrastructure

> **Takeaway: whichever reading is right, the unsolved engineering is now in the layer that runs *many* agents — plan for fleet operations as a real function.**

Share of active-panel contribution attention, February → July 2026



Across a panel of 66,704 practitioners, normalized by the active panel, the single-agent tools that dominated early 2026 collapsed as a share of contribution attention while infrastructure for running *many* agents rose from near zero.[12] What rose describes itself as runtimes, fleet environments, coordination layers — things you need only after you stop evaluating one agent and start operating several.

Two confounders, stated. The panel is designed to roll — recruiting each month from the projects the crowd concentrates on — and as of Revision 6 every member it has ever admitted has been collected across the full window, so newcomers are represented rather than missing. What remains is that the active population still moves month to month, which is why shares are normalized by it rather than reported raw. And **contribution is not usage**: contributors go where the open problems are, so a single-agent tool that matured, or whose category was won by closed-source products invisible to public contribution data, produces the same falling line as one being abandoned. Both readings land in the same place operationally: the open problems moved up a layer.

DIRECTION

Whatever gets bought or built, assume that within two to three years the organization running it is supervising many automated decision-makers at once, with fleet operations — deployment, versioning, monitoring, rollback, cost control — as a staffed function.

FINDING 9 MODERATE EVIDENCE

The tooling for supervising those fleets is fragmented and unsettled

> **Takeaway: the operating model is arriving before its standard implementation. Adopt agent capability through a layer that absorbs the churn.**

In the same practitioner ecosystem the data counts **391** distinct projects addressing agent harnesses, **190** for multi-runtime operation, **134** for supervision burden and **81** for fleet operations — no project close to dominant.[13] Every hot category breeds projects, and no cross-category base rate exists yet, so the counts carry less weight than the absence of consolidation. Hundreds of competing answers with no winner is what a category looks like *before* it standardizes — container orchestration and data warehousing looked the same at the equivalent stage, and early bets on the wrong answer were paid for twice.

DIRECTION

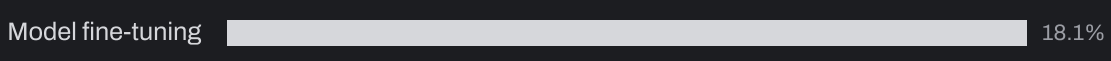
Do not wire decision logic directly to whichever agent framework is current this year. A platform that absorbs that churn is a benefit worth naming in the business case — it is invisible until the second migration. When OpenAI launched its own decision model, the wording of the question mattered more than the vendor.

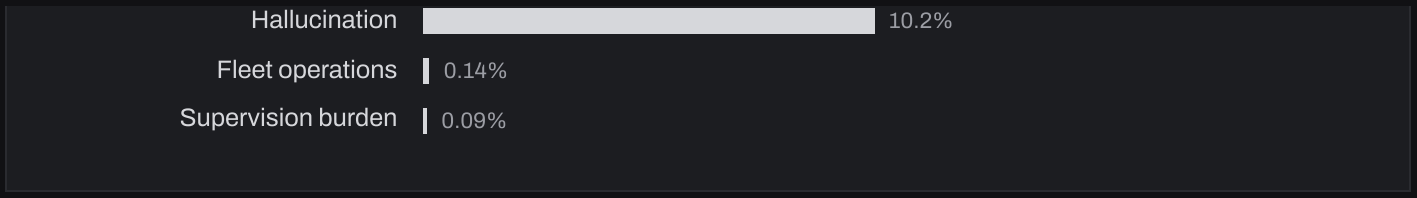
FINDING 10 STRONG EVIDENCE

The research community has barely started on the operational problems

> **Takeaway: there is no peer-reviewed playbook for running agent fleets safely at scale. Anyone claiming a known-good pattern is describing a roadmap.**

Share of recent AI-tooling papers by topic, May–July 2026





The problems practitioners spend their days on — running many agents reliably, supervising them affordably — are effectively absent from the literature that would normally say how to do them well.[14] Two qualifications: the concept vocabulary here predates fleet operations as a named topic, so part of this is plausibly under-detection; and peer review lags practice structurally, so part is latency. Neither changes the operational point — the guidance does not exist *today*, when the decision has to be made.

DIRECTION

Weight this in the risk assessment, not the technology assessment. Being early in a domain with no literature is legitimate — done knowingly and resourced accordingly, never assumed safe because it is fashionable.

