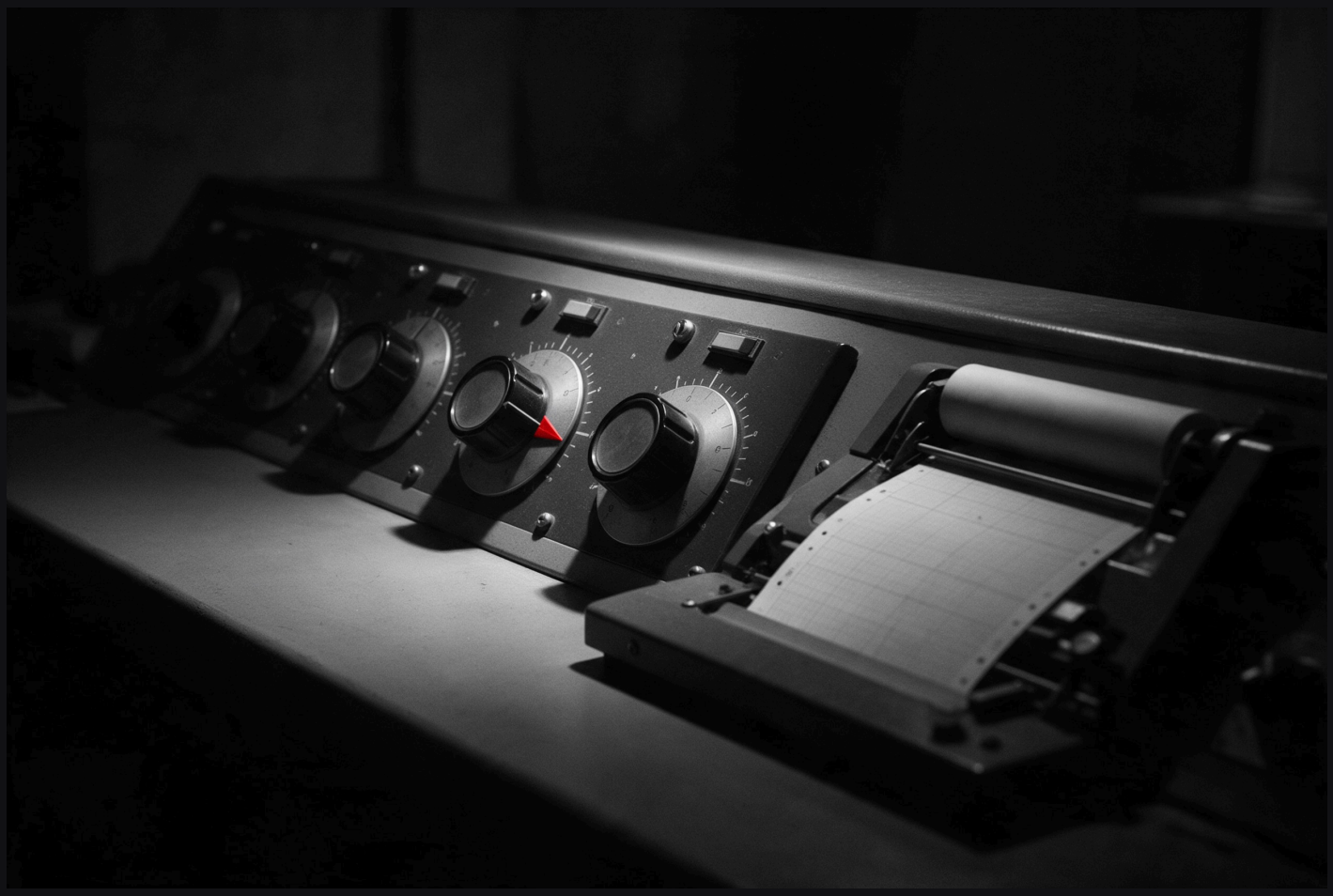


# Letting AI agents tune the system: adopted fast, measured almost never

There's a strategy being pitched everywhere right now: leave the decision engine deterministic and auditable, and put a team of AI agents above it, offline, to tune it between runs. This report reads the published evidence on what a company would actually get. The promise being made turns out to be the one thing nobody has measured.

AUG 19, 2026 · REVISION 4 · LAB.WORTHINGTON.CLOUD

# Letting AI agents tune the system: adopted fast, measured almost never

There's a strategy being pitched everywhere right now: leave the decision engine deterministic and auditable, and put a team of AI agents above it, offline, to tune it between runs. This report reads the published evidence on what a company would actually get. The promise being made turns out to be the one thing nobody has measured.

003 Field Report Aug 19, 2026 Revision 4

Instrument and questions: Rick Worthington

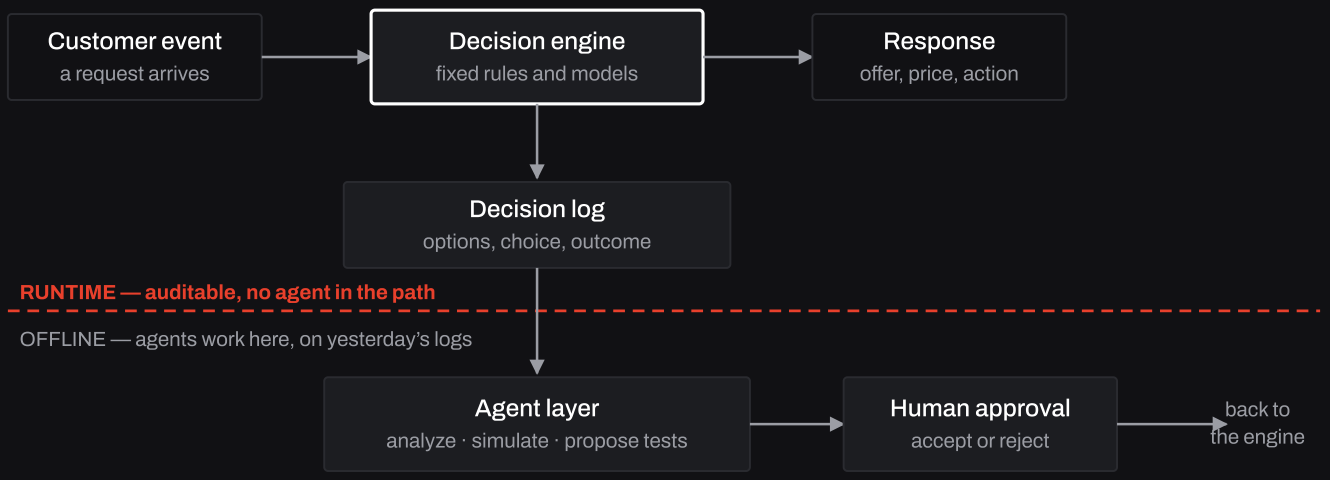
Analysis and prose: Agent

THE PROBLEM

## A strategy with a boundary in it

There is a specific proposal circulating about how a company should use AI agents on the decisions it makes about customers, and it is more careful than most.

The proposal goes like this. A decision engine — the rules-and-models system that picks which offer to show, which price to quote, whether to approve an application — stays exactly as it is: deterministic, inspectable, with no AI agent called while a customer is waiting. Every decision it makes is logged: what the options were, which one was chosen, what was suppressed and why, and what the customer did next. Then, entirely offline, a team of AI agents reads those logs. They look for patterns, run simulations, propose experiments, and suggest changes to the engine's configuration. A person approves or rejects each proposal before anything ships.




The boundary in red is what makes this strategy different from putting a model in the decision path: no agent is called while a customer is waiting, so the record of why a decision was made stays reproducible.

The strategy, drawn straight. Nothing at the moment of decision changes: a fixed engine answers the customer and writes a log. The agents live off to the side, reading those logs and proposing changes that a person approves before anything ships.

The boundary is the interesting part. Putting a language model inside the decision path destroys the property that makes a decision engine defensible: you can no longer reproduce why a particular customer got a particular answer. This proposal keeps that property and puts the intelligence somewhere it cannot contaminate the record. That is a genuinely thoughtful piece of design, and it is why the strategy deserves a real examination rather than a dismissal.

The examination is worth doing because the strategy is not hypothetical. In the filings of a 529-company panel of large public corporations, the number describing AI agents alongside decision-making or optimization language went from three companies in the first eight months of 2024 to 107 in the same window of 2026 [3]. That is one company in five, describing this to investors under legal liability. Something is being adopted at speed, and the question of what it buys is worth an evidence base rather than a brochure.

