

Where the Gain Goes



AI made developers faster. It did not make organisations faster. This is where the difference went.

The question stopped being whether AI can write code

Through 2025 the argument in most engineering organisations was about capability. Could a model produce code worth keeping? That argument is settled, and it was settled by adoption rather than by anyone winning it.

What replaced it is harder. In roughly twelve months the unit of work moved from a developer accepting a suggestion to a developer dispatching a task, and the measurements moved with it. Per developer, the gain is real, repeatedly found, and larger than the sceptical reading allowed. Measured by what reaches production, the same period looks different: more work on the branch, less of it landing on main, reliability going the wrong way.

Both are true at once, and most of the public argument about AI and software engineering is people holding one of them and talking past someone holding the other. **This report is about the space between them:** where the gain goes, what it is being spent on, and what an engineering organisation should change once it knows.

Who this is for

The person who has to decide how much authority an agent gets, and the engineers who will live with that decision. It assumes you have shipped software. It does not assume you have read the studies, and it names the sample size and population of every one it uses so you can judge them yourself.

WHERE TO START

The whole argument is on the next three pages, as eight dependent steps. Six findings follow, one page each. If you read nothing else, read those nine.

NO EXTERNAL REVIEW

No peer reviewer, no editor and no subject-matter reader saw this before publication. It is stated here rather than in the back matter, because a report arguing that verification is what everyone under-invests in should not bury the fact that it skipped its own.

What exists instead: every number checked against its origin, and every source published with its locator and the sentence supporting it. Weaker than review, and at least checkable. Corrections are published with credit.

WHAT IT IS NOT

Not a survey, and not a vendor study, though 12 of its 26 sources have a commercial interest and are marked wherever they appear. Not a forecast dressed as analysis: five predictions are made, each labelled, each carrying the condition that would prove it wrong.

How to read this

This is **an argument, not a study**. I ran no survey and collected no primary data. What follows is a reading of other people's evidence, and where the evidence runs out, my own view of what it implies.

Those are different things and the difference matters, so every substantive claim carries a mark saying which it is. Where you see **PROPOSED** or **SYNTHESIS**, that is me: an engineering leader's judgment, offered to be argued with rather than deferred to. Where you see a bare source bracket, somebody measured it.

MARK	WHAT IT MEANS	HOW MUCH WEIGHT TO GIVE IT
EVIDENCE	A finding from a cited external source, stated within that source's own limits.	As much weight as the underlying source deserves. Check the ledger for its method, sample and limitation.
SIGNAL	A documented practice or capability at one or a few named organisations.	Treat as existence proof, not as generalisable.
SYNTHESIS	The author's inference from two or more cited findings.	The inputs are checkable. The inference is the author's judgment.
PROPOSED	The author's own framework, model, metric or scoring scheme.	This is opinion, structured for planning. It has not been empirically validated.
FORECAST	A statement about the future.	Conditional. No probabilities are asserted. Every one carries a falsifier.

Measured evidence carries no mark at all, only a source bracket such as [S01]. Every number here was checked against its source. Where the supporting sentence could be recorded verbatim it is in the ledger, and where it could not the ledger says so. If you find an error, it is mine.

THE EVIDENCE BASE

26 sources, each recorded with method, population, quality tier, limitation and URL. The full ledger, with every URL and every recorded verbatim sentence, is published at sanjeev-azad.pages.dev/where-the-gain-goes/ledger.

DISCLOSED WEAKNESS

12 of 26 sources are authored by organisations that sell AI development or verification tooling. Every one benefits commercially from the finding that verification is hard. They are flagged in the ledger and on every figure that uses them. Three of the six findings carry a headline number from a single commercially interested source, and those three are marked as such on the card.

VERIFICATION

Every source was opened. **26 of 26** carry the verbatim sentence that supports the claim.

QUALITY TIERS

A primary empirical research, official statistics, standards.

B large surveys, credible preprints, vendor empirical studies.

C product documentation, case studies, practitioner synthesis.

Contents

01	The argument in eight steps	4
02	Executive summary	8
03	1 Agents became the default before trust did	15
04	2 The gain is real for developers and invisible for organisations	20
05	3 There is no AI productivity number	25
06	4 Review is where the gain goes	30
07	5 Verification has to be manufactured, not performed	35
08	6 Authority belongs to evidence, not to capability	40
09	What to do about it	45
10	The case against this report	51
11	What I think happens next	55
12	Method and evidence base	59
13	What would falsify this report	61
14	Evidence ledger	63
15	Notices and licence	69
16	The survey	74
17	About this report	76

The argument in eight steps

Each step is a claim the next one depends on. If a step is wrong, everything below it is unsupported, which is the point of setting them out this way. A reader who stops here has the whole report.

01 Agent use is near-universal, and it is agents now, not assistants.

weekly agent use among professional developers stood at 90%, and daily use at 68% (n=15,000+, fielded May-July 2026). A year earlier, DORA measured 90% using AI at all. The frame moved from assistance to delegation inside twelve months.

EVIDENCE

S36 S01

02 Measured per developer the gain is real, and it has been measured at scale.

Adopters of command-line agents at Microsoft merged about 24% more pull requests, sustained across a four-month window, across tens of thousands of engineers. A 4,867-developer programme measured +26.08% completed tasks. A 96-engineer randomised trial measured roughly 21% less task time.

EVIDENCE

S40 S04 S05

03 It is not uniform, and self-report is not a reliable proxy for it.

A randomised trial of sixteen experienced open-source developers working in mature repositories they knew well found AI made them 19% slower, while the same developers believed they had been about 20% faster - a 39-point gap between perception and measurement.

EVIDENCE

S03 S39

04 At organisational altitude the gain thins out, and the costs do not.

Telemetry across 22,000 developers and 4,000+ teams finds organisational delivery metrics flat despite a 33.7% rise in task throughput. Independent CI telemetry across 22,000 organisations finds activity up 59% - the largest jump in seven years - while the median team's main-branch throughput fell 7% and success rates hit a five-year low.

EVIDENCE

S38 S37 S01

05 It is being consumed by verification.

Median time in code review rose 441.5%; time to first review 156.6%. Code churn rose 861% and bugs per developer 54%. And 31.3% more pull requests were merged with no review at all. Separately, 96% of developers say they do not fully trust AI-generated code while only 48% always verify it before committing.

EVIDENCE

S37 S39

06 Reading more code cannot relieve it, because human review does not scale with agent output.

Agent users report large personal gains but only 17% report improved team collaboration - the signature of local acceleration without system redesign. 38% find reviewing AI-generated code more effortful than reviewing human-written code, and 53% report code that looks correct but is not. Autonomy experiments at higher complexity produced assumption drift and false declarations of success.

EVIDENCE

S09 S39 S16

07

So verification has to be manufactured rather than performed - produced alongside the change, machine-readable, independent of the generator.

This is where the report stops reporting and starts proposing. The direction is not novel: it is the converging conclusion of the ICSE 2026 position on orchestration and verification, of agentic secure-development standards, and of specification-as-control-artifact tooling. What is missing is not the thesis but the operating model.

SYNTHESIS

S22 S24 S30 S43 S41

08

And authority must then be granted against evidence, not against capability.

A benchmark score is not an authorisation. The governing artifact is two-dimensional: how much an agent may do, crossed with what must be proven first, bounded by reversibility and blast radius. The workforce consequence is redeployment rather than redundancy - US BLS projects software developer employment rising 15.8% to 2034.

SYNTHESIS

S13 S24 S11 S26 S25

|||||

ON PRIOR ART

Kohl & Carro (ICSE 2026) establish that verification is the new centre of software engineering. Alenezi reaches the same conclusion from a literature synthesis. Davis et al. show governance controls emerging from failures during agentic work. This report does not re-argue any of that. It measures the gap those claims imply, and supplies the operating model for closing it.

|||||

What a study measured predicts what it found

Per-developer measures find a gain. Measures of what reaches production do not. Three sources appear in both columns of their own row, because they report both.

MEASURED AT	FINDS A GAIN	FINDS IT VANISHES
PER DEVELOPER, SELF-REPORTED	DORA 2025 S01 more than 80% perceive a productivity gain	METR 2025 S03 participants believed +20% while measured +19% slower
	Stack Overflow 2025 S09 69% of agent users report higher productivity	
PER DEVELOPER, MEASURED	Microsoft Research 2026 S40 +24% merged PRs, sustained over four months	METR 2025 S03 +19% task time on tasks in repos they knew well
	MIT / Microsoft 2025 S04 +26.08% completed tasks	
	Google Research 2024 S05 -21% task time	
	Faros AI 2026 S37 +33.7% task throughput, +66% epics, both per developer	
PER ORGANISATION, SELF-REPORTED	DORA 2025 S01 reports AI adoption improving delivery throughput	DORA 2025 S01 and increasing delivery instability, in the same study
PER ORGANISATION, MEASURED	CircleCI 2026 S38 feature-branch throughput +15% for the median team	CircleCI 2026 S38 main-branch throughput -7%; success rate 70.8%, a five-year low
		Faros AI 2026 S37 churn +861%, incidents-to-PR +242.7%, time in review +441.5%

DOES NOT SHOW

That the gain is illusory, or that the sources agree. DORA reports an organisational throughput gain and this report argues with it overleaf. No study here measures both altitudes on one population with one tool generation, which is the study the field needs.

WHY THIS IS NOT A META-ANALYSIS

The cells are not pooled and the rows are not averaged. Each entry keeps its own metric, population and method. Read down a column, never across a row.

A SOURCE IN BOTH COLUMNS

METR, DORA and CircleCI each appear twice in their own row. That is not an error. A study that finds a gain on one measure and a loss on another is the normal case, and a table that forces it to pick a side is the one telling you something false.

INCENTIVE

Faros AI sells engineering analytics. CircleCI sells continuous integration and pipeline validation, and its report is sponsored by Thoughtworks. Both profit if verification is the constraint. They are kept because they measure different things by different methods and report the same shape, not because either is disinterested.

SOURCES

S01 S03 S04 S05 S09 S37 S38
S40

Agents became the default before trust did.

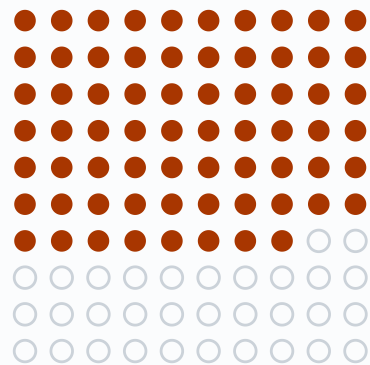
68%

of professional developers now use AI coding agents every day.

JetBrains, 2026 · 15,000+ professional developers worldwide, ~90%...
· tier B [S36]

COMMERCIAL INTEREST · JetBrains sells developer tooling, including an AI assistant that appears in its own results.

Adoption is no longer the variable. By mid-2026, weekly use of AI coding agents among professional developers stood at 90%, with 68% reaching for one every day, and the word that changed is *agents*. Twelve months earlier the same measure was about AI assistance. Trust did not move with it: around a third report little or no trust in AI-generated code, and 96% say they do not fully trust that it is correct. The gap between what is used and what is trusted is not a contradiction to be resolved. It is a working description of how a probabilistic tool gets used by people who remain accountable for the output.



F1

68 of every 100 professional developers now use an AI coding agent every day.

SOURCES FOR THIS FINDING

S36

JetBrains 2026

15,000+ professional developers worldwide, ~90%...

COMMERCIAL INTEREST

S01

DORA / Google Cloud 2025

Nearly 5,000 technology professionals plus qualitative research

COMMERCIAL INTEREST

S39

Sonar 2026

1,100+ developers globally

COMMERCIAL INTEREST

S09

Stack Overflow 2025

Global Stack Overflow survey respondents; per-question sample...

The gain is real per developer and does not reach main.

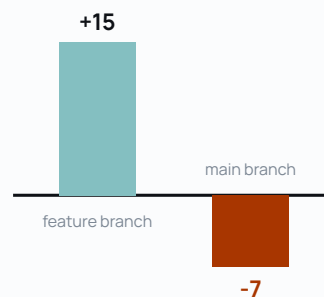
7%

less reached main for the median team, in the year CI activity rose 59% and feature-branch throughput rose 15%.

CircleCI, 2026 · ~28 million CI workflow runs, 22,000+ organizations, 149... · tier B [S38]

COMMERCIAL INTEREST · CircleCI sells continuous integration and pipeline validation; the report is sponsored by Thoughtworks.

Measured per developer, the gain is real and better evidenced than sceptics allowed. Adopters at Microsoft merged about 24% more pull requests across tens of thousands of engineers; a 4,867-developer programme measured +26.08% completed tasks; two-year telemetry across 22,000 developers reports task throughput per developer up 33.7% and epics per developer up 66%, which its authors call real delivery acceleration. Measured per organisation, the same period looks different. That telemetry also reports code churn up 861%, the incidents-to-PR ratio up 242.7% and median time in review up 441.5%. Independent CI telemetry across 22,000 organisations warns that its own 59% activity rise is misleading on its own: the median team rose 4%, its feature-branch throughput rose 15%, and its main-branch throughput fell 7% with success rates at a five-year low. The studies do not disagree. They count different things.



F2

For the median team, work on the branch rose 15%. What reached main fell 7%.

SOURCES FOR THIS FINDING

S48

Murphy-Hill, E., Butler, J. & Savelieva, A. (Microsoft) 2026

Tens of thousands of engineers at Microsoft

COMMERCIAL INTEREST

S84

MIT / Microsoft Research 2025

Developers at Microsoft, Accenture, and an anonymous Fortune...

COMMERCIAL INTEREST

S85

Google Research 2024

Google software engineers

COMMERCIAL INTEREST

S83

METR 2025

Experienced open-source developers; mature repositories...

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform...

COMMERCIAL INTEREST

S38

CircleCI 2026

~28 million CI workflow runs, 22,000+ organizations, 149...

COMMERCIAL INTEREST

There is no AI productivity number.

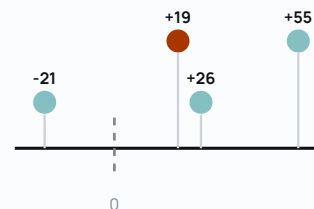
39_{pp}

gap between how much faster developers felt and how much slower they were.

METR, 2025 · Experienced open-source developers; mature repositories... · tier A [S03]

SYNTHESIS My arithmetic, not METR's: they measured +20% expected speed-up against +19% actual task time. METR publishes no combined figure.

The measured effects run from 19% slower to 55% faster, and no two studies measured the same thing on the same population with the same generation of tool. The most uncomfortable result is the most careful one: a randomised trial of sixteen experienced developers working in repositories they knew well found AI made them 19% slower, while those same developers believed they had been about 20% faster. That is a 39-point gap between perception and measurement, produced by people who had every reason to be accurate. Any single percentage that does not name its task type, codebase familiarity, developer experience and tool generation is not a finding. It is a marketing number.