PART E

Buy versus build: the evidence leans rent, and it is thinner than it looks

The honest section header: this is where the evidence is weakest, on both sides — and it is graded here rather than quoted.

FINDING 11 CONTESTED EXTERNAL STUDY

The most-quoted study favors buying — and is too weak to lean on

> **Takeaway: no number from this study belongs in anyone's business case. The defensible statement is smaller: there is no documented base of successful internal builds, so the burden of proof sits on building.**

MIT's *GenAI Divide* (2025) reported that **95%** of enterprise AI pilots produced no measurable P&L impact, with purchased solutions succeeding materially more often than internal builds. Its methodology — 52 interviews, 153 survey responses, a six-month ROI window, underlying data unpublished — has been publicly and credibly criticized, and secondary reports state its build-versus-buy multiple inconsistently. The criticism is cited here rather than left to be discovered.[15]

What it remains good for, used carefully: its *diagnosis* of why pilots stall — tools that cannot retain feedback, adapt to context, or improve over time — describes a missing learning loop, which is what a decision layer exists to provide. A coherence argument, not proof.

DIRECTION

Assess every proposal — including the one this report argues for — against the base rate that most enterprise AI initiatives demonstrate no measured value, whatever the exact figure. Spend scarce engineering on what is genuinely proprietary: the data, the integration, the measurement.

FINDING 12 OLD ESTIMATE, ONE CURRENT AUDIT

Most of the cost of enterprise software is running it, not writing it

> **Takeaway: evaluate any build on five-year total cost of ownership, never on build cost or time-to-first-demo — the majority of the money is in operations and is routinely omitted.**

A widely repeated estimate puts **40–80%** of software lifecycle cost after launch, in operations and maintenance. It traces to studies and textbooks published between 1980 and 2002, not to a modern measurement.[16] The closest current measurement points the same way: a 2025 federal audit found **79%** of the largest agencies' planned IT spending going to operations and maintenance of existing systems — an annual budget split rather than one system's lifetime, but measured.[24] A 2026 academic analysis

restates the estimate and places regulated, auditable, mission-critical systems firmly in the "buy" category — because certified compliance track record is the expensive part.[16] It bears directly on "AI makes building cheap": cheaper development addresses the smaller share of lifetime cost, while AI-generated code adds new cost categories of its own (governance of generated artifacts, quality assurance, compliance documentation). The paper is conceptual by its author's own statement — no case studies, no measured outcomes — and is weighted as a reasoned model from a source with nothing to sell, not as empirical proof.

DIRECTION

Require any build proposal to estimate operating cost explicitly: support, compliance evidence, model governance, upgrades, staffing, across a five-year life.

FINDING 13 CAUTION

Nearly all published ROI evidence in this category was paid for by the vendors

> **Takeaway: import no vendor's return figure. Replace it with a contract term: the relationship continues only while the buyer's own measured return clears the cost of the alternative.**

The headline returns circulating for decisioning platforms — including a widely quoted 442% ROI — come from vendor-commissioned studies with vendor-selected reference customers.[17] Not necessarily false; inadmissible as independent evidence, and any reviewer will find the commissioning line quickly. Cost comparisons deserve the same skepticism in reverse: three-year TCO studies putting decisioning platforms far above lighter alternatives are largely published by the lighter alternatives. Both are marketing; neither anchors a case.

DIRECTION

Measure the return on a baseline you produced yourself — converting the weakest evidence in the category into the strongest term in the contract.

FINDING 14 OPEN QUESTION

No published record exists of a company successfully replacing a bought platform with its own

> **Takeaway: "buy now, build in-house later" is the one strategy with no supporting evidence in either direction. Preserve the exit as an option; never commit to it as a plan.**

Both directions were searched — organizations that failed to build and then bought, and organizations that bought and then successfully insourced. Neither is documented to a standard that survives scrutiny, and the reason is structural: purchases are material and must be disclosed; replacing a vendor with internal engineering is not a reportable event. In this corpus, "reducing third-party reliance" language is almost entirely tariffs and supply chains. A later search, in I heard software was dead, found the same gap for business software and one exception in infrastructure.[1]

DIRECTION

Preserve the option contractually and architecturally at signature, when it is cheap: the decision log in the buyer's own store in an open format; rules and configuration exportable in a documented schema; the renewal test defined before go-live.

What the evidence supports

Seven positions for anyone facing this decision, each traceable to findings above — none resting on a number that failed re-analysis.