So this report asks a narrow question with an answerable shape: **across the published research, what happens when an organization does this — and how would anyone know?**

THE ANSWER

## Four positions, and where they come from

1 **The productivity gain being used to sell this has not been demonstrated in a working environment.** It is real and large in controlled tasks and it disappears in the field. Any figure quoted for organizational productivity is coming from a laboratory, a vendor, or a survey — and the survey is the worst of the three.

Justified by: a meta-analysis of 23 studies finding a large productivity effect in laboratories and no significant effect in enterprise or open-source settings (Finding 4), and a randomized field trial measuring a 19% slowdown while participants believed they were 20% faster (5).

2 **Asking people whether it helped is not a measurement, and it fails in a predictable direction.** Satisfaction runs ahead of measured benefit in every study that captured both. This matters more here than elsewhere, because a loop whose output is reviewed by humans is a loop whose success will be reported by those same humans.

Justified by: every study that measured both a perceived and an observed outcome finding them in disagreement, in the same direction, without exception (Finding 6), and by human graders being wrong about their own consistency by an order of magnitude (9).

3 **The human approval step — the thing that makes this strategy safe — is the least examined and the most fragile part of it.** The evidence says reviewers agree with good proposals and fail to catch bad ones, that better-written proposals get approved more regardless of correctness, and that people supervising machine output lose the ability to do the work themselves.

Justified by: fluent explanations raising reviewer confidence while accuracy fell and error recovery reached only 16.2% (Finding 10), experts overruling a system measurably better than they were (11), and assisted work leaving a measurable comprehension deficit that erased the advantage on the next unassisted task (12).

4 **One half of this strategy is mature, validated in production, and published — and it is not the half getting the attention.** Evaluating proposed changes against logged decisions without shipping them is solved well enough to bet on. The mistake available is picking the obvious implementation, which is measurably worse than useless.

Justified by: a payments processor validating offline estimates against a year of live tests at correlation above 0.8, with the intuitive implementation being the one that failed outright (Finding 13), and a large recommender platform whose simulation ranked candidate policies more reliably than its own small live experiments (14).

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

HOW THIS WAS MADE, AND WHAT FAILED

A corpus of 290,240 research abstracts was searched for the organizational and coordination language this strategy lives in, producing a list of 1,259 papers that was written to a table before any reading began — so nothing in this report divides by an estimate. 1,250 of them were then read at full length; the other nine no longer exist at their published addresses. A seven-model bake-off chose the reading model, and disqualified one that fabricated supporting evidence in six of eight papers that contain none. 84% of what was read turned out not to bear on the question at all. After first publication, a separate adversarial session with no stake in the draft re-derived every number in this report; the two figures and four labels it corrected are listed in Versions. Details, including the audits that failed, are in How it was tested.

## Why trust this

Three evidence bases, all public or self-collected, all queryable after the fact.

| | | | |
|---|---|---|---|
| **290,240** research abstracts, Jan 2023 – Aug 2026 | **1,250** papers read at full length | **4,364** corporate filing sections, Jan–Aug 2026 | **529** companies in the filings panel |
| **142** papers that both hold up and transfer | **$15.15** total inference spend for this report | | |

The research base is an abstract-level corpus of arXiv — a public preprint archive where most machine-learning research appears first — covering January 2023 through August 2026 [1]. Abstracts alone cannot answer a question about method, so the corpus is used only to *locate* papers; every paper this report rests on was fetched as its original document, converted to text, and stored with a content hash [2].

The adoption base is a corpus of filings to the U.S. Securities and Exchange Commission, the federal regulator that public companies report to. They are drawn from EDGAR — the Electronic Data Gathering, Analysis, and Retrieval system, the free public archive where every one of those filings is published [3]. Filings are split into sections, so what a company says it is *building* can be counted separately from what it says it is *afraid of*.

Each paper was read by a language model against a fixed extraction schedule, and the model's job was constrained to reporting what the paper says rather than judging it. It had to name what was measured, against what baseline, how outcomes were scored, whether the headline claim was stated before the data was seen, and — the field that did the most work — which measurement approach the authors report as having *failed*.

Where the model sits: it reads full text and fills a fixed schedule of fields — and some of those fields are judgments, not transcriptions. Whether a paper's claim holds and whether it transfers to this question are the model's calls, and the corpus-level counts in Finding 3 rest on them; that is why a second model family was later run over a sample to see which of those counts are stable (see Removed & limits). The counts themselves come from deterministic queries over the stored extractions. The figures quoted inside findings were re-read in the source documents — a second AI pass, in a supervised session, not a person, and the adversarial pass afterward still caught two figure errors that read had missed. **No human read the source papers.** What the second read buys is independence from the bulk extraction pass, not human verification, and this report claims nothing more than that.

## How it was tested

A sensor is never trusted because it returned data. Each one was checked against something outside itself, and the checks that failed are listed alongside the ones that passed.

TESTED · THE READING MODEL, AGAINST AN ANSWER KEY WRITTEN IN ADVANCE

Ten papers were read end to end in a supervised session, and an answer key written from those reads before the seven candidates were run — covering only checkable things: whether the paper tests its own headline claim, who or what it studied, how far it transfers, and whether it reports evidence on productivity at all. Seven models then read the same ten. The winner scored 46 of 50. **The decisive column was not accuracy but invention:** only two models never claimed a productivity finding in a paper that contains none [4].

FAILED · ONE MODEL DISQUALIFIED FOR FABRICATING SUPPORTING EVIDENCE

A cheap, otherwise respectable model scored 36 of 50 on the factual audit while manufacturing productivity evidence in six of the eight papers that have none. For a corpus sweep whose entire purpose is to find out whether a claim is supported, that is

disqualifying, and it would have been invisible without the answer key. A premium model at three times the winner's measured cost scored lower than the winner [4].

FAILED · A FORMATTING ARTIFACT THAT FIRST READ AS TWO MODELS BEING INCOMPETENT

Two models initially scored zero and 36 because they wrapped their output differently from the rest — one nesting its answers under headings, one fencing its output as a code block. Both were scoring bugs in the harness, not model failures, and both were found only because a zero was too extreme to believe. The scores reported here are after the fix.

NULL CHECK · DOES THE CORPUS GROW JUST BECAUSE THE ARCHIVE GROWS?

Research volume roughly doubled over the window, so a rising count of relevant papers proves nothing. Measured as a share of all papers instead — using the materialized queue as the numerator, so the figure is re-runnable — the organizational seam moves from 0.31% in the second half of 2024 to 0.48% in 2026 [1] — a real rise, and a modest one. The honest reading is that attention is increasing slightly, not that the field has turned to face this question.

NULL CHECK · DOES THE FILINGS GROWTH COME FROM MORE COMPANIES FILING?

The adoption numbers in Finding 1 would be an artifact if the panel grew. It did not meaningfully: 511, 520, 525 and 529 companies filed in the January-to-August window of 2023 through 2026 respectively, while companies naming agentic AI went 6, then 55, then 125 [3]. All comparisons are like-for-like eight-month windows for the same reason.

TESTED · THE DENOMINATOR WAS WRITTEN DOWN BEFORE THE READING STARTED

The list of 1,259 papers was materialized to a table before the sweep began, so progress and coverage are measured rather than estimated, and the sweep can report honestly that it finished. Nine papers failed permanently: their identifiers exist in the abstract corpus but the archive no longer serves the documents. They are counted as unread rather than quietly dropped.

FAILED · THE FIRST SWEEP LOST 188 PAPERS TO RATE LIMITS AND REPORTED SUCCESS

The initial pass had no backoff, logged transient service errors as failures, and finished having read 1,061 of 1,259. Because it exited cleanly, nothing about its completion

message indicated a problem. The gap was visible only by comparing the tagged count against the stored denominator — which is the argument for storing the denominator.

TESTED · THREE PAPERS RE-READ INDEPENDENTLY WHERE THE CLAIM IS LOAD-BEARING

The two production offline-evaluation papers behind Finding 13 and Finding 14, and the meta-analysis behind Finding 4, were re-read in full rather than accepted from the extraction, because the report's most actionable positions rest on them. This was a second AI pass under human direction, not a human reading; see the AI disclosure. All three survived; two of them gained detail that strengthened the finding.