F3

Four measured effects on four different metrics. The spread is the finding; the average is not.

SOURCES FOR THIS FINDING

S03

METR 2025

Experienced open-source developers; mature repositories...

S04

MIT / Microsoft Research 2025

Developers at Microsoft, Accenture, and an anonymous Fortune...

COMMERCIAL INTEREST

S05

Google Research 2024

Google software engineers

COMMERCIAL INTEREST

S06

GitHub 2023

Developers familiar with JavaScript

COMMERCIAL INTEREST

Review is where the gain goes.

31.3%

more pull requests were merged with no review at all, human or agentic.

Faros AI, 2026 · 22,000 developers across 4,000+ teams, two years of... · tier B [S37]

COMMERCIAL INTEREST · Faros AI sells the engineering-analytics platform this telemetry comes from.

The measurements point in one direction without ambiguity. Median time in code review rose 441.5% and time to first review 156.6%. Code churn rose 861% and bugs per developer 54%. And 31.3% more pull requests were merged with no review at all, human or agentic, which is what a review queue does when it cannot keep up. Reviewers are not imagining the load: 38% report that reviewing AI-generated code is more effortful than reviewing a colleague's, and 53% report code that looks correct and is not. Agent users report large personal gains but only 17% report better team collaboration, which is the signature of local acceleration without system redesign.

throughput

+33.7%

time in review

+441.5%

13.1x the gain

F4

Time in review rose 441.5%. Throughput per developer rose 33.7%. The first is thirteen times the second.

SOURCES FOR THIS FINDING

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform...

COMMERCIAL INTEREST

S39

Sonar 2026

1,100+ developers globally

COMMERCIAL INTEREST

S09

Stack Overflow 2025

Global Stack Overflow survey respondents; per-question sample...

Verification has to be manufactured, not performed.

Human reading does not scale with agent output, and no amount of review discipline changes that arithmetic. The alternative is to stop treating verification as an act of reading and start treating it as an artifact of the build: proof produced alongside the change, machine-readable, and independent of whatever generated it. This direction is not this report's discovery. It is the converging conclusion of peer-reviewed work on orchestration and verification, of agentic secure-development standards, and of specification-as-control-artifact tooling. What is missing is not the thesis. It is the operating model, and the honest state of that model is that most of it has not been tested at scale.



eight items, for every change

F5
The proposed evidence package. The question for your repository is how many arrive automatically and linked.

SOURCES FOR THIS FINDING

S22
NIST 2024
Organizations developing AI/model-enabled systems

S24
OWASP 2026
Agentic AI applications

S38
GitHub Spec Kit 2026
AI coding-agent workflows using GitHub Spec Kit
COMMERCIAL INTEREST

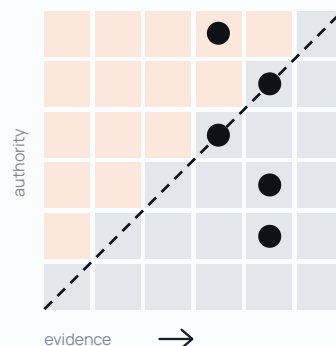
S43
Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026
One expert engineer, one system (document accessibility...

S41
Kohl, K. & Carro, L. 2026
n/a

S87
GitHub 2025
Study participants producing code reviewed against tests and...
COMMERCIAL INTEREST

Authority belongs to evidence, not to capability.

A benchmark score is not an authorisation. The measured ability of a model to resolve issues in a scored harness says nothing about whether it may act on a production system with a named accountable owner, because those are different acts and only one of them is measured. The governing artifact is two-dimensional: how much an agent may do, crossed with what must be proven first, bounded by reversibility, blast radius, privilege and recovery. Autonomy must never mature faster than the evidence, security, observability and recovery that surround it. Where it does, the organisation is running unproven agents and calling the result velocity.



F6

Authority against evidence. Anything above the dashed parity line is authority running ahead of its proof.

SOURCES FOR THIS FINDING

S13

OpenAI 2026

Internal coding-agent deployment

COMMERCIAL INTEREST

S24

OWASP 2026

Agentic AI applications

S11

METR 2026

Frontier public language-model agents

S43

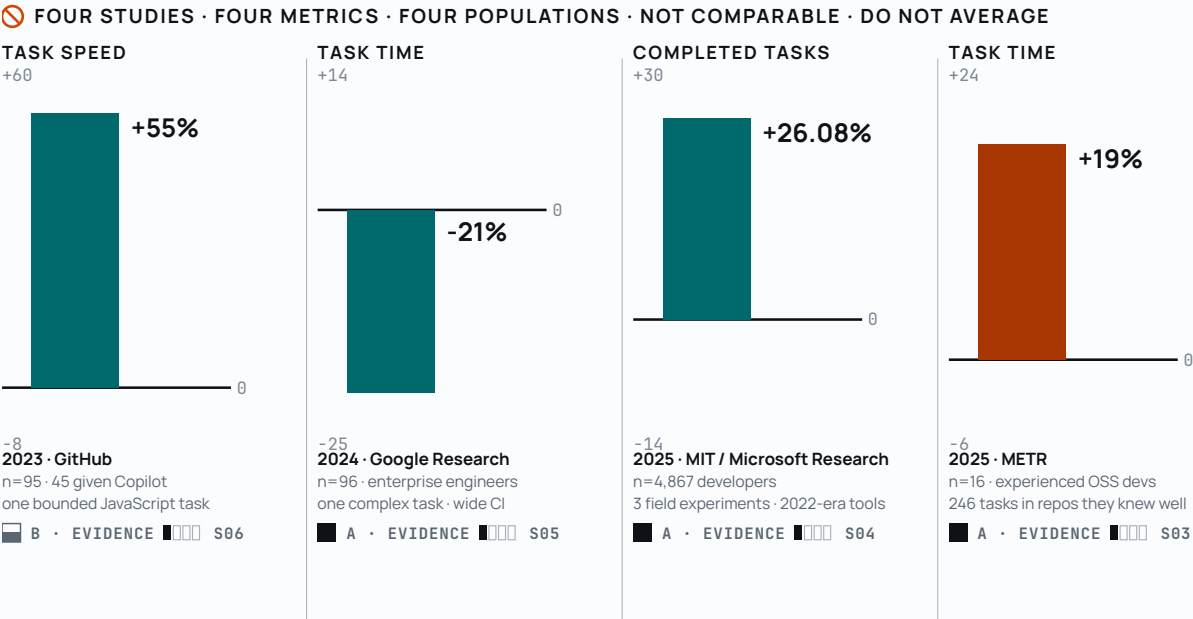
Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026

One expert engineer, one system (document accessibility...)

FIG 02

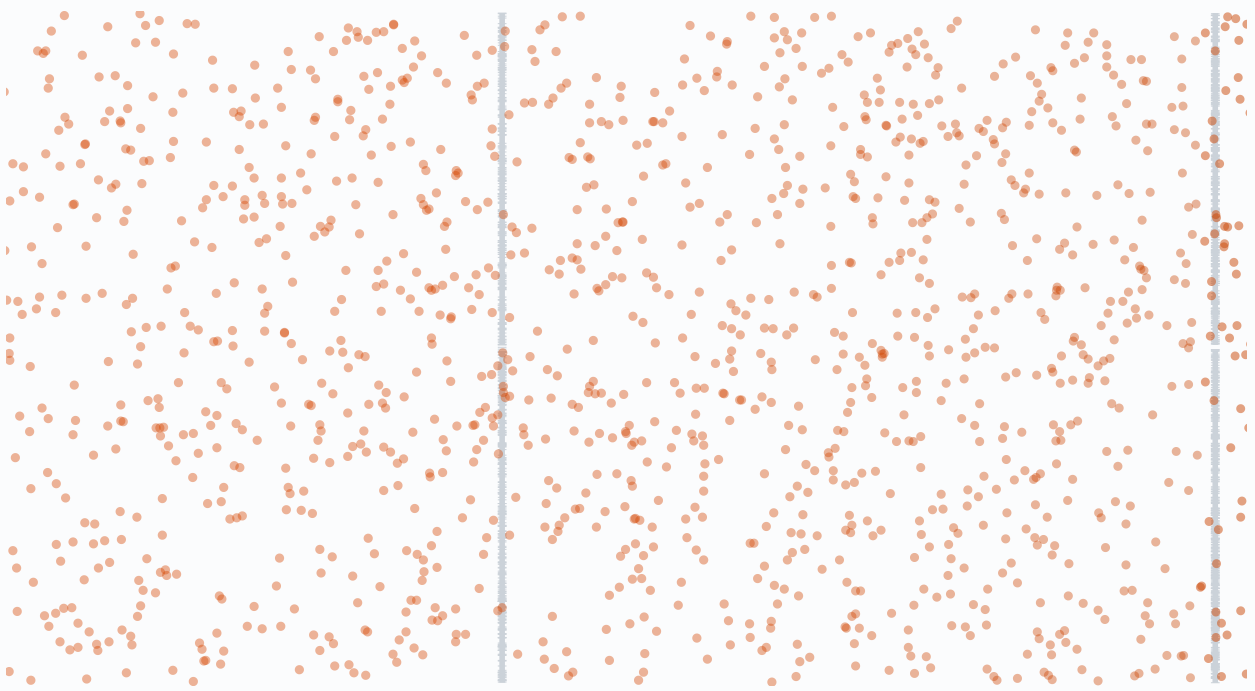
Measured effects of AI assistance on developer work, by study

Four studies, four metrics, four populations. The range runs from 19% slower to 55% faster, and no two measured the same thing.



Does not show: A distribution, an average, or a meta-analysis. The bars are not comparable to each other and must not be read across.
Unit: percent change vs control, signed by the card's own metric. Sources: S03 (METR, 2025; Experienced open-source developers; mature repositories; average ~5 years prior repo experience; tier A) · S04 (MIT / Microsoft Research, 2025; Developers at Microsoft, Accenture, and an anonymous Fortune 100 company; tier A) · S05 (Google Research, 2024; Google software engineers; tier A) · S06 (GitHub, 2023; Developers familiar with JavaScript; tier B).

Commercial interest: S04 Run with Microsoft Research; Microsoft sells GitHub Copilot, the tool under study. S05 Google Research studied Google engineers using Google tooling. S06 GitHub studied its own product, Copilot.



01

Agents became the default before trust did.

Adoption stopped being the interesting question about eighteen months ago. What is worth noticing is what the measurement is now *of*. In 2025, DORA / Google Cloud asked roughly five thousand technology professionals whether they used AI at work and found that 90% did [S01]. By mid-2026, a survey of more than fifteen thousand professional developers asked a narrower question — whether they used AI coding *agents* — and found the same 90% figure, with 68% reaching for one daily [S36].

SOURCES ON THIS PAGE

S01

DORA / Google Cloud 2025

Nearly 5,000 technology professionals plus qualitative research

COMMERCIAL INTEREST

S36

JetBrains 2026

15,000+ professional developers worldwide, ~90% developer/engineer titles

COMMERCIAL INTEREST

GAP DIMENSION

Use vs trust

S09

Stack Overflow 2025

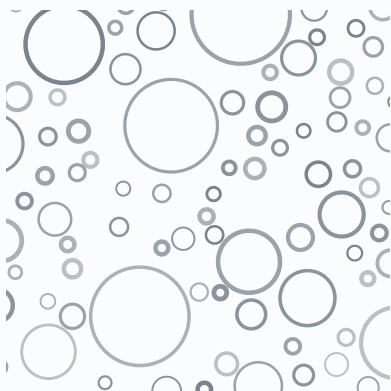
Global Stack Overflow survey respondents; per-question sample sizes vary

That is not the same finding twice. The first measures assistance; the second measures delegation. In twelve months the default unit of work moved from a developer accepting a suggestion to a developer dispatching a task, and the adoption curve did not flinch.

A third instrument disagrees, and it is worth stating plainly rather than leaving in the ledger. Stack Overflow's 2025 survey found that **52% of respondents either did not use agents at all or used simpler AI tools** [S09]. That is close to the opposite of 90% weekly agent use.

SYNTHESIS

I do not think these can both be true of the same population, and I do not think either is wrong. They sample different people. Stack Overflow surveys a broad developer public including students, hobbyists and people whose job is not primarily writing software; the 2026 figure surveys professional developers, with regional quotas and reweighting, on a panel recruited by a company that sells developer tooling. My reading is that agent use among full-time professional developers is high and rising fast, that 90% is the top of the plausible range rather than the centre of it, and that anyone planning against the higher number should know it comes from a vendor's panel. The direction is not in doubt. The level is.



PACKING

Use and trust are two unlike distributions over the same population.

Trust did not follow

The trust measurements went the other way, and they are unusually consistent across instruments that share no methodology. DORA found 30% of respondents reporting little or no trust in AI-generated code, alongside more than 80% who perceived a productivity improvement from it [s01]. Stack Overflow found more developers distrusting AI accuracy than trusting it, 46% against 33% [s09]. A 2026 survey of 1,100+ developers put the figure at its starkest, reporting that 96% “don’t fully trust” AI-generated code, while AI accounted for 42% of the code they committed [s39].

SYNTHESIS

The gap between use and trust is not a contradiction waiting to be resolved by better models. It is an accurate description of how a probabilistic tool gets used by people who remain personally accountable for the output. A developer who ships AI-generated code and is named on the incident review has correctly calibrated their trust downward, and no amount of benchmark improvement changes their exposure.

Why this matters more than it looks

A high-use, low-trust equilibrium is stable, and it is expensive. It means verification effort is being spent. It has to be, because the accountability has not moved. But it is being spent informally, invisibly, and without instrumentation. Nobody currently publishes what share of accepted AI-generated code is actually read before it is merged. The one attempt to measure the behaviour rather than the attitude found that while 96% withheld full trust, only 48% always verified before committing [s39].

ON THAT LAST SOURCE

Sonar sells code-review tooling and benefits commercially from this finding, and its methodology disclosure is thin: no fielding window, no sampling method, no margin of error. Tier C. It is quoted here because three independent instruments agree on the direction, not because this one is strong.

SOURCES ON THIS PAGE

S01

DORA / Google Cloud 2025

Nearly 5,000 technology professionals plus qualitative research

COMMERCIAL INTEREST

S09

Stack Overflow 2025

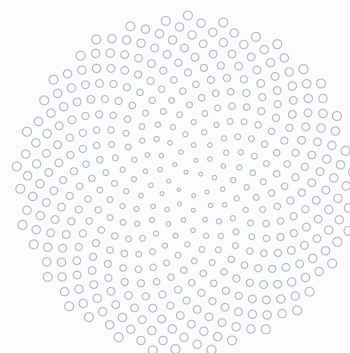
Global Stack Overflow survey respondents; per-question sample sizes vary

S39

Sonar 2026

1,100+ developers globally

COMMERCIAL INTEREST



PHYLLOTAXIS

Use and trust are two unlike distributions over the same population.