Put the burden of proof on removing a decision layer, not on keeping one. Whether buyers are keeping or replacing theirs is not externally knowable (5), the success case for skipping the layer is undocumented (11, 14), and the accountability test any replacement must pass rises yearly (1).

Weight buying over building — for the honest reasons. Not a failure-rate multiple, which the evidence will not support, but most of lifetime cost sitting in operations where certified compliance is the expensive asset (12), and no documented base of successful internal builds to price against (11, 14).

Make the auditable decision record non-negotiable in any architecture, bought or built. It satisfies the outcome test the market discloses, the reason-giving test the examiner applies, and it feeds the next decision (2).

Establish the decisioning baseline before anything is signed — including deployment cadence and hands-on-to-elapsed time, which no external source can supply (3, 4).

Make the commercial relationship contingent on measured value, importing no vendor ROI figure — and use current sell-side uncertainty as the negotiating room to get that structure (6, 13).

Preserve the exit as an option, not a plan: the decision log in the buyer's own store, exportable rules, the renewal test defined before go-live (14).

Adopt agents through the layer, not around it, and budget fleet operations as a real function — the operating model is arriving, its tooling is unsettled, its literature is a year behind its practice (8, 9, 10).

WHAT THIS REPORT IS NOT CLAIMING

That agent architectures will fail — the contribution data reads at least as well as maturation. That enterprises at large kept their decision layers — that was tested, and the data cannot say. That AI should be kept away from customer decisions — it should not; the disagreement is about what it stands on. That any particular vendor

is the answer — selection is a separate exercise. Or that buying is risk-free: it is a different risk, better documented, whose category is itself being re-priced this quarter.

METHOD

Removed & limits

Findings investigated and not published — including three removed after the adversarial re-analysis. A report that shows its dead is harder to ambush than one that buries them.

REMOVED · REV 2 **"Companies kept their decision layer" (a 14-of-16 within-company test).** The cohort of companies describing decisioning early was dominated by decisioning *sellers* describing their own products, plus keyword accidents (an oilfield-services firm's "real-time decision" is about drilling). Genuine buy-side members: fewer than five. A placebo test sealed it: filing language persists at 63–100% for arbitrary technology terms, and decisioning's 87.5% persistence is statistically indistinguishable from *blockchain's* 77.3%. The claim measured drafting stickiness, not architecture.[4]

REMOVED · REV 2 **"AI vendors underwrite compute; buyers don't" (company-level).** The gap vanished inside verbosity bands, and the vendor cohort contained chip and GPU-cloud makers, for whom "building compute" is definitionally true. The underlying intuition — unit price and total bill are different lines — stands, and the daily price sensor[18] now collects the data to test it properly.

REMOVED · REV 2 **"Two thirds of companies have never quantified AI" as a headline level.** The share tracks disclosure volume almost mechanically (8% of thin filers have ever quantified; 93% of heavy filers have). Only the banded gap in Finding 3 survives.

REMOVED · REV 1 **A correlation that confirmed the report's own position.** Data-quality work appeared to predict quantified outcomes (42.3% vs 16.8%); under a verbosity control it vanished (20.0% vs 19.2%).

REMOVED · REV 1 **A trend that was pure composition.** "Quantified AI claims are falling" was a shift in section mix; within every section the rate was flat.

REFUSED **A velocity proxy** built from vague speed language — it would have flattered this report's position and measured nothing (Finding 4).

Bias/fairness term family: 21% false-positive rate, measured (300-passage hand audit, 79% precision). Levels are overstated by about a fifth; trends survive because precision is flat across years and identical across regulated/unregulated groups.[2]

Cohorts are assigned by disclosure language, not verified deployment — throughout. This document therefore claims nothing about what any buyer has deployed, only what companies state under liability.

2023 coverage is thin and 2026 is a partial year — which is why every trend uses a fixed panel. A flat 2026 line may be truncation rather than a turn.

Practitioner findings carry stated confounders — vocabulary lag, contribution-versus-usage — listed in Findings 8–10 where they apply. Panel aging was one of them until Revision 6 collected the whole roll.

Licensed analyst research is referenced only through publicly announced results. No proprietary scoring, wording or figures appear here.

APPENDIX

Citations

Two kinds: **Data** — this project's own datasets and query code, stored, versioned, re-runnable for audit — and **Ext** — external sources, graded where used.