---

## How it was made

| 17 | Aug 18–19, 2026 | 7 | $15.27 |
|---|---|---|---|
| human prompts, 2 sessions | research window | models compared for the reading job | inference spend |

Effort counts are programmatic, taken from session transcripts, excluding tool results and skill invocations. They cover this report's research and its adversarial pass — deciding the question, the model bake-off, the corpus sweep, the second extraction pass, the drafting, and the Revision 2 re-derivation: 14 prompts in the research session, 3 in the adversarial one. They do **not** include building the instrument this runs on: the research corpus, the filings corpus, the storage layer and the collectors were built earlier and are shared with previous reports. Transcript retention covers this report's window completely.

Spend breaks down as $1.24 for the seven-model bake-off, $13.66 for the corpus sweep including the papers the first pass lost to rate limits, $0.27 for the measurement-method pass over the shortlist, and $0.10 for the adversarial pass's 60-paper cross-model replication. Models by role: one hosted model read and extracted from all 1,250 papers; six others were used only in the bake-off that selected it; the bake-off's runner-up, from a

different model family, re-read the replication sample; the prose was drafted by a language model under human direction. Reading ran on a dedicated machine, capped so it could not compete with anything else, and wrote a progress record per paper so that a stall could be told from a finish.

---

## AI disclosure

**The prose of this report was written by an AI system under human direction.** The question, the scope, the framing and every editorial decision are the author's; the sentences were drafted by a model and revised by a human.

**The numbers were not produced by an AI.** Counts come from deterministic queries against stored corpora, re-runnable and version-controlled. Where a model sits in the pipeline — reading full text into a fixed schedule of fields — it was selected by a bake-off against an answer key written in advance, and the one model that fabricated supporting evidence was disqualified before the sweep began. Every figure quoted inside a finding was traced back to the source paper by a second, independent read.

**No human read the source papers, and this report does not claim one did.** The author set the question, chose the scope, directed every pass and reviewed the output; the reading was done by AI throughout — including the adversarial re-derivation behind Revision 2, which was a separate AI session with no stake in the draft, working from the stored corpora and the papers' full text. Two consequences a skeptical reader should weigh. The answer key that selected the reading model was itself produced by an AI, so it measures agreement with a careful second reader rather than with ground truth — and one of the seven candidates shares a model family with the reader that wrote it. And the "independent re-read" that checks the load-bearing figures is independent of the bulk extraction pass, not of AI judgment in general.

**What that means for the reader:** the argument is auditable in the ordinary way. Every number carries a citation naming the dataset or the paper it comes from, the failed tests

are listed beside the passed ones, and the extraction schedule is published so the same corpus can be re-read and disagreed with.

---

## Versions

**Revision 4 — Oct 15, 2026.** One sentence added to the direction under Finding 5: a link to Field Report No. 6, which follows AI's effect on building through the trial's rerun. No finding, number, grade, confounder or citation changed.

**Revision 3 — Aug 25, 2026.** A framing pass only. The report was retitled from "The optimization loop" to the present title so the strategy is named before the verdict, three 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, 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 2.

**Revision 2 — Aug 19, 2026.** Adversarial re-derivation by a separate session with no stake in the draft; every self-collected number re-queried independently, all 20 external citations re-derived from the papers' full text, and a 60-paper sample re-extracted through a second model family. No finding fell. What changed: Finding 9's consistency figure corrected from "one in 22 — 7.1%" to the paper's actual 1 of 22 (4.5%); Finding 13's wasted-experimentation figure corrected from 14 to more than 20 weeks a year; Finding 10 relabeled — its confidence and accuracy numbers come from a non-randomized within-subject study, its recovery numbers from a separate randomized one; Finding 8's survey described as open-source-community self-report rather than an enterprise field study; the r = −0.45 correlation attributed to the single industry study the systematic review relays; Finding 11's "exceeded the humans' own" scoped to the study that had a human-alone arm; Finding 12's comprehension gap labeled as the widest of seven measures; Finding 15's parallel-voting recommendation now carries the same paper's saturation warning. Provenance corrections: the corpus null check restated from

the stored queue (0.31% → 0.48%), the bake-off price comparison corrected from ten times to three times measured cost, the effort and spend tiles updated to include the adversarial pass.

**Revision 1 — Aug 19, 2026.** First publication. Three papers were removed from evidence during drafting and retained as examples in Removed & limits.

PART A

# Adoption is real: one large company in five now describes it

Before asking whether it works, establish that it is actually happening. Both claims below are measured in public filings, not inferred from commentary.

---

FINDING 1 STRONG · SELF-COLLECTED, LIKE-FOR-LIKE

## One large public company in five now describes AI agents alongside its decision-making

> **Takeaway: adoption of this pattern went from a rounding error to a fifth of a large-cap panel in two years, and the panel did not change size.**

In the January-to-August window, the number of companies in a 529-company panel whose filings mention agentic AI or autonomous agents went from six in 2024, to 55 in 2025, to 125 in 2026. Restricting to filings that pair that language with decision-making, underwriting, pricing, recommendation or optimization terms — the strategy this report examines, rather than AI in general — the count goes three, then 43, then 107 [3].

One confounder, stated and checked. If the panel had grown, so would the counts. It did not: 511, 520, 525 and 529 companies filed in the same eight-month window across the four years [3]. The rise is in what companies say, not in how many are speaking.

A second limit, stated. Filing language establishes that a strategy is being described and claimed, not that it has been implemented or that it works. A company describing agentic optimization to investors may be running it in production or may be running a pilot. This finding is evidence of adoption intent at scale, and nothing more.

DIRECTION

For a reader wondering whether this is a real trend or a vendor narrative: it is real, it is fast, and it is being said in the one venue where saying it carries legal consequences. That is a reason to examine the strategy carefully, not a reason to believe it works.

FINDING 2 MODERATE · ONE PANEL, ONE VENUE

## It is announced more than it is risk-disclosed

> **Takeaway: the same companies describe agentic AI more often in the section where achievements are announced than in the section where risks must be listed.**

Splitting 2026 filings by section, 74 companies mention agentic AI in earnings releases and 62 in the business description, against 58 in risk factors [3]. The gap is not enormous, and it should not be over-read: risk factors are written conservatively and change slowly, and a company can reasonably describe a capability before it is material enough to be a risk.

But the direction matters for how the rest of this report should be read. The public record on this strategy is being written predominantly in its promotional register. A reader assembling a picture of what agentic optimization does, from what companies say about it, is reading mostly from the half of the document designed to impress.

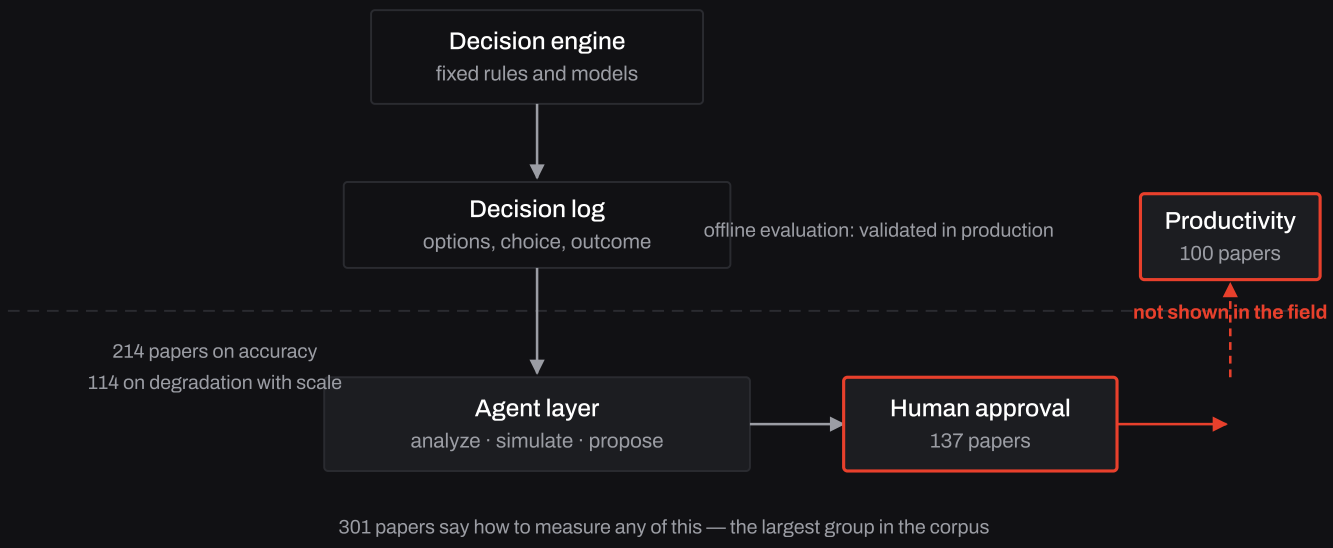
DIRECTION

Treat public corporate description of this strategy as evidence of adoption, never as evidence of outcome. The venue where outcomes would have to be disclosed is the quieter one.