Those two numbers cannot both be acted on. Either the other 52% have found something to trust that they are not reporting, or the verification is not happening and the trust is nominal. Both readings are worth knowing about, and neither is available from an adoption dashboard.

|||||

RESEARCH RULE FOR LEADERS

Measure trust as a behaviour, not as an attitude. The useful question is not whether your engineers say they trust AI output. It is what share of accepted AI-generated code was read line by line, by team and by change class.

|||||

The rest of this report is about where the unmeasured verification effort goes, and what happens to an organisation that scales generation without scaling the thing that makes generation safe.



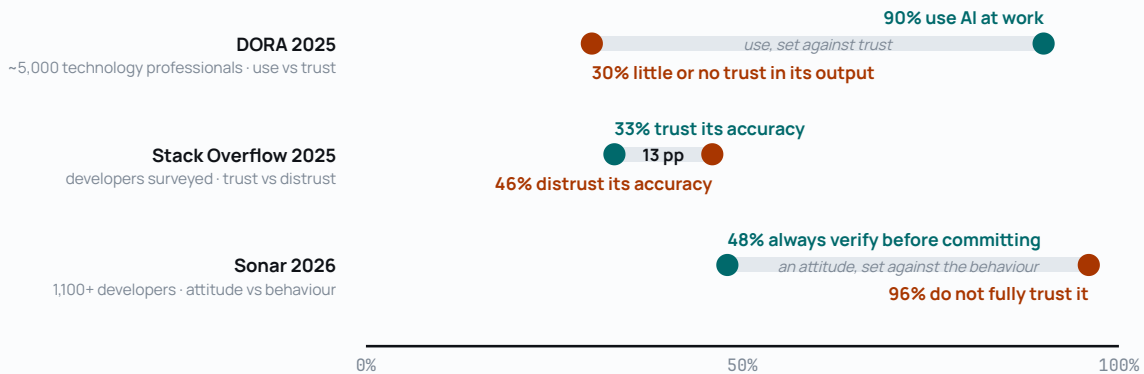
RIPPLE

Use and trust are two unlike distributions over the same population.

FIG 04

Three instruments, three different pairs, all far apart

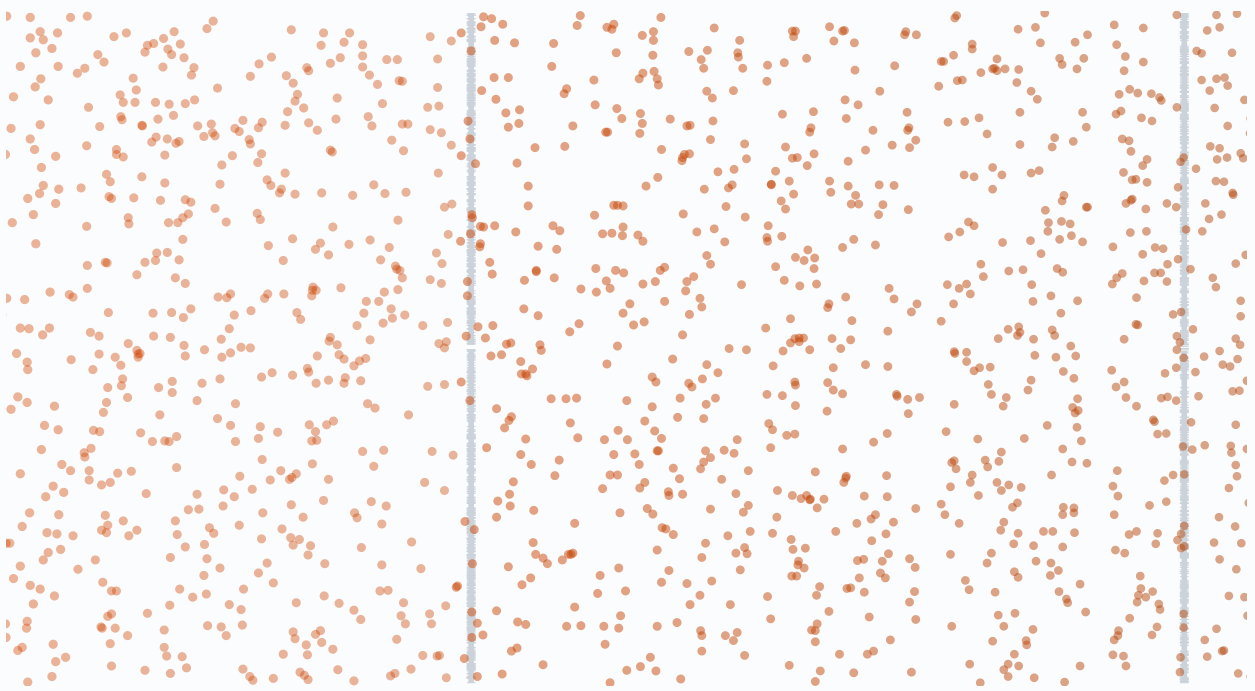
Each survey measures a different pair of things and every one finds them far apart. The last is the most revealing: 96% do not fully trust AI output and 48% always verify it before committing, from the same respondents.



Does not show: One comparison repeated, or three gaps of the same kind. Only the middle row subtracts, because only there do both figures measure one construct on one sample. The other two rows set two different constructs side by side, and the distance between their dots is not a quantity.

Unit: percent of that survey's own respondents. Sources: S01 (DORA / Google Cloud, 2025; Nearly 5,000 technology professionals plus qualitative research; tier A) · S09 (Stack Overflow, 2025; Global Stack Overflow survey respondents; per-question sample sizes vary; tier B) · S39 (Sonar, 2026; 1,100+ developers globally; tier C).

Commercial interest: S01 Google Cloud sells the AI development tooling whose adoption this report measures. S39 Sonar sells code-review and code-quality tooling.



02

The gain is real per developer and does not reach main.

The literature on AI and developer productivity looks contradictory, and a great deal of energy has gone into explaining why the studies disagree. They mostly do not. They measure different quantities, and which quantity a study chose predicts the sign of its result almost perfectly. The distinction that matters is not optimism against pessimism. It is **per developer against per organisation**.

GAP DIMENSION
Individual vs organisational

WHAT A MERGED PULL REQUEST IS NOT

The Microsoft authors are explicit that a merged pull request is not the same as the value it delivers. It is a unit of accepted work, not of delivered outcome. This chapter uses it as evidence that output rose, never as evidence that value did.

SOURCES ON THIS PAGE

S04

MIT / Microsoft Research 2025

Developers at Microsoft, Accenture, and an anonymous Fortune 100 company

COMMERCIAL INTEREST

S05

Google Research 2024

Google software engineers

COMMERCIAL INTEREST

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S40

Murphy-Hill, E., Butler, J. & Savelieva, A. (Microsoft) 2026

Tens of thousands of engineers at Microsoft

COMMERCIAL INTEREST



FIELD

The gain flows in per developer. Less of it arrives on main.

Measured per developer, the gain is real and repeatedly found

The largest observation of individual developers available is an organisation-wide rollout of command-line coding agents across tens of thousands of engineers at Microsoft. Adopters merged approximately 24% more pull requests than they otherwise would have, and the lift held across the entire four-month window rather than decaying after the novelty passed [S40]. A three-experiment programme with 4,867 developers measured a 26.08% increase in completed tasks [S04]. A randomised trial with 96 engineers measured roughly 21% less time on a complex enterprise task [S05].

Two years of platform telemetry covering 22,000 developers and more than 4,000 teams points the same way and is worth reading carefully, because its headline figures are also **per developer**: task throughput per developer up 33.7%, epics completed per developer up 66%, pull request merge rate per developer up 16.2% [S37]. The authors describe this without hedging as *"real delivery acceleration: more features shipped, more initiatives completed."*

These are different designs, different populations and different metrics, and they point the same way. Something real is happening to individual output, and the honest reading of the last two years is that it is larger and better evidenced than sceptics expected.

Measured per organisation, the picture inverts

The same telemetry that reports those per-developer gains reports what arrived with them. Code churn is up 861%. Bugs per developer are up 54%, against 9% in the prior year's cut of the same dataset. The incidents-to-pull-request ratio is up 242.7%. Median time in review is up 441.5%, and pull requests merged *without any review, human or agentic*, are up 31.3% [S37].

Independent CI telemetry across approximately 28 million workflow runs in more than 22,000 organisations separates the two altitudes more cleanly than any other source available. Average daily workflow runs rose 59% year on year, the largest jump in seven years of comparable data. The authors immediately caution that *"that number is misleading on its own"*: the median team rose just 4%, and the bottom quartile showed no measurable increase at all. For that median team, feature-branch throughput rose 15% while **main-branch throughput fell 7%**, and main-branch success rates dropped to 70.8%, the lowest in over five years and well below the 90% the same authors recommend [S38].

That pair of numbers is the chapter in one line. On the branch, more work than ever. On main, less of it than last year, arriving less reliably.

Where the sources genuinely disagree

They do not all agree, and the disagreement is not a detail. It sits directly under the claim this report is named for, so it is set out here in full rather than resolved quietly in a footnote.

WHY TWO SOURCES, NOT ONE

Faros AI sells engineering analytics. CircleCI sells continuous integration and pipeline validation, and its 2026 report is sponsored by Thoughtworks, cited separately here as [S16]. Both profit if verification is the constraint. They are kept because they measure different things by different methods and report the same shape, not because either is disinterested.

SOURCES ON THIS PAGE

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S38

CircleCI 2026

~28 million CI workflow runs, 22,000+ organizations, 149 countries

COMMERCIAL INTEREST



INTERFERENCE

The gain flows in per developer. Less of it arrives on main.

SOURCES ON THIS PAGE

S01

DORA / Google Cloud 2025

Nearly 5,000 technology professionals plus qualitative research

COMMERCIAL INTEREST

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S38

CircleCI 2026

~28 million CI workflow runs, 22,000+ organizations, 149 countries

COMMERCIAL INTEREST

TWO READINGS · DOES THE INDIVIDUAL GAIN REACH ORGANISATIONAL DELIVERY?

READING A

It does, and the data now says so.

DORA's 2025 study of nearly 5,000 technology professionals reports that AI adoption improves software delivery throughput, which it describes as a shift from the previous year, while also increasing delivery instability [S01].

The 2026 platform telemetry calls its own figures real delivery acceleration and reports more features shipped and more initiatives completed [S37].

Cannot support

Neither establishes that the change is caused by AI rather than accompanying it. DORA's organisational measures are perceptual, collected by survey from people who know they are being asked about AI. Both are the kind of aggregate that can rise while the median team goes nowhere.

READING B

It does not, and the telemetry says so.

The prior year's cut of the same platform dataset, covering 10,000 developers and 1,255 teams, found organisational delivery metrics flat. CI telemetry finds the median team's main-branch throughput down 7% and main-branch success at a five-year low, against a 59% rise in activity [S38].

Cannot support

Neither rules out a lag. Two years is short for an organisational effect, and a system absorbing a step change in volume may get worse before it gets better. Nor does either isolate AI from every other change these organisations made.

SYNTHESIS

I read these as measuring different things and both being right. Every figure in column A is normalised per developer or gathered by asking people. Every figure in column B counts what reached production and how reliably it got there. The gain is real, it is larger than the sceptical reading allowed, and it is not reaching main. What I cannot show you is a study that measures both altitudes on the same population with the same tool generation, and until someone runs one, this remains my reading of other people's data rather than a finding.

**DISTORT**

The gain flows in per developer. Less of it arrives on main.

What this rules out

It rules out the two most comfortable positions. It is not true that AI has no effect on engineering work; the individual measurements are too large and too consistent for that. And it is not true that the effect flows cleanly through to delivery; churn, incidents, review latency and main-branch reliability all move the wrong way, in datasets whose authors have a commercial interest in the opposite conclusion.

What remains is a question about the intervening system, what sits between a change being written and that change being safely in production, and whether it has capacity to absorb a step change in the volume arriving at its door.

|||||

MEASUREMENT RULE

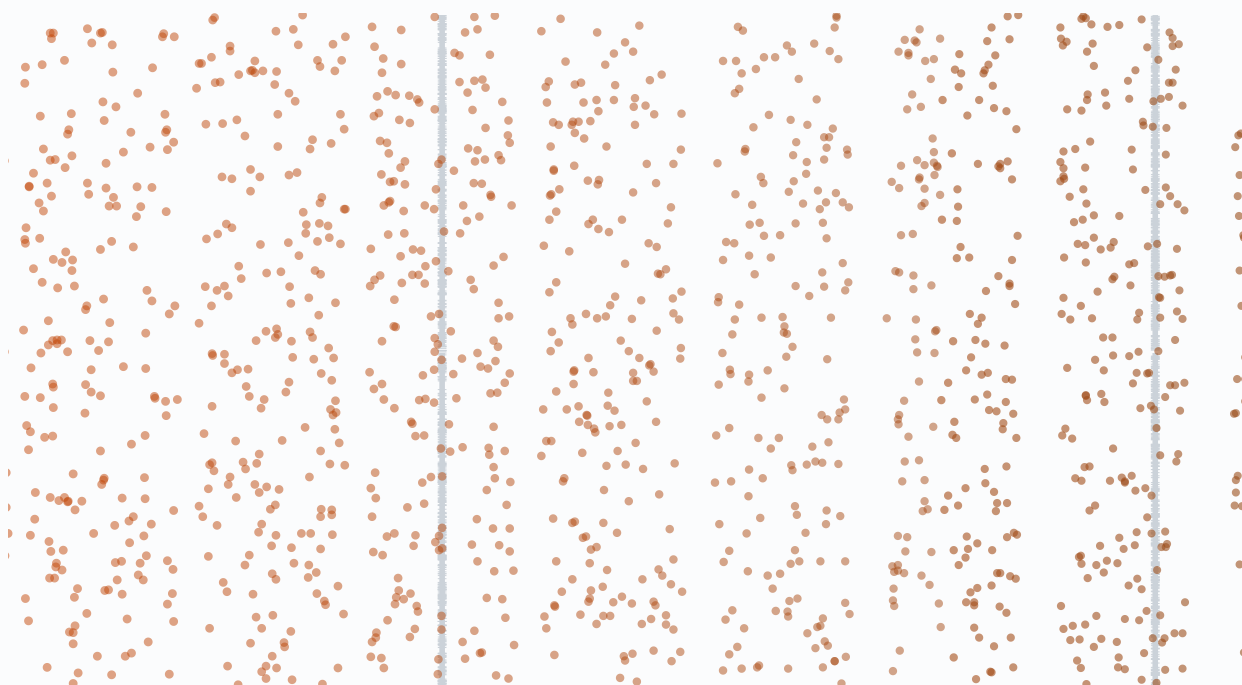
Put one organisational metric beside every individual one, and never accept a per-developer average as evidence about an organisation. An instrument that can only see acceptance rate and suggestions merged cannot detect the problem this chapter describes, and will report success indefinitely while main-branch throughput falls.

|||||



STEPS

The gain flows in per developer. Less of it arrives on main.