DATA The boardroom corpus: primary filings fetched from SEC EDGAR for a 530-company panel (S&P 500 constituents plus 30 AI-sector names), January 2023 – August 2026. 18,634 filings, sectioned (Business / Risk Factors / MD&A / earnings exhibits) into 29,287 AI passages. Every passage is traceable to its EDGAR accession number; the panel itself is stored as data so any historical figure can be reconstructed against the panel that produced it. Collection code, section-split accuracy audit (18/18 Business and Risk Factors in pilot), and skip log retained.

DATA Passage classification: open-weight model (gpt-oss-120b) pinned to a fixed prompt version, applied to 18,200 classifiable passages. Model selected by measuring candidates against a frontier reference whose self-agreement was itself measured first (fields below 92% self-agreement were dropped from use). Noisiest term family (bias/fairness) hand-audited on a 300-passage sample: 79% precision, flat across years and identical across regulated/unregulated groups.

DATA Accountability trend: balanced 245-company panel, within-section (Item 1A) measurement, query code `accountability.py`, run Aug 11, 2026.

DATA Adversarial re-analysis, Aug 12, 2026: independent re-derivation of the accountability trend (independent term lists, stricter panel: 32%→88%), placebo persistence panel (mainframe, on-premise, blockchain, big data, metaverse), verbosity-banded re-runs, widened-vocabulary architecture split, and cohort membership audits. Full query outputs stored alongside the original analysis code (`REDTEAM.md`, `attack1.json`, `attack2.json`).

DATA Explainability scan: term-family counts across the full 29,287-passage corpus plus adjacent-vocabulary probes (adverse action, model risk, audit trail, human-in-the-loop), Aug 11–12, 2026.

DATA Dependency-versus-quantification: classifier dependency and evidence fields, company-level, banded by AI-passage volume, run Aug 12, 2026.

DATA Velocity vocabulary scan: deployment-cadence term families across the corpus, including the time-to-market contamination check (semiconductor/EDA filers), `cohorts.py`, Aug 11, 2026. Reported as a null.

DATA Forward architecture split: companies with ≥3 AI passages, trailing twelve months, agent-architecture vs decisioning-platform term families; vendor-cohort language shares from the same corpus, `cohorts.py`.

EXT Appian Corporation, Q1 and Q2 2026 results: cloud subscription revenue +25% and +23% year-over-year (Q1 (https://www.investing.com/news/company-news/appian-q1-2026-slides-cloud-growth-accelerates-to-2year-high-93CH-4668691), Q2 (https://www.investing.com/news/company-news/appian-q2-2026-slides-cloud-growth-hits-23-shares-surge-126-93CH-4847296)); FY2025 cloud growth 19% from company filings.

EXT Pegasystems Inc., Q2 2026 results and earnings call: revenue –1% year-over-year, total ACV growth 7%, management attribution of client purchase delays to AI-market changes (earnings call coverage (https://www.investing.com/news/transcripts/earnings-call-transcript-pegasystems-q2-2026-results-miss-as-shares-sink-19-93CH-4806035), results summary (https://www.tradingview.com/news/urn:summary_document_report:quartr.com:3671960:0-pega-revenue-declined-1-year-over-year-as-ai-market-volatility-slowed-growth-but-pega-cloud-rose-32/)).

EXT The Forrester Wave™: AI Platforms, Q3 2026 — publicly announced results only (Forrester announcement (https://www.forrester.com/blogs/the-forrester-wave-ai-platforms-q3-2026-is-live-prepare-to-recalibrate/), vendor announcement (https://www.pega.com/about/news/press-releases/pega-named-leader-ai-platforms-independent-research-firm)). No licensed content reproduced.

DATA Practitioner panel: 66,704 practitioners, seeded from 141M public GitHub events and extended each month by rolling recruitment, contribution activity collected monthly via the GitHub GraphQL API (pull requests and issues — star data was withdrawn by GitHub in Q2 2026 and is not used), February–July 2026, the sensor validated at 90% recall on a 500-account sample of the opening cohort against the one month of full-firehose ground truth available, never re-measured on the panel as a whole. Shares normalized by the active panel. Every recruited member has been collected across the full window as of Revision 6; 20,330 recorded any activity inside it. Revisions 1–4 stated the panel as 66,704 while only 29,499 had been collected; Revision 5 corrected the figure downward to the collected cohort, and Revision 6 restores it by collecting the remainder, which moved this finding's endpoints from 10.9%→1.2% and 5.2%→1.6% and 0.03%→2.45% to 9.1%→1.3% and 4.2%→1.5% and 0.05%→2.7%.