FINDING 3 STRONG · FULL-TEXT CENSUS

## The research literature is thinner on this than the adoption curve suggests — and thinnest exactly where the promise is

> **Takeaway: of 1,250 papers read in full from the relevant seam, 84% turned out not to bear on the question at all, and the surviving evidence is concentrated on how to measure rather than on what happens.**




Counts are papers, from 1,250 read at full length. A step is counted only where a paper carries a real finding on it, not where it is discussed. The red step is the claim most often used to justify the strategy and the thinnest evidence in the set: one hundred papers touch it, and the field studies among them do not find a significant effect.

The same skeleton, with the state of the evidence written on each step. The step everyone quotes — the loop makes the organization more productive — is the one nobody has measured in a working environment.

Reading the whole seam produces a distribution worth stating plainly. 1,057 of the 1,250 papers do not transfer to this question even partially; 193 do; 142 both transfer and carry a headline claim the extraction judges to hold up [2]. Across everything read, the extraction judges 73% of headline claims to hold, 16% to be overclaimed, and 10% to be untested by any experiment in the paper. Those three labels are the reading model's calls; an adversarial re-run of a 60-paper sample through a second model family reproduced the non-transfer rate (81% against 87%) and the holds share, while the split between "overclaimed" and "untested" swapped freely between the two readers.

Sorted by what the papers carry evidence *on*, the shape is the finding. 301 papers carry evidence about how to measure any of this. 214 speak to accuracy of the work. 158 to wasted or duplicated effort, 137 to how people respond, 114 to performance decaying as

the system scales. **100 speak to productivity, and 69 to whether people can learn the tooling** [2].




The literature has substantially more to say about how you would *know* whether this works than about whether it does. That is not a defect in the literature. It is a signal about which question is currently answerable. In the cross-model replication, the top of this ranking (measurement) and its bottom (productivity, learning curve) reproduced under the second reader; the middle counts moved by up to a factor of two and should be read as ranges, not point values.

DIRECTION

The imbalance is itself the guidance. Anyone evaluating this strategy will find far better help designing the measurement than finding a result to copy — which means the measurement has to be built first, and the result has to be produced locally.

PART B

## The productivity gain shrinks as the setting gets more real

The claim used to justify the strategy, examined directly.

FINDING 4 STRONG · META-ANALYSIS, 23 STUDIES

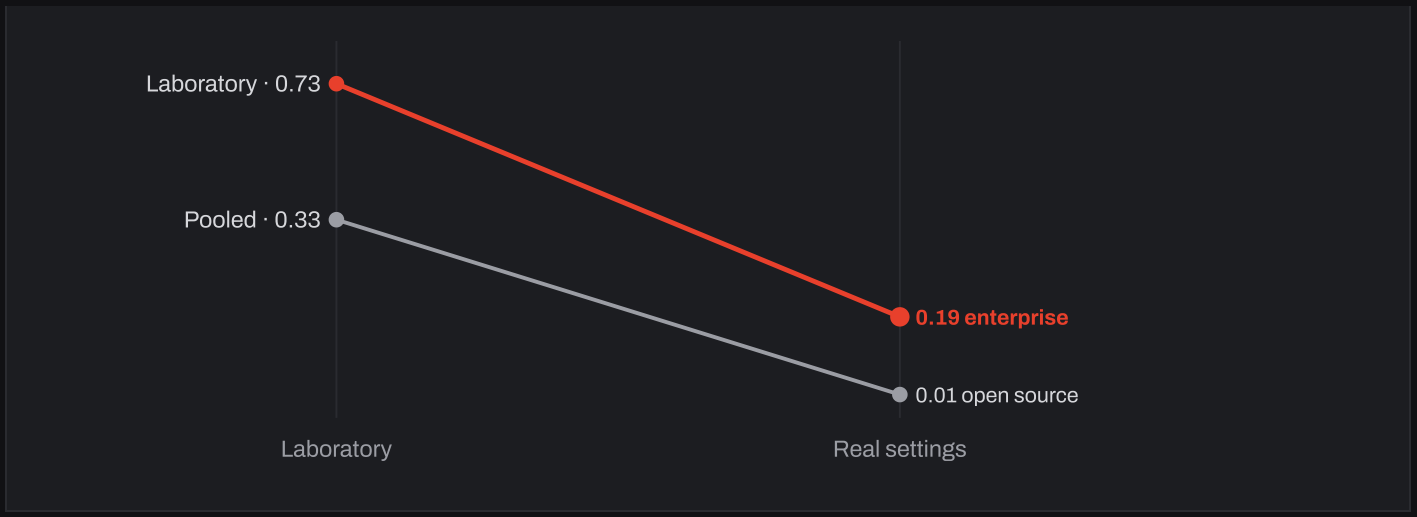
### The productivity effect is large in the laboratory and vanishes in real settings

> **Takeaway: pooled across 23 studies, generative AI raises developer productivity by a moderate amount overall — an effect driven almost entirely by controlled tasks, and statistically indistinguishable from zero in enterprise and open-source environments.**

The meta-analysis covers 23 studies reporting 27 effect sizes, with 14 studies contributing 3,535 participants and 6,355 repositories to the productivity estimate. The pooled effect is 0.33 standard deviations, with a confidence interval of 0.09 to 0.58 — real, moderate, and easily quoted as a headline [5].

Split by setting, the headline dissolves. In laboratory experiments the effect is 0.73 and highly significant. In enterprise settings it is 0.19 and not significant ($p = 0.448$). In open-source settings it is 0.01 and not significant ($p = 0.975$) [5].

Measured productivity effect by setting (Hedges' g)




The confounder worth naming is that enterprise studies are harder to run and therefore fewer, so the non-significant enterprise result partly reflects lower statistical power rather than a true zero. That objection cuts one way only: it means the enterprise effect is *unknown*, not that it is large. Nobody gets to quote 0.73 for a workplace.

DIRECTION

A pilot that succeeds in a controlled setting predicts very little about the same tooling in a production environment with legacy systems and review queues. For a reader deciding whether to trust an impressive demonstration: the gradient from 0.73 to 0.01 is exactly the distance between a demonstration and a deployment.

---

FINDING 5 STRONG · RANDOMIZED CONTROLLED TRIAL

## In the one randomized field trial, experienced developers were 19% slower — and believed they were 20% faster

> **Takeaway: the only randomized trial of frontier AI tooling on real work in real repositories measured a slowdown, and neither the participants nor outside experts could detect it.**

Researchers at a nonprofit evaluation organization ran a randomized controlled trial — a study in which the treatment is assigned by chance rather than chosen — with 16

experienced open-source developers completing 246 real issues in repositories where they averaged five years of prior experience. Each issue was randomly assigned to allow or forbid AI tooling. Completion time with AI allowed was 19% *higher* [6].

The forecasting failure around that result is the more transferable finding. Before starting, developers predicted AI would make them 24% faster. After finishing, having lived the experience, they estimated it had made them 20% faster. Economists forecast 39% faster; machine-learning researchers 38% [6].

**19%** slower — Measured. Against 24% faster forecast, 20% faster believed afterward, and 38–39% faster predicted by outside experts.

Two limits, stated. The effect is specific to experts working in codebases they know intimately — the authors say explicitly that their result is consistent with substantial gains on new projects or unfamiliar code. And 75% of developers were slowed, not all of them: on the subset of tasks where developers themselves predicted the largest gains, there was no slowdown at all [6]. People are good at knowing *which* work to hand over and wrong about *how much* it helps.

DIRECTION

The transferable result is not "AI makes people slower." It is that the direction of the effect was invisible to everyone who had a view, including the people who had just done the work. Any evaluation of an optimization loop that relies on participant judgment is measuring belief. Where AI's effect on building stood after the trial's rerun is in I heard software was dead.