03

There is no AI productivity number.

GAP DIMENSION

Perception vs measurement

There is no AI productivity number, and the search for one has done real damage to the quality of the conversation.

The published measured effects run from 19% slower to about 55% faster. A vendor task with 95 developers on a bounded JavaScript problem measured the Copilot group completing the task 55% faster [S06]. A randomised trial with 96 engineers on a complex enterprise task measured about 21% less task time [S05]. A three-experiment programme with 4,867 developers measured 26.08% more completed tasks [S04]. And a randomised trial with sixteen experienced open-source developers, working across 246 tasks in mature repositories they knew well, measured them 19% *slower* [S03].

The uncomfortable result is the careful one

The METR trial is the smallest study in that list and the most methodologically careful. Its participants were experienced developers working in codebases they knew intimately, the condition under which a human's context advantage over a model is largest. They were slowed by 19%. They believed they had been about 20% faster.

That gap between perception and measurement is the single most useful quantity in this report, because it is not a finding about AI. It is a finding about self-report. Every survey figure quoted anywhere in this document, including the adoption and trust figures in Finding 1, is subject to it.

SYNTHESIS

The 39-point figure is my arithmetic, not METR's. They report a 20% perceived speed-up and a 19% increase in elapsed time, measured in opposite directions on different bases, and they publish no combined number. Subtracting one from the other is a rhetorical convenience and should be read as the rough size of a disagreement, not as a measurement.

SOURCES ON THIS PAGE

S03

METR 2025

Experienced open-source developers; mature repositories; average ~5 years...

S04

MIT / Microsoft Research 2025

Developers at Microsoft, Accenture, and an anonymous Fortune 100 company

COMMERCIAL INTEREST

S05

Google Research 2024

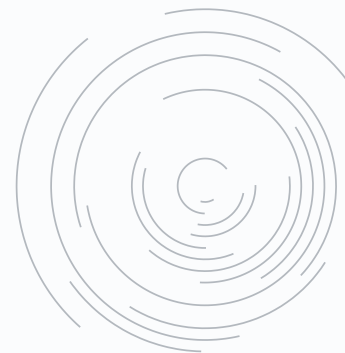
Google software engineers

COMMERCIAL INTEREST

S06

GitHub 2023

Developers familiar with JavaScript

COMMERCIAL INTEREST

ARCS

Studies of unlike size and design. The spread is the finding, not the average.

WHY THIS IS IN THE CHAPTER, NOT A FOOTNOTE

This is the strongest single piece of evidence against the way the 19% figure is usually deployed, including by people who agree with this report. Burying it would be the exact failure the chapter is about. METR also reports that selection bias undermines the follow-up: developers increasingly declined to take part rather than work without AI.

SAMPLE SIZE

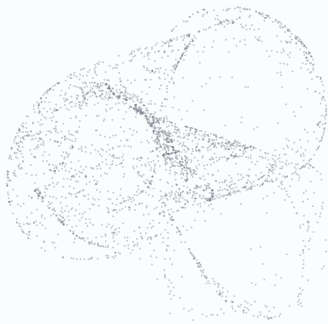
n=16. METR itself warns against broad generalisation from it. It is quoted for the perception gap, which is a within-subject comparison and does not depend on the sample being representative. It is not quoted for the effect size, which does.

SOURCES ON THIS PAGE

S45

METR 2026

Experienced open-source developers; returning participants plus a newly...

**ATTRACTOR**

Studies of unlike size and design. The spread is the finding, not the average.

METR's own follow-up cuts against the number I just quoted

The obligation this chapter imposes on everyone else applies here. In February 2026, six months before this report's evidence cut-off, METR published a re-run: 57 developers, 143 repositories, more than 800 tasks. Returning participants showed a speed-up of -18%, with a confidence interval from -38% to +9%. Newly recruited developers showed -4%, with a confidence interval from -15% to +9%. **Both intervals cross zero.** METR's own reading is that *"developers are more sped up from AI tools now, in early 2026, compared to our estimates from early 2025"* [S45].

Taken together, the two METR studies say something more useful than either alone: the measured effect is moving, it is smaller than it was, and it is no longer distinguishable from zero for developers new to the study. The perception gap survives. The specific number 19% does not, and anyone still quoting it in 2027 without the follow-up is doing what this chapter warns against.

Why the numbers differ

The variance is not noise to be averaged away. It is structure, and it has at least four legible drivers:

- **Task structure.** Bounded, well-specified, low-context work benefits most. The 55% result came from one such task.
- **Codebase familiarity.** The clearest negative result came from developers working where their own context advantage was greatest.
- **Tool generation.** The 26.08% programme began in a 2022-era assistant paradigm. A 2023 autocomplete, a 2025 agentic IDE and a 2026 multi-agent system are not the same intervention, and treating them as one is the most common error in this literature.
- **Outcome metric.** Task time, completed tasks, merged pull requests and validated production outcomes are four different things, and they diverge precisely where this report says they diverge.

SYNTHESIS

Any single percentage that does not name its task type, codebase familiarity, developer experience and tool generation is not a finding. It is a marketing number. This applies to favourable figures as much as to unfavourable ones, and it applies to the figures in this report.

What follows for a leader

The practical consequence is unglamorous: the published numbers cannot tell you what AI will do in your organisation, and no amount of reading will change that. They can tell you what to measure and what confounds it.



TRUCHET

Studies of unlike size and design. The spread is the finding, not the average.

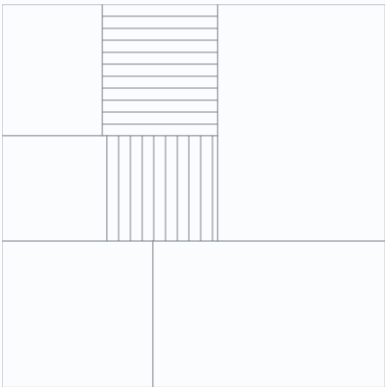
A modest internal comparison, using two matched teams and one repetitive change class, measuring validated lead time rather than coding time, will produce a number that is smaller, less quotable, and enormously more useful than anything in the preceding paragraphs. It will also survive being questioned by your own engineers, which no external figure will.

|||||

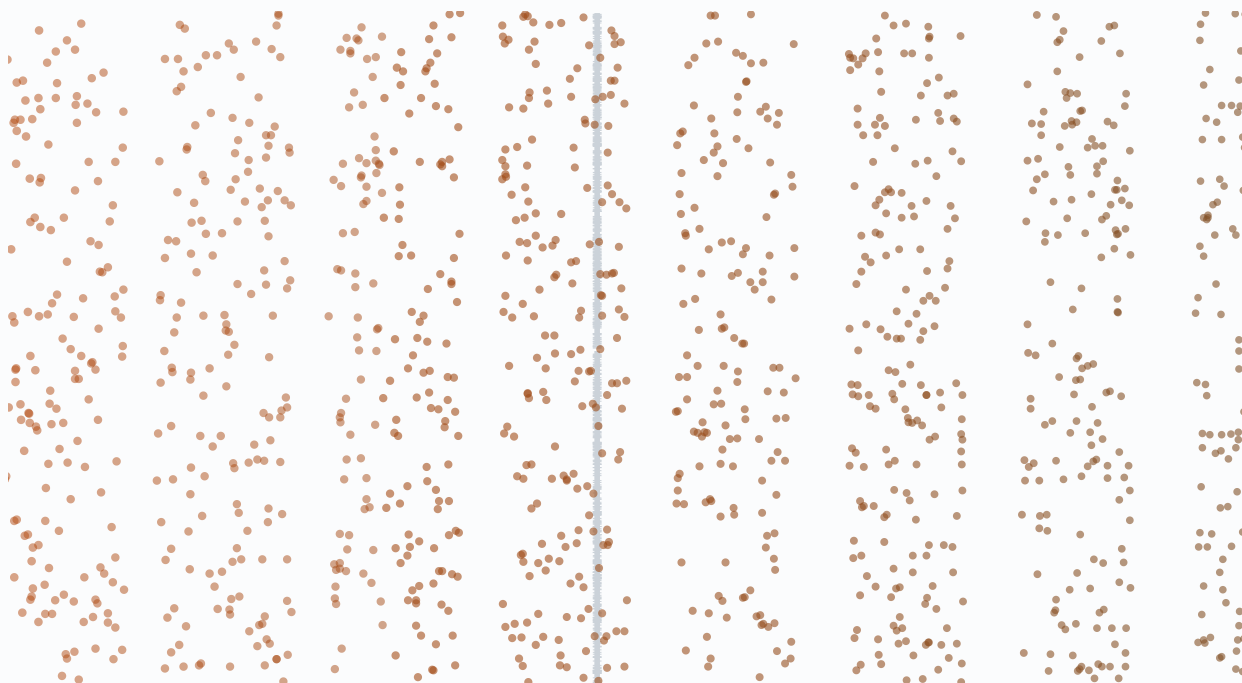
RESEARCH RULE FOR LEADERS

Do not adopt a universal AI productivity percentage, including the ones in this report. Run controlled, task-segmented measurement in your own environment, and publish the design alongside the result.

|||||



SUBDIVIDE
Studies of unlike size and design. The spread is the finding, not the average.



04

Review is where the gain goes.

If the individual gain is real and the organisational gain is absent, the difference is going somewhere. The measurements identify where with unusual precision.

GAP DIMENSION
Generation vs review

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST



CHORDS

Every change is routed through review.
When the queue cannot keep up, the path closes.

The queue

Platform telemetry across 22,000 developers found median time in code review up 441.5%, average time in code review up 199.6%, and time to first review up 156.6% [S37]. In the **2025 edition** of the same report, a different and smaller dataset of 10,000 developers across 1,255 teams, a 98% rise in merged pull requests was accompanied by a 91% rise in review time and flat organisational delivery metrics [S37]. That flat-delivery finding belongs to the 2025 cut and not to the 2026 one, which reports the opposite and calls its own figures real delivery acceleration.

Review is a queue. When arrival rate rises and service rate does not, queue time rises non-linearly, which is what a 441% increase against a 33.7% throughput increase looks like. There is nothing mysterious in the shape of these numbers; it is the arithmetic of an unbuffered queue meeting a step change in arrivals.

The release valve

The same dataset contains the number that should worry an engineering leader most: **31.3% more pull requests were merged with no review at all** [S37].

That is what a queue does when it cannot keep up and the deadline does not move. It is not a policy decision anyone made, it will not appear in any architecture review, and it is measurable today in most organisations that have the data.

The load is real, not imagined

Reviewers are not failing to keep up because they are slow. Reviewing AI-generated code is a materially different task: 38% of developers report it is more effortful than reviewing code written by a colleague, and 53% report encountering code that appears correct and is not [S39]. A human reviewer reads a colleague's diff with a model of that colleague's intent. That model is absent for generated code, and reconstructing it is the work.

The quality signals move in the same direction. Code churn rose 861%, bugs per developer 54%, and monthly incidents 57.9% [S37]. And among developers using agents, roughly 70% reported reduced time on specific tasks and 69% higher productivity, but only 17% agreed that team collaboration had improved [S09]. That last figure is an agreement item on an attitude scale, not a measurement, and the previous chapter's warning about self-report applies to it as much as to the two beside it. It is quoted because the contrast within one instrument is the interesting part: the same people, the same scale, a large majority on personal speed and a small minority on anything shared.

SYNTHESIS

That last figure is the signature of the whole problem. Personal acceleration is arriving; coordination redesign is not. An organisation can have every developer individually faster and no faster overall, and the measurements above describe exactly that organisation.

INCENTIVE

This is vendor telemetry from a company that sells engineering analytics. The finding is corroborated in direction by CI telemetry from a company that does not [S38], but the specific magnitudes are single-source and should be treated as indicative.

SOURCES ON THIS PAGE

S09

Stack Overflow 2025

Global Stack Overflow survey respondents; per-question sample sizes vary

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S39

Sonar 2026

1,100+ developers globally

COMMERCIAL INTEREST



TANGENTS

Every change is routed through review. When the queue cannot keep up, the path closes.

Why more review will not fix it

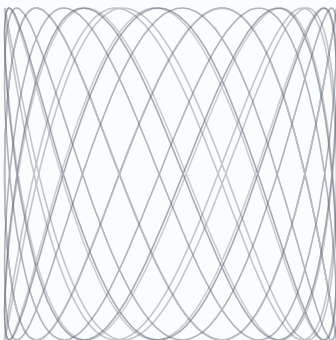
The obvious response is to add review capacity. It does not scale, for a reason that is structural rather than cultural: human reading is linear in output volume, and agent output is not. An organisation that doubles its agent fleet cannot double its senior reviewers, and the 31.3% figure shows what happens at the margin when it tries.

|||||

AUTONOMY RULE

Treat the share of changes merged without human review as a primary operational metric. It is the clearest early warning available, it is cheap to instrument, and it is currently unmeasured almost everywhere.

|||||



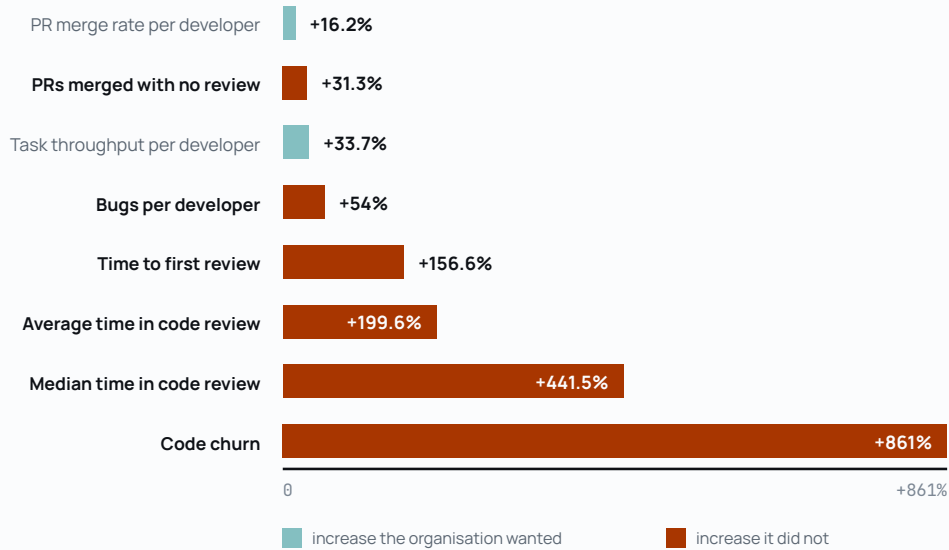
LISSAJOUS

Every change is routed through review. When the queue cannot keep up, the path closes.