DATA Project classification: each repository labeled by a language model from its name, description, topics and language against a fixed prompt version; low-confidence labels excluded. Counted over repositories the practitioner panel actually contributed to inside the window. ⚠ Revisions 1–5 stated these counts over the whole label table, which includes repositories carried in from the earlier star-based collection that the panel never touched — an inherited scoping error, not a panel-size question, and the reason the figures fall despite the panel growing: 497→391 agent harnesses, 211→190 multi-runtime, 175→134 supervision burden, 87→81 fleet operations. Corrected in Revision 6. Ordering and the absence of a dominant project are unchanged under either scope.

DATA Research corpus: 290,240 arXiv papers, January 2023 – present, gated to the AI-tooling subset (a moving series, re-computed per window, not a constant), topic shares measured May–July 2026.

EXT MIT NANDA, *The GenAI Divide: State of AI in Business 2025* — publicly reported findings, cited together with published methodological criticism (BigDATAwire (https://www.hpcwire.com/bigdatawire/2025/09/04/mit-report-flags-95-genai-failure-rate-but-critics-say-it-oversimplifies/), Marketing AI Institute (https://www.marketingaiinstitute.com/blog/mit-study-ai-pilots)). Graded: contested; used only for its diagnosis and the burden-of-proof point, never for a precise figure.

EXT *The Buy-or-Build Decision, Revisited: How Agentic AI Changes the Economics of Enterprise Software*, arXiv:2604.26482, April 2026. Not vendor-affiliated; conceptual by its author's own statement. It restates the 40–80% lifecycle estimate as "a widely cited heuristic," citing Lientz & Swanson (1980), Boehm (1981) and Glass (2002); the figure is theirs, not the paper's own estimate. Graded: reasoned analysis, not empirical proof.

EXT Vendor-commissioned economic-impact and total-cost-of-ownership studies, identified by their own disclosed commissioning statements. Graded: inadmissible as independent evidence; cited only as a caution.

DATA Inference price sensor: daily snapshot of a 406-model public API catalog (prices, context windows, capability flags) into the same analytical store, begun Aug 11, 2026. Day-one medians: $0.55 per million input tokens (paid models), 84% tool-use support, 262,144-token median context window.

DATA Upstream collection audit, Aug 5, 2026: per-event-type monthly volumes across the public event stream, 2025-10 to 2026-06, cross-checked against the independent public archive of the same stream to separate an upstream collection artifact from a local loading error. The finding fixed the usable analysis window and the rankings-not-levels rule for everything built afterwards.

DATA Instrument audit pack, Aug 5, 2026 — written to be handed to a reader with no prior context, with a stated bias disclosure and a section naming the author's own weakest work. Contains the repository-identity and rename-rate audit (1,840 repositories, 124 aliases), the self-referential null check (0.96×), the sustainable-pacing measurements for each collector, and the decision to cancel the social cross-reference analysis at 0.4% panel density.

DATA Source-withdrawal confirmation, Aug 2026: three independent checks — live event-stream volumes by type, same-hour-of-day sampling of the historical archive across months, and direct probes of the affected and neighboring API endpoints — run before the sensor was replaced.

DATA Out-of-sample refutation, Aug 8, 2026: method frozen before the test window was touched; 46 training breakouts, 55 held-out breakouts; controls matched on activity band and, decisively, on topic. In-sample 8.7× against an activity-matched stranger; out-of-sample earliness 1.03× against a same-topic control; the 2.57× population effect replicates. Published in full as a refutation.

DATA Effort accounting: human prompts counted programmatically from Claude Code session transcripts, excluding tool results and sub-agent turns, across the sessions whose work is the

instrument (Aug 3–13, 2026). Transcript retention covers the project's entire life, so the count is complete rather than sampled.

EXT U.S. Government Accountability Office, *Information Technology: Agencies Need to Plan for Modernizing Critical Decades-Old Legacy Systems*, GAO-25-107795, July 17, 2025 (report (https://www.gao.gov/assets/gao-25-107795.pdf)): about $83 billion (79%) of planned fiscal 2025 IT spending at the 24 CFO Act agencies was intended for operations and maintenance, per the federal IT Dashboard. Graded: government measurement of an annual budget split; not a lifecycle share for any one system.