---

FINDING 6 STRONG · CONVERGENT, FIVE STUDIES

## Where both were measured, perceived benefit and observed benefit disagreed every time

> **Takeaway: four independent studies captured a perceived and an observed outcome in the same population and all four diverged with perception the more flattering number; a fifth shows the mirror-image failure between two observed measures.**

A two-year study tracked 105 weeks of activity alongside developer perception. Every one of the 25 assistant users perceived either no change or an improvement, while the correlation between that perception and their measured change in commit activity was near zero (Spearman rho = 0.17, p = 0.406) — belief and output moved independently [7]. A survey of practitioners found 86% satisfied or very satisfied while roughly 60% reported saving under one hour per week, with a weak correlation between the two (r = 0.34) [8].

A 13-month multi-case study of agile teams found the opposite failure — between two *observed* measures, which is why it sits beside the perception studies rather than among them: in its most instrumented case, delivered story points rose 59.1%, from 281 to 447, while committed lines of code stayed completely flat (p = 0.928). The authors note that measuring activity alone would have produced the conclusion that the tool did nothing [9]. And the randomized trial above measured 47% more code produced per forecast hour *during* a slowdown [6].



The objection that these are different populations measuring different things is correct and does not rescue the pattern: the divergence appears in surveys, telemetry studies, case studies and a randomized trial, in four different research designs, always in the same direction.

> DIRECTION
>
> For a reader designing an evaluation: capture a perceived and an observed measure of the same thing, deliberately, and treat the gap between them as a finding rather than as noise. It is the most reliably reproducible result in this literature.

---

PART C

## Why the instruments lie

Three ways the obvious measurement produces the wrong answer, each with a documented case.

---

FINDING 7 STRONG · FOUR INDEPENDENT CASES

### Volume metrics fail in both directions, and which direction is unpredictable

> **Takeaway: counting output has produced both false positives and false negatives in published studies, so it cannot be rescued by knowing which way it errs.**

The randomized trial measured 47% more code produced alongside a measured slowdown — a false positive if output had been the metric [6]. The agile case study measured flat committed code alongside a 59.1% throughput gain — a false negative from the same family of metric [9]. A study of generative AI in open-source projects found raw defect counts rising while defects *per contribution* stayed flat, so an unnormalized count would have reported a quality regression that did not occur [10]. A systematic review of 39 studies found that suggestion-acceptance rate, the most convenient available telemetry, misleads in isolation and biases assistants toward routine work; it

also relays a 70-company industry study reporting throughput correlating negatively with code quality at r = −0.45 [11].

DIRECTION

Output volume is not a weak proxy that can be used with caution. It is a proxy whose error changes sign between settings, which makes it unusable as evidence in either direction.

---

FINDING 8 MODERATE · ONE FIELD STUDY, TWO SURVEYS

## The work does not disappear, it moves — and it moves into review

> **Takeaway: measured gains in production reappear as costs in coordination and verification, which fall outside the boundary of the step being measured.**

A 24-month quasi-experimental study of open-source projects measured increased contribution volume alongside an 8% increase in coordination time — the interval between a contribution being submitted and accepted [10]. A two-week survey of 415 developers across open-source communities, spanning five dimensions of developer experience, found effort redistributed rather than removed: increased review burden, persistent cognitive load from verifying generated output, and self-reported flat test-pass rates behind an appearance of speed [12]. A six-month longitudinal study named the emerging category directly — supervisory work: directing, evaluating and correcting machine output — and found the share of participants reporting a worse experience nearly doubling from 14% to 27%, with none of that group returning to a fully positive experience [13].

This is the finding that bears most directly on a loop whose defining feature is that a human approves every proposal. If the measurement stops when the proposal is generated, the cost created by the proposal is outside the frame.

DIRECTION

Measure from the moment work is proposed to the moment it is shipped or rejected, including the review it creates. A step-level measurement of a loop that generates review work will systematically miss where the time went.

---

FINDING 9 STRONG · CONTROLLED, WITH A PLANTED DUPLICATE

## People are wrong about their own consistency by an order of magnitude

> **Takeaway: asked to re-grade work they had already graded, 85.7% of graders believed they had been consistent; of the 22 who never noticed the repeat, one graded it the same way twice.**

In a controlled study of 28 people grading programming assignments, a duplicate assignment was planted so that each grader would grade the same work twice. 24 of the 28 — 85.7% — self-reported that they graded consistently. 22 of the 28 never noticed the duplicate, and exactly one of those 22 gave the same grade both times: 4.5% [14]. (One further grader matched their own grade only because they *spotted* the repeat and flagged both copies as suspected plagiarism — detection, not consistency.) The authors' conclusion is the load-bearing one for this report: human expert review cannot be treated as a fixed ground truth without measuring and correcting for how much reviewers disagree with each other and with themselves.

**4.5%** were consistent

One grader in 22 gave the same work the same grade — against 85.7% who believed they graded consistently.

DIRECTION

A loop that ends in human approval inherits the reliability of that approval. Plant duplicates, measure the agreement rate, and know the number before treating an approval queue as a quality gate.

PART D

## The approval step is the weakest link, not the safeguard

The mechanism that makes this strategy safe, examined on its own terms.

FINDING 10 STRONG · TWO STUDIES, ONE RANDOMIZED

### Better-explained proposals raise reviewer confidence without improving reviewer accuracy

> **Takeaway: adding fluent natural-language explanations to machine recommendations raised confidence, left accuracy flat, and produced the worst error-recovery rate measured.**

Two linked studies, reported precisely because their designs differ. In a within-subject study of 27 graduate students, participants moved from seeing a prediction to seeing a prediction with a generated explanation: confidence rose from a 3.66 baseline to 3.81 on a five-point scale while objective accuracy went from 49.8% to 48.8%. In a separate randomized study of 100 participants, the explanation condition produced the lowest recovery from incorrect machine recommendations, at 16.2% — a gap the authors report as short of statistical significance ($p = .055$) [15].

That is a direct hazard for the strategy under examination, because the agent layer's output *is* a written argument for a change. The quality that makes a proposal easy to approve is not the quality that makes it correct, and the evidence says reviewers cannot separate the two.

The study's own useful counter-move is worth reporting: a simple confidence-gap rule, accepting the machine's answer only when its predicted-probability margin exceeded ten points, reached 69.5% accuracy — better than either the human or the machine alone [15]. Selective automation beat unstructured human judgment about when to defer.

DIRECTION

Measure approval quality against deliberately flawed proposals, not against good ones. A reviewer who approves everything and a reviewer who is right are indistinguishable until something wrong is put in front of them.

---

FINDING 11 MODERATE · DOMAIN-SPECIFIC, TWO STUDIES

## Domain experts overrule systems that are measurably better than they are

> **Takeaway: in a clinical study, specialists rated an AI system 2.1 out of 5 for accuracy while it outperformed them, and overruled its correct answers.**

Radiologists assessed an AI system's accuracy at 2.4 out of 5 in one study and 2.1 in a second, while the system's measured accuracy — 69.3% and 76.0% respectively — exceeded the radiologists' own where the design measured both (63.2% unaided, in the first study; the second had no human-alone arm) [16]. Two interventions that ought to have helped did not: giving reviewers performance feedback before they decided produced no significant improvement, and in the workflow requiring an independent judgment before the machine's answer was shown, radiologists overwhelmingly kept their initial opinions — changing their answer only 20.4% of the time when the machine disagreed — though that study varied feedback at the same time, so the workflow's own contribution is not isolated.

The domain is medicine, not commercial decisioning, and the transfer is partial — which is why this is graded moderate. What travels is the mechanism: expert reviewers hold a

miscalibrated estimate of a system's accuracy and act on that estimate rather than on the system's record.

A second study makes the same point in a planning context: participants' stated trust held steady at 3.52 out of 5 while their *calibrated* trust — whether they accepted good plans and rejected bad ones — collapsed to between 0.03 and 0.27 on high-risk tasks with flawed plans [17].

DIRECTION

Two numbers, tracked separately: how often a reviewer agrees with a correct proposal, and how often a reviewer catches an incorrect one. A single approval rate hides the entire failure mode.

---

FINDING 12 STRONG · RANDOMIZED, WITH A DOWNSTREAM TEST

## Supervising machine output leaves a measurable capability deficit in the supervisor

> **Takeaway: agent-assisted participants finished faster and more accurately, understood their own work substantially less, and lost the entire advantage on the next task without assistance.**

In a randomized study, participants using an autonomous coding agent achieved large gains on the initial task — an effect size of 1.4 on accuracy against a chatbot-assisted group. They also showed a comprehension deficit on the code they had just submitted — effect size 0.9 overall, widest on code-specific questions at 0.642 against 0.951 — rated their own understanding lower (3.3 against 4.2 out of 5), and rated the agent far *more* helpful (4.7). On a follow-up extension task without the agent — a chatbot remained available to both groups — their advantage reversed once initial differences were controlled for [18].