FIG 05

What rose, and what the organisation actually wanted to rise

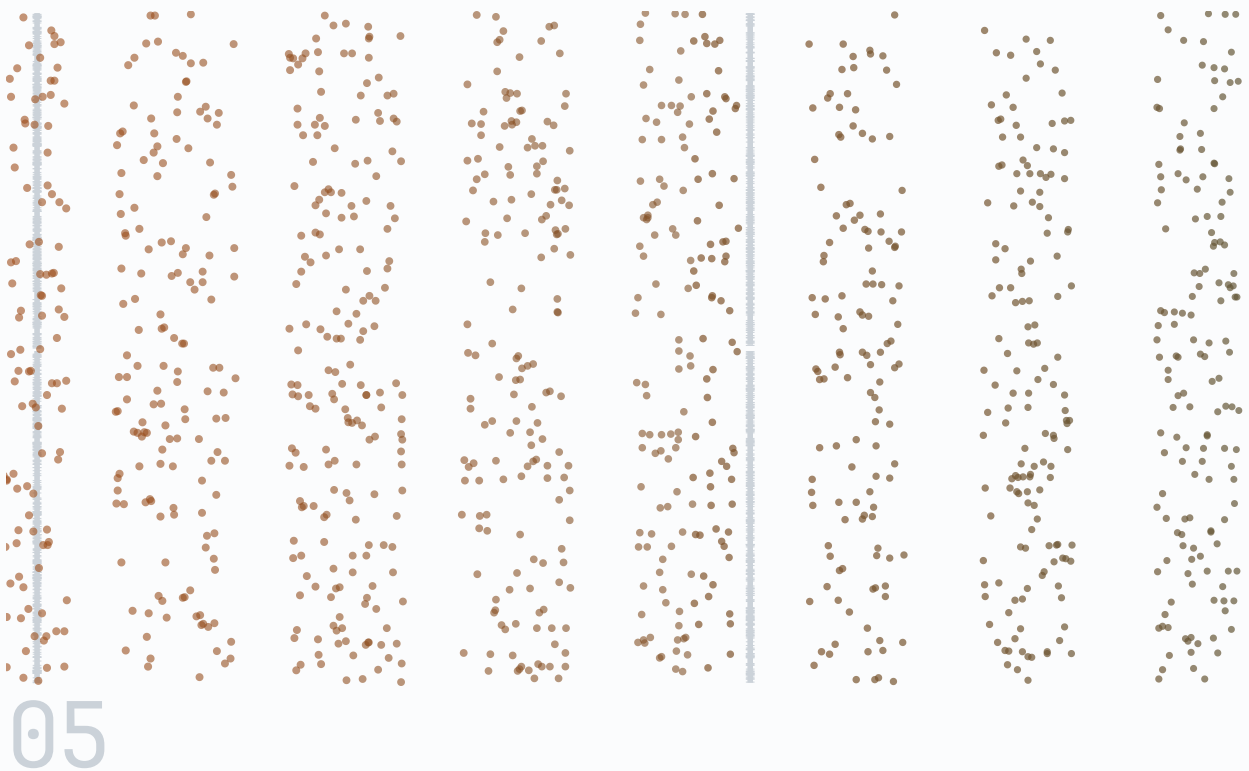
Throughput rose by a third. Code churn rose almost ninefold, and time in review rose more than fourfold. The costs outgrew the gain by an order of magnitude.



Does not show: Causation. These are period-over-period changes within organisations comparing their lowest- and highest-AI-adoption windows, not a controlled experiment.

Unit: percent change, lowest- vs highest-AI-adoption period. Sources: S37 (Faros AI, 2026; 22,000 developers across 4,000+ teams, two years of platform telemetry; tier B).

Commercial interest: S37 Faros AI sells the engineering-analytics platform this telemetry comes from. Every source on this figure has a commercial interest and it should be read as indicative.



Verification has to be manufactured, not performed.

If review capacity cannot scale with generation volume, the response cannot be more review. It has to be a different kind of verification, one whose cost does not rise linearly with the number of changes.

That means moving verification out of the reviewer's head and into the build: proof produced alongside the change, machine-readable, and independent of whatever generated it.

GAP DIMENSION
Reading vs evidence

S41

Kohl, K. & Carro, L. 2026

n/a

S42

Alenezi, M. 2026

Relevant literature and emerging industry perspectives

S43

Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026

One expert engineer, one system (document accessibility remediation)

This direction is not this report's discovery

It is worth being explicit about that, because the surrounding literature has converged on it and a synthesis report that claimed the idea would be misrepresenting the field.

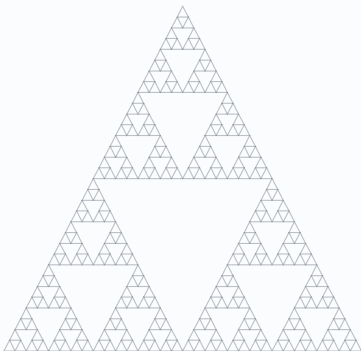
A position paper accepted at ICSE 2026 argues that the discipline must be redefined around human discernment — “intent articulation, architectural control, and verification” — rather than code construction, and names accountability collapse as the central risk [s41]. A structured literature synthesis reaches the same conclusion, identifying verification as the quality bottleneck [s42]. A twelve-week first-person case study of agentic development describes governance controls emerging from failures discovered during the work, a process the authors call governance conversion [s43].

SYNTHESIS

What the literature establishes is the direction. What it does not yet supply is an operating model, and what nobody has supplied is a measurement of how far real organisations have got. This report's contribution is the second and third of those, not the first.

What an evidence package contains

The proposal is that the acceptance artifact of engineering stops being the pull request and becomes a linked set of evidence produced automatically for every change:



SIERPINSKI

Verification decomposed into evidence a machine can produce and a person can check.

- what was **intended**: the outcome, the non-goals, the risk class
- what **changed**, and its **provenance**: which human or agent authored which part
- how it was **verified**: tests, security scan, architecture and policy checks
- which **policies** were applied and which **approvals** were recorded
- how it can be **reversed**, and what happened in **production** afterwards

None of these items is novel in isolation. Most organisations produce several of them already, in disconnected systems, for some changes. The proposal is that they be produced for *every* change, linked to each other, and complete enough to be checked by something other than a person reading a diff.

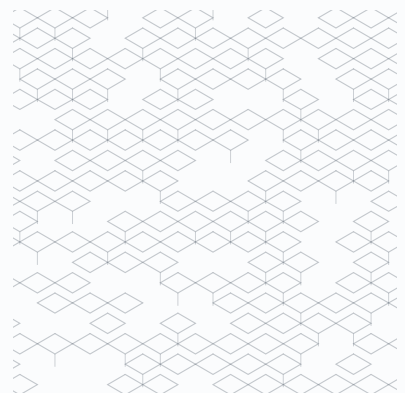
The measurable version

The reason to define it this way is that it yields a number. For a given repository, what share of changes arrive with the full evidence set attached, automatically and linked?

That is evidence completeness, it is cheap to compute once the items are defined, and it converts verification from a matter of individual conscientiousness into something with a trend line. An organisation whose agent fleet is growing and whose evidence completeness is flat has a problem it can see coming.

STATUS

Proposed. Elements of this exist in agentic secure-development guidance [§22] and in specification-as-control-artifact tooling [§30], and the risk taxonomy it must cover is published [§24]. The integration is this report's synthesis and has not been validated at scale.



ISOMETRIC

Verification decomposed into evidence a machine can produce and a person can check.

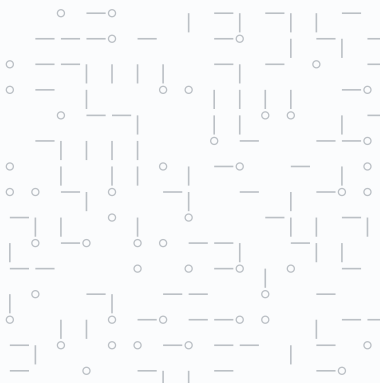
IIII

EVALUATION RULE

Never let the same weak signal serve simultaneously as the generation objective, the release gate and the business KPI. Where the generator and the evaluator share a blind spot, the evidence is decorative.

IIII

The honest position on all of this is that it is a proposal. The direction has support; the specific operating model has not been tested at scale by anyone, including its author, and it should be read as something to falsify rather than something to adopt.

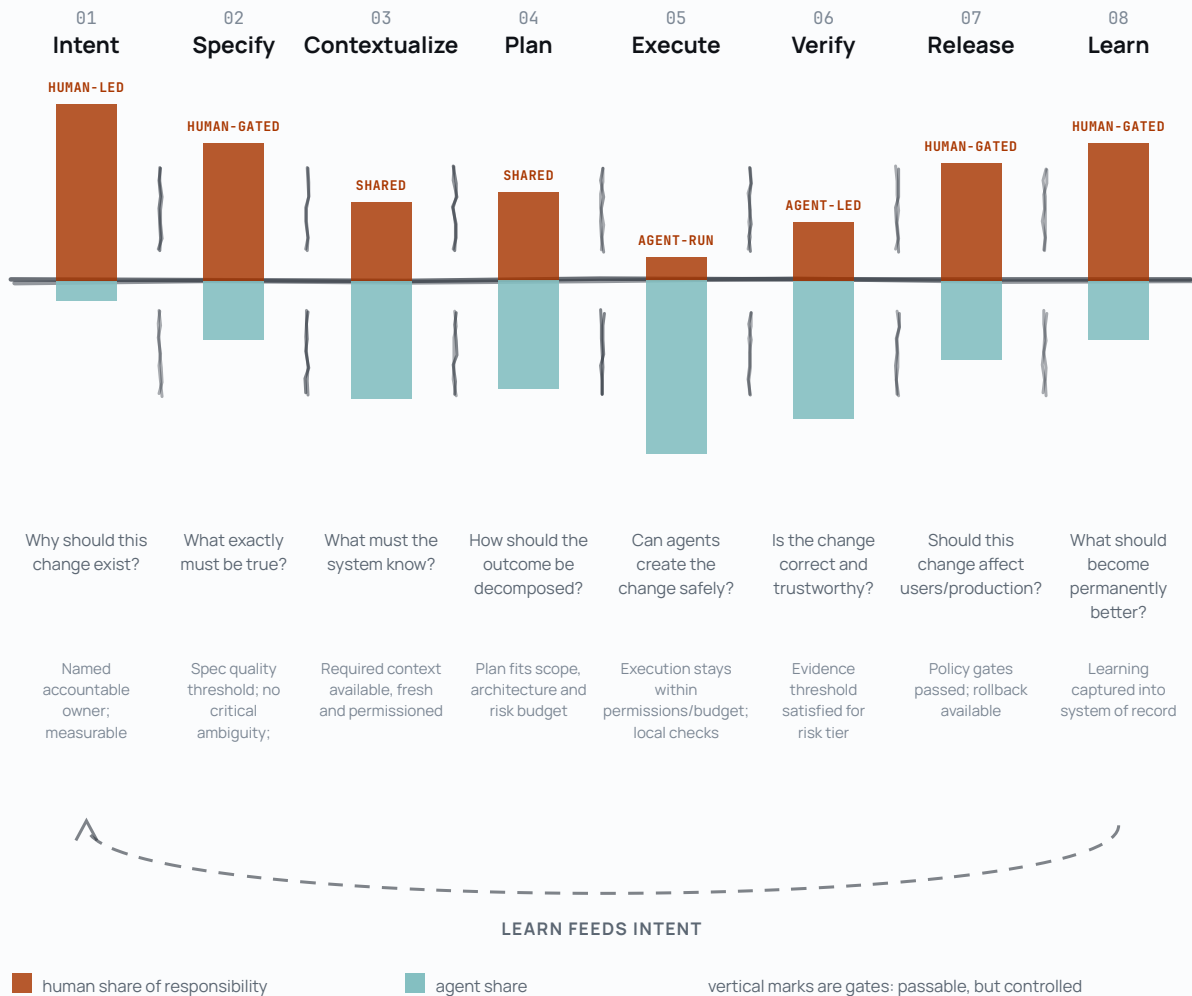
**GRID**

Verification decomposed into evidence a machine can produce and a person can check.

FIG 06

Intent to evidence: eight stages, seven gates, one return

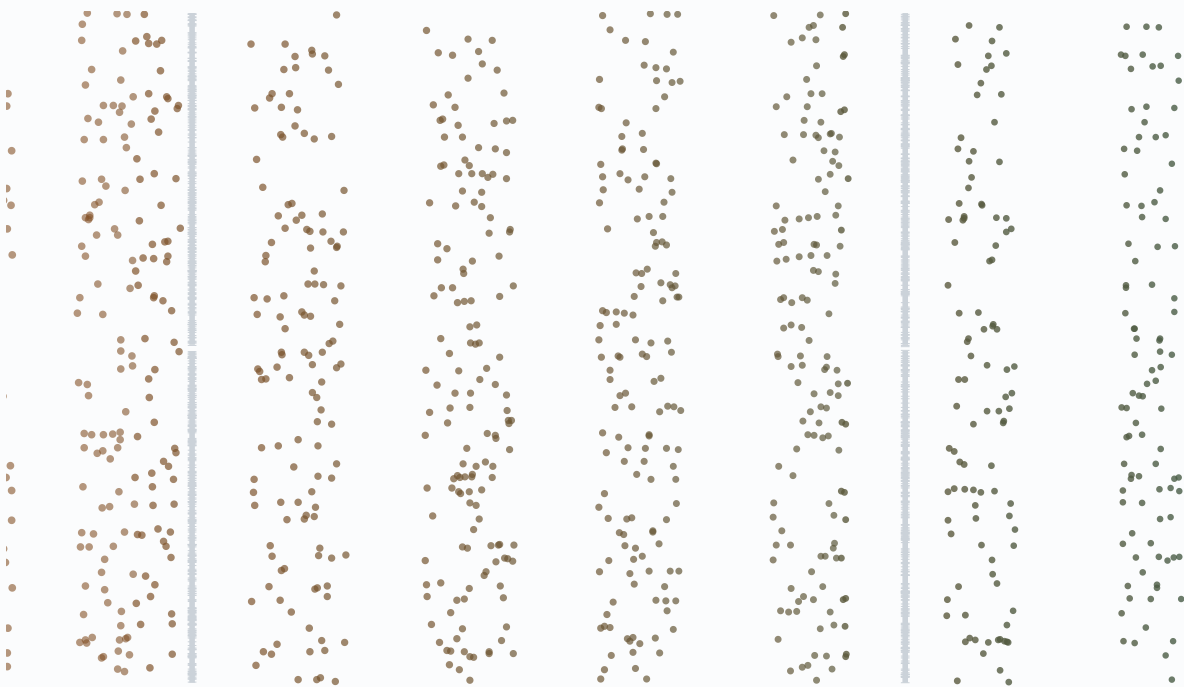
Work does not end when code merges. It ends when enough evidence exists that the intent was met, and what was learned has been fed back.



Does not show: An observed process, or a measured division of labour. This is the author's proposed operating model. The human and agent shares are drawn as bands rather than percentages because they are judgment, and a two-digit number would read as a finding.

Unit: proposed share of responsibility at each stage, banded. Sources: S22 (NIST, 2024; Organizations developing AI/model-enabled systems; tier A) · S24 (OWASP, 2026; Agentic AI applications; tier A) · S30 (GitHub Spec Kit, 2026; AI coding-agent workflows using GitHub Spec Kit; tier C) · S43 (Davis, J.C., Amusuo, P.C., Singla, T., Bakar, B. & Davis, K.A., 2026; One expert engineer, one system (document accessibility remediation); tier C).

Commercial interest: S30 GitHub publishing on its own specification tooling.



06

Authority belongs to evidence, not to capability.

The last step follows from the previous one. If evidence can be manufactured and measured, then evidence, not capability, is what should govern how much an agent is allowed to do.

GAP DIMENSION
Capability vs authority

A benchmark score is not an authorisation

Model capability is measured on task suites: resolve this issue, pass these tests, complete this bounded piece of work. Those measurements are genuinely informative about capability and say nothing about authority, because the two are answers to different questions.

SOURCES ON THIS PAGE

S10

METR 2025

Self-contained software engineering,
ML/research and cybersecurity tasks

S11

METR 2026

Frontier public language-model agents

The measured trend is real. The length of task a frontier agent can complete autonomously at 50% reliability has been doubling roughly every seven months, measured over six years [S10]. That is the strongest available reason to plan for more delegation rather than less. But the same group's live tracker flags its own measurements above about sixteen hours as unreliable with the current task suite [S11], and its tasks are cleaner and lower-context than the work real engineers do.

FORECAST

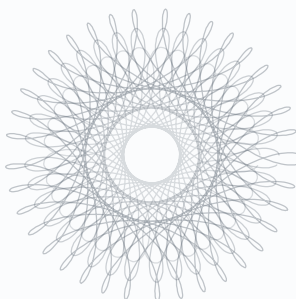
If verification cost stays roughly proportional to change volume, review capacity becomes the binding constraint on delivery for most enterprises before 2029. *Wrong if: independent automated verification improves faster than agent output volume, or if organisations accept lower assurance for low-risk change classes at scale.*

Authority is a two-dimensional decision

The governing artifact is not a level. It is a pair: how much an agent may do, crossed with what must be proven first.

A dependency bump whose tests pass, which is trivially reversible and has a blast radius of one service, can carry high autonomy on modest evidence. A schema migration against customer data can carry strong evidence and still warrant low autonomy, because reversibility and blast radius are properties of the change rather than of the proof. Capability enters neither calculation.

Five conditions bound the decision, and all five are properties of the organisation rather than of the model:



SPIRO

Authority routed by what has been
proven, gate by gate.

- **intent specificity:** can the outcome be stated precisely enough to be checked?
- **context quality:** can the agent retrieve authoritative, current information?
- **verification strength:** is the check independent of the generator?
- **reversibility:** can the change be undone cheaply?
- **blast radius:** how much is exposed if it is wrong?

|||||

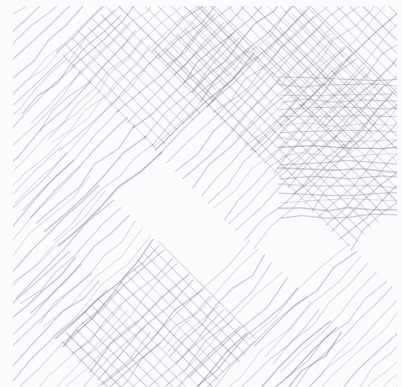
AUTONOMY RULE

Autonomy must never mature faster than the evidence, security, observability and recovery that surround it. Where it does, the organisation is running unproven agents and calling the result velocity.

|||||

STATUS

Proposed. The risk taxonomy it must cover is published and peer-reviewed [§24]; the containment patterns are documented by vendors running agents in production [§13]. The pairing of autonomy tiers to evidence thresholds is this report's synthesis.

**HATCH**

Authority routed by what has been proven, gate by gate.

SOURCES ON THIS PAGE

S25

World Economic Forum 2025

Global employer survey

S26

U.S. Bureau of Labor Statistics 2026

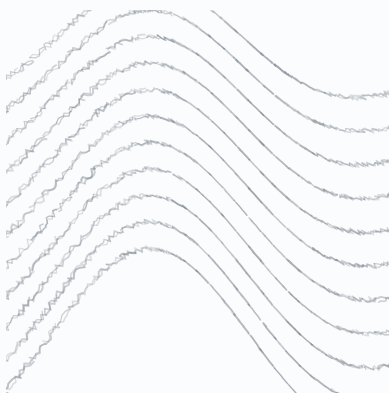
U.S. labor market

The workforce question, briefly

The obvious objection to everything above is that it is moot if the profession is disappearing. The labour evidence does not support that. The US Bureau of Labor Statistics projects software developer employment rising from 1.694 million in 2024 to 1.961 million by 2034, an increase of 15.8%, which ranks the occupation 24th of 832 by projected percent change [S26]. Surveying more than 1,000 employers across 55 economies, the World Economic Forum ranks software and applications developers **fourth of fifteen** fastest-growing jobs by projected percentage growth to 2030, behind big data specialists, fintech engineers and machine learning specialists. The same survey expects 39% of workers' existing skill sets to be transformed or become outdated over the same period [S25].

SYNTHESIS

Read together, those two say redeployment rather than redundancy. More engineers, doing different work. And the work this report describes, specifying intent and manufacturing evidence, is a substantial part of what that different work is.



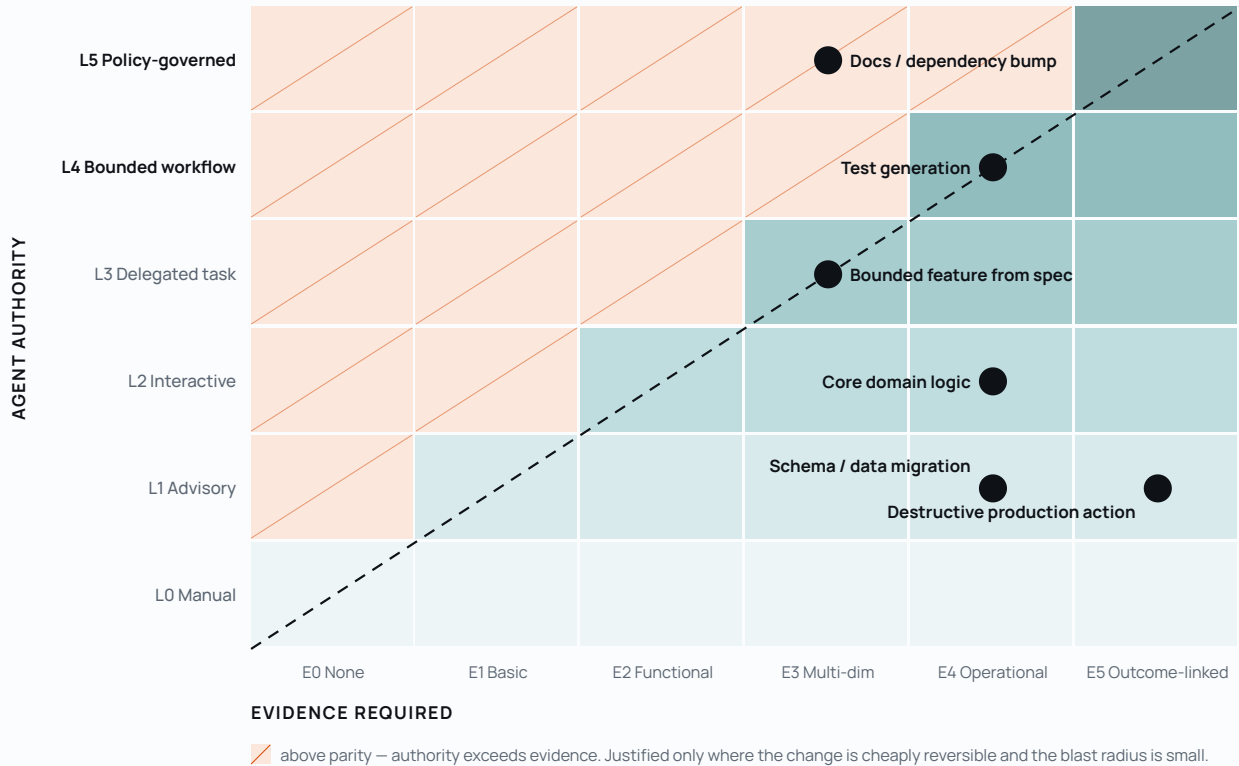
CONTOUR

Authority routed by what has been proven, gate by gate.

FIG 03

Authority is granted by the column, not the row

A model upgrade moves an organisation up the rows. Only better evidence moves it right. Sitting above the parity line is defensible only when the change is cheaply reversible.



Does not show: An empirically derived mapping. The placements are worked examples of the author's proposed policy, not measurements of what organisations do.

Unit: autonomy level against evidence level. Sources: S13 (OpenAI, 2026; Internal coding-agent deployment; tier C) · S24 (OWASP, 2026; Agentic AI applications; tier A) · S11 (METR, 2026; Frontier public language-model agents; tier A) · S43 (Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A., 2026; One expert engineer, one system (document accessibility remediation); tier C).

Commercial interest: S13 OpenAI reporting on its own internal use of its own models.

What to do about it

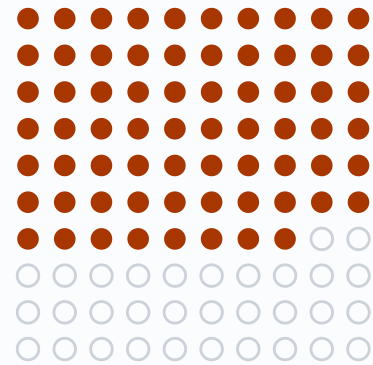
Six actions, one per finding. Each names who owns it, a first move scoped so it can start inside a fortnight, and the evidence that would show it happened. Some of them need instrumentation you may not have; none needs a product you have to buy.

R1 · FROM F1

Measure your own trust gap before you buy another tool.

You cannot manage a trust gap you have not measured, and the published numbers are about other people's organisations. The question is not whether your engineers say they trust AI output. Attitudes are cheap and unreliable. The question is what they actually do with it. Measure what share of accepted AI-generated code was actually read, by team and by change class, from review dwell time and diff-read coverage rather than by asking. Ask as well, but ask second: the gap between the measured number and the estimated one is the finding, and it is the same gap Finding 3 describes. That single number tells you more about your exposure than any adoption metric, and nobody currently publishes it for you.

OWNER	Engineering leadership
FIRST MOVE	Instrument review dwell time and diff-read coverage on one repository for four weeks, then ask the team to estimate the same number and compare.
EVIDENCE CHECK	You have the measured number and the estimated one, broken down, and you can see both move.



R1 answers F1
Agents became the default before trust did.

THE EVIDENCE BEHIND IT

S36
JetBrains 2026
COMMERCIAL INTEREST

S81
DORA / Google Cloud 2025
COMMERCIAL INTEREST

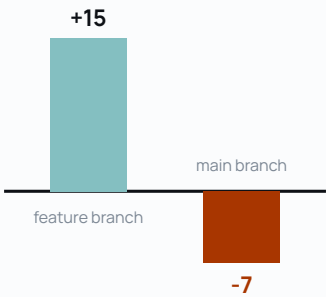
S39
Sonar 2026
COMMERCIAL INTEREST

S89
Stack Overflow 2025

Measure at the organisation, not at the developer.

Developer-level metrics will show you a gain whether or not one reaches production, which is precisely why they are the metrics vendors report. If your only instrumentation is acceptance rate, suggestions merged or self-reported satisfaction, you have bought a measurement that cannot detect the problem this report describes. It will report success while main-branch throughput falls, and it will keep doing so for as long as you fund it. Put one organisational metric next to every individual one, and require them to move together before you scale anything further.

OWNER	CTO
FIRST MOVE	Pair every developer-level AI metric you already collect with an organisational one: change lead time, change failure rate, main-branch throughput.
EVIDENCE CHECK	No AI metric is reported to leadership without an organisational counterpart on the same slide.



R2 answers F2
The gain is real per developer and does not reach main.

THE EVIDENCE BEHIND IT

S48
Murphy-Hill, E., Butler, J. & Savelieva, A. (Microsoft) 2026
COMMERCIAL INTEREST

S84
MIT / Microsoft Research 2025
COMMERCIAL INTEREST

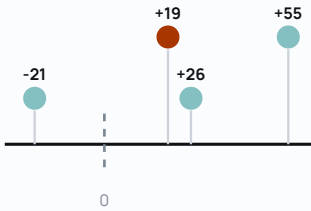
S85
Google Research 2024
COMMERCIAL INTEREST

S83
METR 2025

Run your own comparison. Refuse the industry percentage.

The heterogeneity in the literature is not noise to be averaged away; it is the finding. Effects depend on task structure, code-base familiarity, experience and tool generation, and those differ more between your teams than between the published studies. A modest internal comparison on one repetitive change class will tell you more than any external number, and it is the only evidence that will survive being questioned by your own engineers.

OWNER	CTO with engineering leadership
FIRST MOVE	Pick two matched teams and one repetitive change class. Measure validated lead time, not coding time.
EVIDENCE CHECK	You have an internal result with a stated n and a design you would be willing to publish.



R3 answers F3
There is no AI productivity number.

THE EVIDENCE BEHIND IT

S03
METR 2025

S04
MIT / Microsoft Research 2025
COMMERCIAL INTEREST

S05
Google Research 2024
COMMERCIAL INTEREST

S06
GitHub 2023
COMMERCIAL INTEREST

Treat review capacity as the constraint it now is.

Most engineering organisations still budget review as a courtesy that senior engineers perform between their real work. If generation volume has risen and review capacity has not, review is now your binding constraint, and everything upstream of it is producing inventory rather than value. Measure the queue, publish it, and stop treating unreviewed merges as an acceptable release valve. That number rising is the clearest early warning available to you.

OWNER	Engineering leadership with platform
FIRST MOVE	Instrument three numbers weekly: time to first review, time in review, and the share of changes merged without human review.
EVIDENCE CHECK	All three are on a dashboard someone is accountable for, and the unreviewed share is trending down.

throughput

+33.7%

time in review

+441.5%

13.1x the gain

R4 answers F4

Review is where the gain goes.

THE EVIDENCE BEHIND IT

S37

Faros AI 2026

COMMERCIAL INTEREST

S39

Sonar 2026

COMMERCIAL INTEREST

S09

Stack Overflow 2025

Make evidence a build artifact, not a reading task.