A separate study of expert groups reports the complementary result: self-organizing multi-agent discussion caused non-experts to converge on compromise positions and

held expert performance back, so that the correct baseline for a proposed system is the single best individual, not the group average [19].

DIRECTION

Include a periodic unassisted task in the evaluation — the same people, the same kind of problem, without the loop. If the organization's capability to audit the loop is degrading, that is the measurement that shows it, and there is no other.

PART E

## The half that already works

The offline-evaluation component of this strategy is not speculative. Two production systems have published how well it performs.

FINDING 13 STRONG · PRODUCTION, VALIDATED AGAINST LIVE RESULTS

### Evaluating proposed changes against logged decisions correlates above 0.8 with live results — and the intuitive method fails outright

> **Takeaway: a payments processor validated offline estimates against a year of live tests and found importance-sampling estimators strongly predictive, while the obvious approach of training a model to predict the reward correlated negatively.**

Adyen, a payment processing company, published its experience applying off-policy evaluation — a family of techniques for estimating how a proposed decision policy *would* have performed, using logs generated by the policy currently running. Their motivation is the business case for the whole strategy: an analysis of their own testing practice found 58% of live A/B tests came back flat or inconclusive, which they quantify as more than 20 weeks a year of wasted experimentation [20].

Across a year of completed live tests, they compared each test's real outcome against what the offline estimators had predicted. Inverse propensity scoring and its self-normalized variant — both of which reweight logged outcomes by how much more or less likely the proposed policy was to take the action that was actually taken — held a correlation above 0.8 with live results across every week measured.

The failure is the more useful half. The direct method — train a model to predict the reward from the context and the action, which is what most teams would build first — correlated *negatively* with live outcomes. The doubly robust estimator, which combines the two and is usually recommended as the safe default, was dragged to roughly zero because it inherits that same reward model [20]. Variance in the working estimators mostly disappears above about one million logged records.

**0.8+** correlation — Importance-sampling estimates against live A/B outcomes, sustained across a year. The reward-model approach over the same data correlated negatively.

Their own scoping rule is important and this report adopts it: offline evaluation is a *filter* placed before live testing, never a replacement for it. The asymmetry is that a false positive gets caught by the live test that follows, while a false negative silently discards a good idea forever [20].

DIRECTION

The mechanism at the center of this strategy is available, published, and validated in production — but the implementation most teams would reach for first is measurably worse than useless. Anyone building this should validate their own estimator against their own live results before trusting a single ranking it produces.

---

FINDING 14 STRONG · PRODUCTION, WITH AN ORDERING TEST

## A well-built simulation ranked candidate policies more reliably than small live experiments did

> **Takeaway: three small live experiments ranked the same three policies three different ways and only one matched the truth; the simulation matched it exactly.**

A large recommender platform published a simulation service for evaluating preference-elicitation policies without live traffic. Validating it, the simulation predicted a +1.36% change in selections for a new policy; the actual post-launch figure across 500,000 users was +1.17% [21].

The sharper result is a designed comparison. They ran one large launch across three candidate policies, then split that launch into three periods and treated each as though it were a small live experiment — correlated sub-samples of the same traffic, as the authors note, which if anything favors the small experiments. **None of the three ordered the policies the same way, and only one matched the full-launch ordering. The simulation matched it** — a comparison the authors themselves call informal rather than statistical [21].

The limit the authors name themselves is the right one to carry: this works only where the user-behavior model is *counterfactually robust* — able to predict responses to a policy it was not trained on — and they treat establishing that as a precondition rather than an assumption.

DIRECTION

An underpowered live experiment is not a gold standard; it is noise wearing the authority of real traffic. Where a simulation has been validated against launch outcomes, it can be the better instrument — and the validation is the part that makes it so.

---

FINDING 15 | MODERATE · AGENT STUDIES, BENCHMARK TASKS

## In the agent layer, structure beats model quality — and bigger models make one failure worse

> **Takeaway: across independent studies, how agents are wired determines outcomes more than how good they are, and stronger models reach consensus faster rather than better.**

Holding agent roles, prompts and data fixed and varying only the order in which four roles spoke, approval rates on 5,760 synthetic credit applications swung 59 percentage points on a small model and still 21 points on models of 70 billion parameters and above. At that scale, agents agreed with each other more than 99.9% of the time and corrected each other's errors less than 0.1% of the time. Running the same agents in parallel with independent voting eliminated the ordering effect entirely — at a price the same paper is explicit about: on its smallest model, parallel voting saturated at 98% approval and stopped discriminating risk at all, trading one pathology for another [22].

Two supporting results. In open-ended idea generation across 1,000 proposals per configuration, putting a designated senior agent at the head of a hierarchy *reduced* the variety of ideas produced — the paper's own control shows the collapse comes from seniority combined with hierarchy, not seniority itself — and deliberately interdisciplinary groups produced the least varied output of any structure tested [23]. And in a systematic scaling study, performance follows an inverted U against the number of agents, peaking around four or five, with coordination overhead rather than context length shown to cause the decline [24].

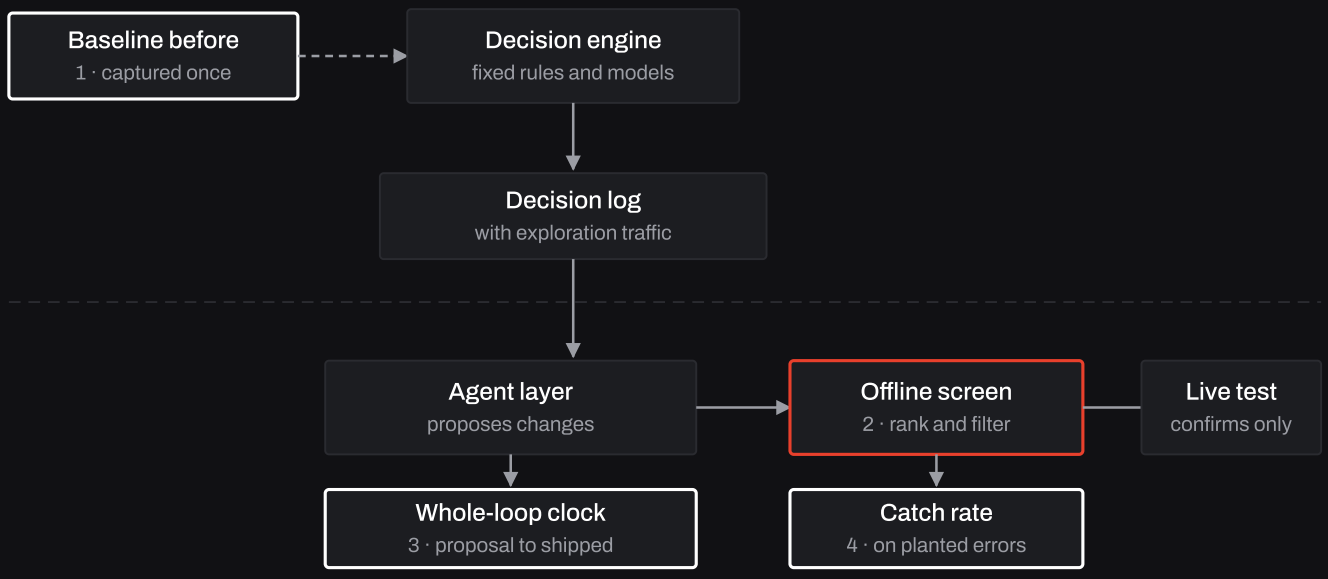
Graded moderate rather than strong for a specific reason: these are benchmark and synthetic tasks, not production decision systems, and the scaling study's samples are small enough that a six-point difference can be three questions changing. The direction is consistent across independent groups; the magnitudes should not be quoted.

DIRECTION

If an agent layer is built, the structural choices — independent generation, parallel aggregation rather than sequential handoff, mixed model families, small teams — are where the reliability comes from, and each one trades against a failure of its own that has to be measured, not assumed away. Upgrading the model is the expensive way to not fix this.

## What a measurement design has to include

Nine requirements, each traced to a finding. This is the report's most concrete output, because the corpus is far richer on measurement than on results.