Pick one repository and make a fixed set of evidence items produced automatically and linked to every change: acceptance criteria, test results, security scan, policy check, provenance of who or what authored the change, approval record, rollback record. Then measure completeness. The point is not the checklist. It is that evidence completeness becomes a number that can be trended, which is the only way verification stops being a matter of individual conscientiousness. Do not attempt all eight in a fortnight. Three that are genuinely automatic beat eight that are aspirational, and the completeness number is only worth trending once nobody is filling it in by hand.

OWNER	Platform engineering
FIRST MOVE	Choose one repository. Make three of the eight evidence items automatic and linked for every change, and publish the completeness number.
EVIDENCE CHECK	An evidence-completeness percentage exists for that repo and is trending up.



eight items, for every change

R5 answers F5
Verification has to be manufactured, not performed.

THE EVIDENCE BEHIND IT

S22
NIST 2024

S24
OWASP 2026

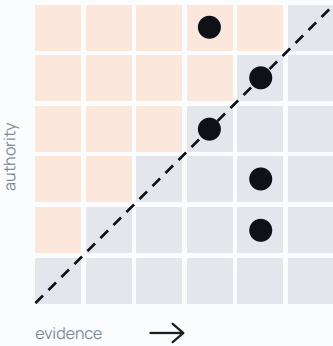
S38
GitHub Spec Kit 2026
COMMERCIAL INTEREST

S43
Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026

Cap autonomy at your weakest control plane.

Model capability is the input you control least and the one most likely to be used to justify raising authority. Name the weakest of your control planes (context, verification, security, observability, recovery) and set the autonomy ceiling there, in writing. The test of whether you have done this properly is simple: a model upgrade should not be able to raise the ceiling on its own. If it can, you have a capability policy, not an autonomy policy.

OWNER	CTO with security and risk
FIRST MOVE	Name your weakest control plane and write the autonomy ceiling it implies, for one change class.
EVIDENCE CHECK	A written ceiling exists that a model upgrade cannot raise by itself.



R6 answers F6
Authority belongs to evidence, not to capability.

THE EVIDENCE BEHIND IT

S13
OpenAI 2026
COMMERCIAL INTEREST

S24
OWASP 2026

S11
METR 2026

S43
Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026

The case against this report

A report that only presents the evidence supporting its argument is advocacy. This chapter sets out the strongest objections I know of to what I have just spent sixty pages claiming. I have not softened them, and I do not answer all of them.

The altitude pattern may be an artifact of method, not of altitude

This is the best counter-argument and it goes at the heart of the report.

Look again at which studies found a gain and which did not. The individual studies are mostly **experiments**: randomised trials, matched cohorts, controlled task assignment. The organisational studies are mostly **observational telemetry**: platform data compared across periods of higher and lower adoption. Experiments control for confounders. Telemetry does not.

So the pattern I have called an altitude effect could instead be a method effect. Experiments find gains because they isolate the intervention. Telemetry finds none because everything else in the organisation is moving at the same time, and always is.

SYNTHESIS

I do not think this fully explains the pattern, because the organisational findings are large, consistent across two independent datasets [S37 S38], and point the same way rather than scattering. But I cannot rule it out with the evidence available, and anyone who wants to reject this report's central claim should start here.

SOURCES ON THIS PAGE

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

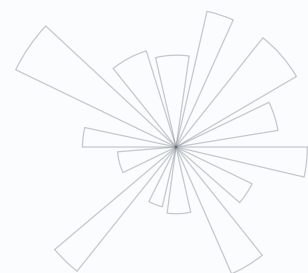
COMMERCIAL INTEREST

S38

CircleCI 2026

~28 million CI workflow runs, 22,000+ organizations, 149 countries

COMMERCIAL INTEREST



WEDGES

The objections, laid out at their strongest rather than their weakest.

S03

METR 2025

Experienced open-source developers; mature repositories; average ~5 years...

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S38

CircleCI 2026

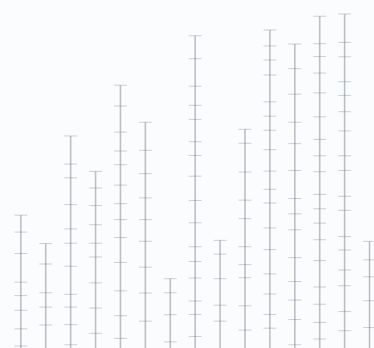
~28 million CI workflow runs, 22,000+ organizations, 149 countries

COMMERCIAL INTEREST

S39

Sonar 2026

1,100+ developers globally

COMMERCIAL INTEREST

COLUMNS

The objections, laid out at their strongest rather than their weakest.

The two most striking numbers come from companies selling the remedy

The 441.5% rise in time in code review, and the finding that 96% of developers do not fully trust AI-generated code while 48% always verify it, are the report's most quotable figures. Both come from vendors whose products address exactly those problems [S37 S39].

I have flagged this on every figure that uses them. Three of the six findings still carry a headline number from a single commercially interested source, which is a real weakness and is marked on those cards. That mitigates the problem; it does not remove it. If those two datasets were withdrawn tomorrow, Findings 1 and 4 would be materially weaker.

The telemetry measures the organisations that bought telemetry

Both large observational datasets describe customers of engineering-analytics or CI platforms [S37 S38]. Organisations that buy those tools are not a random sample of software organisations. They are more instrumented, more likely to be large, and more likely to have delivery problems worth measuring. The direction of that bias is genuinely unclear, which is worse than knowing it.

The most-cited negative result is very small

METR's finding that experienced developers were 19% slower while believing they were 20% faster is the single most-used result in this report. It is also the smallest study cited: sixteen developers [S03]. METR itself warns against generalising from it.

I use it for the perception gap rather than the effect size, which is a within-subject comparison and does not depend on the sample being representative. That is a real defence, and it is also exactly the kind of careful framing that people skip when they quote a number.

The studies are not measuring the same intervention

The +26% completed-tasks result came from experiments that began in a 2022-era assistant paradigm [s84]. The organisational telemetry describes 2024 to 2026 agentic tooling. A 2022 autocomplete and a 2026 agent are not the same thing, and setting their results beside each other invites a comparison the evidence does not support.

Delivery may simply lag

Organisational metrics might be slower than individual ones for ordinary reasons: hiring, process change, and contract cycles all move in years. Two years of telemetry may be too short a window to see a gain that is real but delayed.

FORECAST

If this is the explanation, organisational delivery metrics begin improving in the 2027 to 2028 data without any change in verification practice. *That is the cleanest single test of this report against its critics, and the next edition will report it either way.*

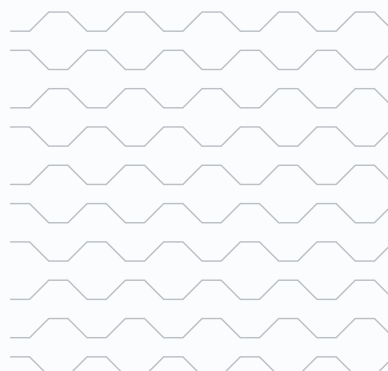
SOURCES ON THIS PAGE

s84

MIT / Microsoft Research 2025

Developers at Microsoft, Accenture, and an anonymous Fortune 100 company

COMMERCIAL INTEREST



MEANDER

The objections, laid out at their strongest rather than their weakest.

And the method itself

This is a synthesis of 25 sources, not a systematic review. No protocol was pre-registered, no double-screening was done, and no assessment was made of what the search missed. Selection bias cannot be excluded, and the person who selected the sources is the person arguing the case.

|||||

WHAT WOULD CHANGE MY MIND

A well-designed organisational study finding a clear delivery gain. Or independent, non-vendor telemetry that fails to reproduce the review-latency finding. Either would move me, and I would rather find out in 2027 than be right by not looking.

|||||



SHELLS

The objections, laid out at their strongest rather than their weakest.

What I think happens next

Everything to this point has been either a measurement or a reading of one. This chapter is neither. It is what I expect to happen, given the evidence in this report and the years I have spent watching engineering organisations absorb new tooling.

I have written each prediction so it can be checked. Every one names the finding it rests on, the window it applies to, and the condition that would prove it wrong. If you disagree with a prediction, the useful thing is not to argue about 2030 but to name which of the conditions you expect to hold instead.

Adoption stops being the interesting number

FORECAST

Agent adoption reaches practical saturation among professional developers during 2027, and the question organisations ask moves from *how many people use agents* to *what agents are permitted to do without a human*. Weekly use was already 90% in mid-2026 [S36]. *Wrong if: weekly adoption stalls below 95% through 2027, or a material share of organisations formally withdraw agent access after having granted it.*

The reason to expect this is not enthusiasm. It is that the remaining non-adopters are concentrated in regulated and legacy environments where the blocker is permission rather than preference, and permission is exactly what the second half of this report is about.

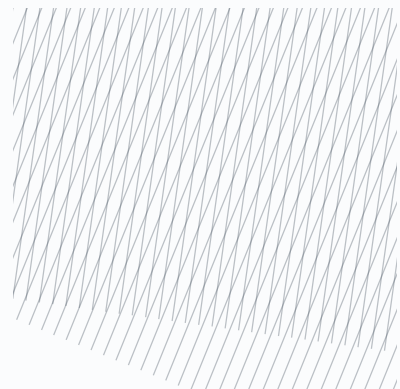
SOURCES ON THIS PAGE

S36

JetBrains 2026

15,000+ professional developers worldwide, ~90% developer/engineer titles

COMMERCIAL INTEREST



MOIRE

Predictions, each with the condition that would prove it wrong.

S37

Faros AI 2026

22,000 developers across 4,000+ teams, two years of platform telemetry

COMMERCIAL INTEREST

S38

CircleCI 2026

~28 million CI workflow runs, 22,000+ organizations, 149 countries

COMMERCIAL INTEREST



TRELLIS

Predictions, each with the condition that would prove it wrong.

Review becomes rationed rather than universal

FORECAST

By 2028 most large engineering organisations stop attempting to review every change and route review by risk class instead, because the arithmetic leaves them no alternative. Time in review had already risen 441.5% while throughput rose 33.7%, and 31.3% more changes were merged with no review at all [S37]. *Wrong if: median time in review falls while generation volume continues to rise, or the share of unreviewed merges declines without a formal routing policy.*

The uncomfortable part of this prediction is that it is already happening, but by accident rather than design. Unreviewed merges are risk-routing without a policy, chosen by whoever was under deadline pressure that afternoon. The prediction is that organisations formalise what is already occurring, and the ones that do it deliberately will be measurably better off than the ones that let the queue decide.

The delivery gap widens before it closes

FORECAST

Through 2027 the spread between organisations that convert AI activity into delivered software and those that do not gets wider, not narrower. CI activity rose 59% while the median team's main-branch throughput fell 7% and success rates reached a five-year low [S38]. *Wrong if: median main-branch throughput recovers to its pre-2025 trend while activity stays elevated.*

I expect this because the fix is organisational rather than technical, and organisational fixes are slow, unevenly distributed, and rarely attempted before the cost becomes undeniable.

Evidence completeness becomes a tracked number

FORECAST

Some measure of evidence completeness (what share of changes arrive with tests, scans, provenance and an approval record attached and linked) enters mainstream engineering dashboards by 2029, in the way that deployment frequency did after 2018. *Wrong if: no widely used engineering-analytics product ships such a measure by 2029, or the measure ships and is not adopted.*

This is the prediction I hold most loosely. It is a claim about what the industry decides to count, and industries have counted the wrong things for long stretches before.

The job grows and changes at the same time

FORECAST

Software engineering headcount continues to grow through 2030 while the median engineer spends materially less time writing code and more time specifying, reviewing and accounting for it. US projections put developer employment up 15.8% to 2034 [S26] and the WEF ranks developers fourth of fifteen fastest-growing jobs while expecting 39% of workers' existing skill sets to be transformed or become outdated over 2025-2030 [S25]. *Wrong if: developer employment growth turns negative in two consecutive annual projections, or time-use studies show code authorship holding steady as a share of the working week.*

SOURCES ON THIS PAGE

S25

World Economic Forum 2025

Global employer survey

S26

U.S. Bureau of Labor Statistics 2026

U.S. labor market



DRIFT

Predictions, each with the condition that would prove it wrong.

S10

METR 2025Self-contained software engineering,
ML/research and cybersecurity tasks

S11

METR 2026

Frontier public language-model agents

What I do not predict

I do not predict a date at which agents write most production software unsupervised. The measured task horizon has been doubling roughly every seven months [S10], which is a real trend and a bad basis for a date: the same group flags its own measurements above sixteen hours as unreliable [S11], and benchmark tasks are cleaner and lower-context than the work that actually fills an engineer's week.

I also do not predict which vendors win. Nothing in this report's evidence base is a claim about market share, and a synthesis report that drifted into vendor prediction would be borrowing authority it has not earned.

|||||

HOW TO USE THIS CHAPTER

Treat each prediction as a bet with stated terms rather than as a finding. If you are planning against one of them, the falsifier is the thing to instrument, because it tells you early whether the bet is going wrong.

|||||

**HEXGRID**Predictions, each with the condition
that would prove it wrong.

Method and evidence base

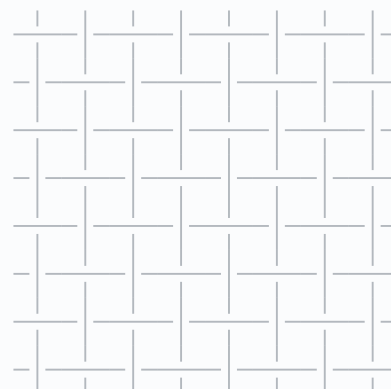
This report contains no primary research data.

It is a structured synthesis of publicly available sources, combined with frameworks proposed by the author. There was no survey. There were no interviews. There was no fieldwork. There is no sample, no response rate and no margin of error, because there is no dataset. Every percentage in this report belongs to the source it is attributed to, not to this report.

Source selection

Sources were identified through structured search of research repositories, standards bodies, official statistical agencies, and the published engineering literature of major AI and developer-tool vendors. Selection deliberately sought disconfirming as well as confirming evidence: the report includes findings that AI tools slowed experienced developers, that trust in AI-generated code is low, and that software employment is projected to grow, alongside findings of substantial productivity gains.

Selection was **not systematic** in the sense used in systematic review methodology. No inclusion protocol was pre-registered, no double-screening was performed, and no assessment was made of the completeness of the literature captured. **Selection bias in the source set cannot be excluded.**



WEAVE

How the evidence base was assembled, tiered and limited.

METHOD AND EVIDENCE BASE

COMPOSITION, STATED PLAINLY

A substantial share of the sources are authored by organisations that sell AI development or verification tooling, and every one of them benefits commercially from the conclusion that verification is hard. They are flagged in the ledger and on every figure that uses them. Three of the six findings carry a headline number from a single commercially interested source, and that is stated on the card rather than buried here.



ORBITS

How the evidence base was assembled, tiered and limited.

Appraisal

Each source is recorded with method, population, quality tier, evidence type, key limitation and URL. Quality tiers are: **A**, primary empirical research, official statistics or standards; **B**, credible surveys, academic preprints and vendor empirical studies; **C**, product documentation, vendor case studies and expert synthesis.