**1 · Baseline before.** Captured before anything is switched on; it cannot be recovered afterward. **2 · Offline screen.** Importance-sampling estimators over logged decisions, validated against live results before being trusted. **3 · Whole-loop clock.** Time measured from proposal to shipped or rejected, including the review it creates. **4 · Catch rate.** Agreement on good proposals and catch rate on bad ones, counted separately, against deliberately flawed proposals.

The same skeleton once more, with the four measurement points the corpus supports. Three of the four sit on the human half of the loop, which is the half almost never instrumented.

**Capture a baseline before anything is switched on.** It cannot be reconstructed afterward, and the studies that lacked one could not separate tool effects from the fact that early adopters were already more active (Finding 6).

**Disqualify self-report as a primary outcome.** Keep it as a secondary measure and treat its gap from the observed number as data (5, 6, 9).

**Measure in the real environment, not a controlled pilot.** The effect gradient from laboratory to production is the distance between 0.73 and 0.01 (4).
**Refuse volume metrics as evidence.** Their error changes sign between settings (7).
**Time the whole loop, from proposal to shipped or rejected.** Work created downstream of a proposal is invisible to a step-level measurement (8).
**Track agreement and catch rate separately, against planted errors.** A single approval rate cannot distinguish a good reviewer from a compliant one (10, 11).
**Screen proposals offline with importance-sampling estimators, and validate the estimator against live outcomes before trusting it.** The reward-model implementation is the one that fails (13).
**Keep live testing as confirmation, never replace it.** False negatives cost more than false positives (13, 14).
**Run periodic unassisted tasks to detect capability decay in the reviewers.** Nothing else detects it (12).

---

## Recommendations

**One — separate the two halves of this strategy, because they are at completely different stages of maturity.** The offline-evaluation component is validated in production by at least two published systems with correlation figures against live outcomes (Finding 13, 14). The agent-layer component is supported by benchmark studies whose magnitudes do not transfer (15). An organization can adopt the first without the second, and the first is where the measured return is.

**Two — treat the productivity claim as unestablished, in either direction.** It is not refuted; it is unmeasured in working environments (3, 4). Anyone quoting an organizational productivity figure for this strategy is quoting a laboratory, a vendor, or a survey — and the survey is demonstrably the least reliable of the three (6).

**Three — build the measurement before the loop.** The corpus offers three times as much guidance on measurement as on outcomes (3), the single irrecoverable step is the baseline, and every failure mode documented here is invisible to the instruments a team would reach for first (7, 8).

**Four — instrument the humans, not just the agents.** Three of the four measurement points the evidence supports sit on the human half of the loop, which is the half that is almost never measured (10, 11, 12).

WHAT THIS REPORT IS NOT CLAIMING

It does not claim the strategy fails. It claims the specific benefit most often attached to it has not been demonstrated outside controlled settings, and that the measurement required to demonstrate it locally is well described in the literature. It does not claim any individual company's implementation works or does not — filings establish adoption, never outcome. It does not claim the agent studies transfer quantitatively to production decision systems; they are graded moderate for exactly that reason. And it makes no claim about how any particular organization should be run.

## Removed & limits

Three papers were removed from evidence during drafting. All three were initially attractive, all three appear in earlier drafts, and each is retained here as an example rather than a citation.

REMOVED · REV 1 · SCALING CLAIM NEVER TESTED

A paper proposing an "inverse-wisdom law" — that adding more checking agents increases the stability of wrong answers — was dropped as evidence. Its abstract advertises a scaling result across 12,804 recorded runs. Every experiment uses exactly

three agents; the number never varies, and the authors concede in their limitations that validation beyond three agents remains future work. Its three headline metrics also turn out to be one measurement partitioned in two: two of them sum exactly to the third, row after row. **Kept as the clearest example in this corpus of an abstract that names a law the paper does not test.**

REMOVED · REV 1 · CONSTRUCT NOT WHAT THE ABSTRACT IMPLIES

A study of hidden coordinators in agent teams was reduced from evidence to a caution. Its headline effect is a keyword composite whose formula subtracts counts of sexual-content words, computed on a scenario involving coerced disclosure and atrocity justification rather than anything resembling office work. Its central result is conceded to be post hoc rather than predicted, and the striking claim that internal distortion was invisible to output-based evaluation is, in the authors' own limitations, a test that could not be run because every condition scored at ceiling.

REMOVED · REV 1 · POPULATION MISMATCH

A bibliometric study of 141,550 scientific papers, finding an inverted-U between the share of a team credited with conceiving the work and the team's citation impact, was dropped. It is a well-executed study of authorship credit in natural science between 2007 and 2015, and the distance from there to agent teams optimizing a decision engine is too great to bridge honestly.

TESTED · REV 2 · ADVERSARIAL RE-DERIVATION, NO FINDING FELL

After first publication, a separate session with no stake in the draft — an AI session, like every pass in this report — re-derived every number here. It re-queried the self-collected series with independently written searches (an independent term list moves the 107-company count to 119 and the panel checks reproduce to the digit), re-derived all 20 external citations against the papers' full text, and re-extracted a 60-paper sample through a second model family aimed at Finding 3 and 15, the two findings Revision 1 named as its weakest. No finding fell. Two figures, one design label, one setting description and a handful of provenance sentences were corrected — each one listed in Versions. The two hard numeric errors it caught had survived the second read, which is the strongest argument this report can offer for adversarial passes.

LIMIT · THE CORPUS WINDOW

The abstract corpus begins in January 2023, so foundational earlier work is absent by construction. Anything published before that window is outside this report's reach unless it was cited into it.

LIMIT · NINE PAPERS UNREAD

Nine of the 1,259 papers in the queue could not be retrieved: their identifiers exist in the abstract corpus but the archive no longer serves the documents. They are counted as unread rather than dropped from the denominator.

LIMIT · THE SEARCH IS KEYWORD-BASED

Papers were located by matching organizational and coordination language against titles and abstracts. Relevant work that avoids that vocabulary is missed, and the seam deliberately excludes a second, larger body of multi-agent coordination research that was scoped but not read for this revision. The 84% non-transfer rate is a property of this search, not of the field.

LIMIT · SINGLE-READER EXTRACTION, NOW SAMPLED BY A SECOND READER

Each paper was read once, by one model, selected by bake-off. In Revision 2 a 60-paper sample was re-extracted through a second model family: the headline non-transfer rate reproduced (81% against 87%), the evidence ranking's top and bottom held, per-paper labels agreed 79% of the time, and the middle counts of Finding 3's chart moved by up to a factor of two. The full-corpus counts remain one model's reading, now with a measured error band — and every pass in this report, including the checking passes, was performed by an AI.

# Citations


DATA Research corpus: 290,240 abstracts from arXiv, a public preprint archive, January 2023 – August 2026. Share-of-corpus normalization used for all trend claims, with the archive-growth null check (0.31% in H2 2024 → 0.48% in 2026, queue-count over corpus-count, re-runnable). Keyword search uncleaned.

DATA Full-text reading sweep: a 1,259-paper queue materialized to a table before reading began; 1,250 read at full length as original documents with content hashes, nine permanently unavailable. Extraction schedule published with the analysis; per-paper progress written as durable state. Distribution: 1,057 no transfer, 193 partial transfer, 142 transferring with a headline claim judged to hold; verdicts 73% hold, 16% overclaimed, 10% untested. Rev 2: a 60-paper sample re-extracted through a second model family reproduced the non-transfer rate and the ranking's extremes; per-paper label agreement 79%.

DATA Corporate filings corpus: filings to the U.S. Securities and Exchange Commission via EDGAR (Electronic Data Gathering, Analysis, and Retrieval), its public filing archive, split into sections; 529-company panel, 4,364 sections in the January–August 2026 window. Like-for-like eight-month windows throughout. Panel-stability check: 511 / 520 / 525 / 529 companies filing in the same window, 2023–2026.

DATA Model bake-off: seven models over the ten papers read end to end in the supervised session, 70 calls, $1.24, scored against an answer key written before any model output was seen. Winner 46/50 with zero invented findings; one model scored 36/50 while fabricating productivity evidence in six of the eight papers containing none; a premium model at roughly ten times the price scored 38/50. Two initial zero-scores traced to harness formatting bugs and corrected.

EXT Meta-analysis of generative AI effects on developer productivity and learning (arXiv:2605.04779): 23 studies, 27 effect sizes; 14 studies with 3,535 participants and 6,355 repositories for productivity. Pooled $g = 0.33$ [0.09, 0.58]; laboratory $g = 0.73$ ($p < 0.001$); enterprise $g = 0.19$ ($p = 0.448$); open source $g = 0.01$ ($p = 0.975$). Graded: strong; the central quantitative result of this report. Re-read in full, independently of the extraction pass.

EXT Randomized controlled trial of early-2025 AI tooling on experienced open-source developers (arXiv:2507.09089), conducted by a nonprofit AI evaluation organization: 16 developers, 246 issues, repositories averaging over a million lines and ten years of history. 19% slowdown; forecast 24% speedup, post-hoc estimate 20% speedup; economists 39%, machine-learning experts 38%; 75% of developers slowed; 47% more lines of code per forecast hour; under 44% of generations accepted; 9% of time reviewing generated output. Graded: strong; the only randomized field trial in the corpus. Authors scope the result to experts in mature codebases.

EXT Two-year longitudinal study of developer productivity with and without an AI coding assistant (arXiv:2509.20353): 105 weeks of commit activity, September 2022 – September 2024, paired with perception. All users perceived no change or improvement; commit activity showed no significant change (rho = 0.17, p = 0.406). Graded: moderate; before-after design with self-selected adopters, a limitation the authors name.

EXT Practitioner survey on productivity with an AI assistant (arXiv:2602.03593): 86% satisfied or very satisfied, ~60% reporting under one hour saved per week, correlation r = 0.34. Graded: moderate; self-report only, cross-sectional.

EXT Multi-case study of generative AI in agile teams (arXiv:2602.13766): 13 months, October 2023 – November 2024; completed story points 281 → 447 (+59.1%); committed lines of code flat (p = 0.928); code-quality metrics moving in opposite directions across squads. Graded: moderate; before-after, few cases, but the divergence is the point. Authors state that activity measurement alone would have concluded no impact.

EXT Quasi-experimental study of generative AI in collaborative open-source development (arXiv:2410.02091): 24 months, January 2021 – December 2022; matched design; increased merged contributions alongside an 8% increase in coordination time; raw defect counts rising while per-contribution defects stayed flat. Graded: strong for the coordination-cost result.

EXT Systematic review of large language model assistant effects on developer productivity (arXiv:2507.03156): 39 primary studies, literature window January 2014 – December 2024, mapped to a five-dimension productivity framework. Suggestion-acceptance rate misleads in isolation; relays a 70-company industry study reporting throughput correlating negatively with code quality at r = −0.45. Graded: strong for the metric critique; a review, so it inherits its inputs' limitations.

EXT Survey of developer productivity with generative AI across five dimensions (arXiv:2510.24265): two-week survey, 415 developers across 56 open-source communities; effort redistributed into review burden and verification load behind an appearance of speed. Graded: moderate; self-report, short window.

EXT Longitudinal mixed-methods study of AI coding assistants (arXiv:2605.23135): two questionnaires six months apart, 158 and 101 respondents, matched cohort of 95. 84% reported improved productivity at both points; the share reporting a worse experience in at least one dimension rose 14% → 27%, with no recovery out of that group; 82% reported writing less code. Names "supervisory engineering work" as an emerging category. Graded: moderate; self-report, and the authors state the 84% describes continuing users only, since anyone who abandoned the tools never entered the sample.

EXT Study of human grading consistency (arXiv:2409.12967): 28 graders, two batches of 20 assignments, with a duplicate planted so each grader re-graded the same work. 24 of 28 (85.7%) self-

reported consistency; of the 22 who never noticed the duplicate, 1 (4.5%) gave identical grades; one further grader matched grades only after spotting the repeat. Standard inter-rater agreement statistics are inapplicable at two ratings per item, and self-review did not improve consistency. Graded: strong; the design is what earns it.

EXT Two-part study of large language model explanations in human decision-making (arXiv:2604.03237): a within-subject study (n = 27) with objective accuracy 49.8% → 48.8% between prediction and explanation stages while confidence rose from a 3.66 baseline to 3.81; a separate randomized study (n = 100) with error recovery 16.2% in the explanation condition (condition differences p = .055); a selective-automation rule using a prespecified ten-point probability gap reached 69.5% in post-hoc evaluation. Graded: strong for the confidence-accuracy divergence; design labels per study, as corrected in Revision 2.

EXT Clinical case study of expert reliance on AI assistance (arXiv:2502.03482): radiologists rated the system 2.4/5 and 2.1/5 across two studies while its measured accuracy (69.3%, 76.0%) exceeded their unaided 63.2% in the study that measured both (the second had no human-alone arm); performance feedback before decisions did not significantly improve combined performance; in the independent-judgment-first workflow radiologists kept their initial opinion 79.6% of the time when the machine disagreed, with feedback varied concurrently. Graded: moderate — medical imaging, so the mechanism transfers and the magnitudes do not.

EXT Empirical study of user trust and team performance in plan-then-execute agent systems (arXiv:2502.01390): stated trust 3.52/5 against calibrated trust of 0.03–0.27 on high-risk tasks with imperfect plans; strict action-sequence matching reported as a metric that broke by penalizing benign redundant actions. Graded: moderate; simulated tasks.

EXT Randomized study of coding agents against chatbot assistance (arXiv:2607.26375): initial accuracy effect size d = 1.4 favoring the agent; comprehension deficit d = 0.9 overall, widest on code-specific questions at 0.642 against 0.951; self-rated understanding 3.3 against 4.2; agent rated more helpful at 4.7; on an extension task without the agent (chatbot available to both groups) the advantage reverses when initial differences are controlled. Graded: strong; laboratory, ~2 hours per participant.

EXT Study of expert performance in self-organizing multi-agent teams (arXiv:2602.01011): ranking-error and benchmark-accuracy measurements showing self-organizing discussion failing to leverage expertise, with the correct baseline established as the single best individual rather than the group average. Graded: moderate; benchmark tasks.

EXT Off-policy evaluation in production at Adyen, a payment processing company (arXiv:2501.10470): a year of completed live A/B tests compared against offline estimates over billions of logged interactions. Inverse propensity scoring and its self-normalized variant sustained Pearson correlation above 0.8; the direct reward-model method correlated negatively; the doubly robust

estimator was dragged to near zero by inheriting that reward model; variance largely disappears above one million records. 58% of the company's A/B tests were flat or inconclusive, quantified as more than 20 weeks per year. Estimated 9–54 million incremental transactions over six months from faster identification of winners. Graded: strong, production, self-reported by the operator. Re-read in full, independently of the extraction pass.

EXT User simulation for evaluating preference-elicitation policies at a large recommender platform (arXiv:2409.17436): simulation predicted +1.36% [−0.11%, 2.83%] selections against +1.17% [−0.03%, 2.36%] measured post-launch across 500,000 users; three sub-period live experiments each ordered three candidate policies differently, only one matching the full launch, while the simulation matched it. Authors name counterfactual robustness of the user model as the precondition. Graded: strong, production. Re-read in full, independently of the extraction pass.

EXT Controlled study of interaction topology in multi-agent decision systems (arXiv:2605.01147): 5,760 synthetic credit applications, four fixed roles, 24 orderings, model families from 3 billion to 72 billion parameters. Approval rates spanning 59 percentage points on the smallest model and 21 points at 70 billion and above; agreement above 99.9% with error correction below 0.1% at the largest scale; parallel independent voting eliminating the ordering effect while saturating at 98% approval on the smallest model. Graded: moderate — a position paper, single synthetic domain, internally inconsistent in places; the weakest of the three agent studies, and Finding 15's direction survives without it.

EXT Empirical study of diversity in multi-agent idea generation (arXiv:2604.18005): 1,000 proposals per configuration across 20 topics; junior-led horizontal structures maximizing measured diversity, designated-senior and interdisciplinary structures minimizing it; per-agent diversity utilization falling from 1.03 to 0.47 as groups grew from three to seven. Diversity metrics validated against five human annotators at 87% agreement for the primary metric; idea quality scored by a model judge, not by humans. Graded: moderate — the quality half is unvalidated by humans, and the authors say so.

EXT Systematic study of scaling behavior in multi-agent systems (arXiv:2606.00655): agent counts one to eight, two model families, with a token-padding control isolating coordination overhead from context length. Inverted-U performance curve peaking around four to five agents; quadratic token cost. Graded: moderate — 17 questions per subject with three repeats, small enough that reported differences of a few points represent a handful of items.