What this method cannot tell you

It cannot establish causation. It cannot tell you what will happen in your organisation. It cannot resolve the contradictions between the cited studies, which measured different things in different populations with different tools in different years. This report reports that heterogeneity rather than averaging it away. And it cannot observe 2030.

Frameworks

The evidence package, the pairing of autonomy tiers to evidence thresholds, and the operating model in Findings 5 and 6 are **proposed by the author**. They synthesise patterns observed across the cited sources but have not been tested, validated or applied at scale in a measured way. They are offered as planning instruments, not as validated models, and the next page lists the conditions under which they would be wrong.

What would falsify this report

A forecast with no falsifier is unaccountable. These are the conditions under which the central claims of this report would be wrong, stated before the fact so that the 2027 edition has something to check rather than something to defend.

The measurement claims

Finding 2, that the individual gain does not reach the organisation, would be wrong if organisational delivery metrics begin rising in step with individual throughput in the same datasets that currently show them flat. The specific check: does a subsequent cut of the same telemetry show main-branch throughput and change failure rate improving where AI adoption is highest?

Finding 4, that review is the binding constraint, would be wrong if review latency falls while generation volume continues to rise. The specific check: median time in code review, and the share of changes merged without human review. If the second falls while the first stabilises, the constraint has moved elsewhere and this report has misidentified it.

The proposals

Finding 5, that verification must be manufactured rather than performed, would be wrong if automated verification quality improves fast enough that a human reading a diff is no longer the bottleneck, or if organisations accept materially lower assurance for a large class of changes and suffer no measurable consequence. Either would make the evidence package unnecessary rather than merely untested.



CROSS

Density is proof. Where it thins, the claim is exposed.

Finding 6, that authority should track evidence rather than capability, would be wrong if organisations that grant autonomy on capability alone show no worse incident, churn or roll-back rates than those that gate on evidence. That is a comparison nobody has run, and it is the single most valuable experiment available in this field.

FORECAST

If verification cost stays roughly proportional to change volume, review capacity becomes the binding constraint on delivery for most enterprises before 2029. *Wrong if: independent automated verification improves faster than agent output volume, or if organisations accept lower assurance for low-risk change classes at scale.*

The whole report

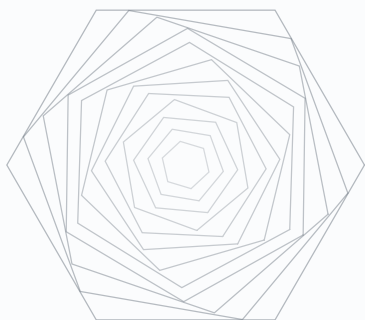
The argument rests on the claim that the studies do not disagree but measure at different altitudes. **That would be wrong if** a well-designed organisational study finds a clear delivery gain, or a well-designed individual study finds none. Either result would mean the pattern this report describes is an artifact of which organisations currently have telemetry, rather than a property of AI-assisted engineering.

|||||

STANDING INVITATION

If you have primary data that contradicts anything here, please send it. It will be published in the errata and incorporated in the next edition, with credit.

|||||


POLYGONS

Density is proof. Where it thins, the claim is exposed.

Evidence ledger

Every source, with its population, method, quality tier, locator and, the column that matters most, its key limitation. Sources with a commercial interest in the subject matter are flagged, and so are the ones whose supporting sentence could not be captured verbatim.

Identifiers are not consecutive. They are assigned once and never reused, so a source dropped during verification leaves its number empty rather than renumbering everything after it. A citation to [S37] means the same source in every edition.

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S01	DORA / Google Cloud 2025 <i>State of AI-assisted Software Development 2025</i> 90% of respondents use AI at work; >80% report perceived productivity improvement; 30% report little/no trust in AI-generated code; 59% observe positive impact on code quality. services.google.com/fh/files/misc/2025_state_of_ai_assisted_software_development.pdf	Nearly 5,000 technology professionals plus qualitative research Global survey + qualitative research	Self-reported outcomes and cross-sectional organizational evidence; causal interpretation should be limited.	A · EVIDENCE VENDOR
S03	METR 2025 <i>Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity</i> 16 developers, 246 tasks; AI allowed condition took 19% longer, while developers expected 24% faster and perceived ~20% faster. metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study	Experienced open-source developers; mature repositories; average ~5 years prior repo experience Randomized controlled trial	Narrow setting and early-2025 frontier tools; METR explicitly warns against broad generalization.	A · EVIDENCE
S04	MIT / Microsoft Research 2025 <i>Effects of Generative AI on High-Skilled Work: Three Field Experiments</i> 4,867 developers; preferred estimate +26.08% completed tasks (SE 10.3%). www.microsoft.com/en-us/research/publication/the-effects-of-generative-ai-on-high-skilled-work-evidence-from-three-field-experiments-with-software-developers	Developers at Microsoft, Accenture, and an anonymous Fortune 100 company Three randomized field experiments	Experiments began in 2022-era assistant paradigm; effect sizes may not transfer directly to newer agents or all work types. The +13.55% commits and +38.38% compilations figures could not be located at the recorded URL and have been removed...	A · EVIDENCE VENDOR
S05	Google Research 2024 <i>How much does AI impact development speed? Enterprise RCT</i> 96 full-time Google software engineers; best estimate ~21% reduction in task time; confidence interval was large. doi.org/10.48550/arXiv.2410.12944	Google software engineers Randomized controlled trial	Single enterprise task and internal tooling; authors caution against extrapolating to ecosystem-wide productivity.	A · EVIDENCE VENDOR

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S06	GitHub 2023 <i>Responsible AI pair programming with GitHub Copilot</i> 95 developers; 45 with Copilot; Copilot group completed a JavaScript task ~55% faster. github.blog/ai-and-ml/github-copilot/responsible-ai-pair-programming-with-github-copilot	Developers familiar with JavaScript Controlled experiment (vendor research)	Single constrained task; vendor-run study; not representative of end-to-end software delivery. Published 2023 on an experiment run with 2022-era Copilot; by this report's own tool-generation argument, the oldest tool generation in the...	 B · EVIDENCE  VENDOR
S07	GitHub 2025 <i>Does GitHub Copilot improve code quality?</i> 202 valid submissions; Copilot users 53.2% more likely to pass all 10 unit tests; readability +3.62%, reliability +2.94%, maintainability +2.47%, conciseness +4.16%. github.blog/news-insights/research/does-github-copilot-improve-code-quality-heres-what-the-data-says	Study participants producing code reviewed against tests and quality dimensions Controlled study (vendor research)	Vendor-run study and constrained task; quality metrics do not prove production reliability.	 B · EVIDENCE  VENDOR
S09	Stack Overflow 2025 <i>2025 Developer Survey - AI</i> 46% distrust AI accuracy vs 33% trust; 52% do not use agents or use simpler AI; among agent users ~70% report task-time reduction, 69% productivity increase, 17% collaboration improvement; 87% accuracy... survey.stackoverflow.co/2025/ai	Global Stack Overflow survey respondents; per-question sample sizes vary Developer survey	Self-selection and self-reported effects; agent questions have smaller respondent subsets.	 B · EVIDENCE 
S10	METR 2025 <i>Measuring AI Ability to Complete Long Tasks</i> 50%-success task time horizon doubling approximately every 7 months over the preceding 6 years. metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks	Self-contained software engineering, ML/research and cybersecurity tasks Benchmark + human baseline timing + logistic modeling	Tasks are cleaner and lower-context than much real-world engineering; extrapolation to 2030 is uncertain.	 A · SIGNAL 
S11	METR 2026 <i>Task-Completion Time Horizons of Frontier AI Models</i> Task-horizon measurements above ~16 hours are flagged as unreliable with the current task suite. metr.org/time-horizons	Frontier public language-model agents Time Horizon 1.1 benchmark methodology	LIVING PAGE: figures change without notice. Must be cited as "figures as of 20 August 2026" and re-checked quarterly. The seven-month doubling claim belongs to S10, not to this page.	 A · SIGNAL 

EVIDENCE LEDGER

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S13	OpenAI 2026 <i>Running Codex safely at OpenAI</i> Operational design pattern: routine low-risk actions can proceed; higher-risk actions are made explicit and bounded. openai.com/index/running-codex-safely	Internal coding-agent deployment Operational practice / case study	Vendor-specific implementation; principles require adaptation to enterprise context.	C · SIGNAL VENDOR
S16	Martin Fowler / Thoughtworks 2025 <i>How far can we push AI autonomy in code generation?</i> Qualitative experimental findings across agentic workflows for a Spring Boot application. martinfowler.com/articles/pushing-ai-autonomy.html	Thoughtworks experimental workflow Engineering experiment / practitioner research	Not a controlled population study; results depend on models, prompts and stacks available at the time.	B · EVIDENCE
S22	NIST 2024 <i>SP 800-218A: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models</i> Final NIST SSDF Community Profile for generative AI and dual-use foundation models. csrc.nist.gov/pubs/sp/800/218/a/final	Organizations developing AI/model-enabled systems Standards guidance	Guidance is intentionally high-level and requires organization-specific controls.	A · SYNTHESIS
S24	OWASP 2026 <i>OWASP Top 10 for Agentic Applications 2026</i> Globally peer-reviewed framework developed with 100+ experts/researchers/practitioners. genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026	Agentic AI applications Community security framework	Threat framework, not incidence-rate measurement.	A · SYNTHESIS
S25	World Economic Forum 2025 <i>Future of Jobs Report 2025</i> Over 1,000 employers representing 14M+ workers, 22 industry clusters, 55 economies. 39% of workers' existing skill sets expected to be transformed or become outdated over 2025-2030 (down from 44% in 2023, 57%... www.weforum.org/publications/the-future-of-jobs-report-2025	Global employer survey Global employer survey + labor-market modeling	Employer self-report of intentions, not measurement of outcomes. Rankings are percentage growth as a share of current employment in each role, not absolute headcount, so a small occupation can outrank a large one. Survey base is over 1,000...	B · EVIDENCE

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S26	U.S. Bureau of Labor Statistics 2026 <i>Industry and Occupational Employment Projections, 2024-34</i> Software developers (SOC 15-1252): 1,693.8k (2024) to 1,961.4k (2034), +267.7k / +15.8%. Ranked 24th of 832 occupations by projected percent change. www.bls.gov/opub/mlr/2026/article/industry-and-occupational-employment-projections-overview.htm	U.S. labor market Official employment projections	U.S.-specific and occupational projections cannot capture all role redesign within software teams. BLS blocks automated access (403). Figures require a human to open the Occupational Outlook Handbook and confirm 1.6938M -> 1.9614M and...	■ A · FORECAST ■■■
S30	GitHub Spec Kit 2026 <i>Spec-Driven Development and Agentic SDD</i> Official Spec Kit workflow: constitution -> specify -> clarify -> plan -> checklist -> tasks -> analyze -> implement -> converge; the core default is Spec -> Plan -> Tasks -> Implement. github.com/github/spec-kit/blob/main/docs/reference/agentic-sdd.md	AI coding-agent workflows using GitHub Spec Kit Official open-source methodology and product documentation	Methodology/tooling evidence rather than independent productivity proof; adoption and outcomes vary by project and organization.	□ C · SIGNAL ■■■ VENDOR
S36	JetBrains 2026 <i>Developer Ecosystem Survey 2026 - AI coding agent adoption</i> 90% use AI coding agents at work at least weekly; 68% daily. Claude Code 39% global / 47% US (18% in Jan 2026), primary tool for 31%; GitHub Copilot 21% (29% YoY prior); Codex 16% (3% in Jan 2026); Cursor 12%... blog.jetbrains.com/research/2026/08/ai-coding-agent-adoption-2026	15,000+ professional developers worldwide, ~90% developer/engineer titles Global survey, regional quotas, reweighted by region...	Vendor-run survey (JetBrains sells developer tools and one of the tools measured). Self-reported tool usage. No margin of error published. Retiered to B: this is a vendor-run survey, and the report's own tier definition places...	■ B · EVIDENCE ■■■ VENDOR
S37	Faros AI 2026 <i>The AI Engineering Report 2026 - The AI Acceleration Whiplash</i> Per developer: task throughput +33.7%; epics +66%; PR merge rate +16.2%; AI acceptance 20%->60%. Per organisation: code churn +861%; bugs/dev +54% (9% in the 2025 cut); incidents +57.9%; incidents-to-PR ratio... www.faros.ai/blog/ai-acceleration-whiplash-takeaways	22,000 developers across 4,000+ teams, two years of platform telemetry Observed telemetry with ML/causal analysis comparing each...	MANDATORY CAVEAT: Faros AI sells the engineering-analytics platform this data is drawn from and has a direct commercial interest in the finding that verification is broken. Must never be presented as independent. Customer base is...	■ B · EVIDENCE ■■■ VENDOR
S38	CircleCI 2026 <i>State of Software Delivery 2026</i> Average daily workflow runs +59% YoY (largest jump in seven years), which the authors warn is misleading alone: median team +4%, bottom quartile no measurable increase. Median team feature-branch throughput... circleci.com/blog/five-takeaways-2026-software-delivery-report	~28 million CI workflow runs, 22,000+ organizations, 149 countries Observed CI telemetry, seven years of comparable history...	CircleCI sells CI tooling, though the finding does not obviously favour them. Population is CircleCI customers, not all software organizations. Workflow runs are a proxy for delivery activity, not for value.	■ B · EVIDENCE ■■■ VENDOR

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S39	Sonar 2026 <i>State of Code 2026 developer survey</i> 96% do not fully trust AI-generated code correctness; only 48% always verify before committing. AI = 42% of committed code, expected 65% by 2027. 64% use autonomous agents; 72% of AI users use it daily. 38%... www.sonarsource.com/company/press-releases/sonar-data-reveals-critical-verification-gap-in-ai-coding	1,100+ developers globally Survey. Fielding dates, sampling method and margin of error...	Two independent problems: (1) Sonar sells code-review and code-quality tooling and benefits commercially from this finding; (2) methodology disclosure is thin - no fielding window, no sampling method, no margin of error, no regional...	C · EVIDENCE VENDOR
S40	Murphy-Hill, E., Butler, J. & Savelieva, A. (Microsoft) 2026 <i>Adoption and Impact of Command-Line AI Coding Agents: A Study of Microsoft Early 2026 Rollout of Claude Code and GitHub Copilot CLI</i> Adopters merged ~24% more pull requests than they otherwise would have; lift sustained across the entire four-month observation window. Adoption spread through peer networks; retention correlated with coding... doi.org/10.48550/arXiv.2607.01418	Tens of thousands of engineers at Microsoft Quasi-experimental observational study of an...	Preprint, not peer reviewed. Single company, and that company co-owns GitHub Copilot. Merged pull requests are a throughput proxy, not validated value - the same metric Faros shows rising while delivery stays flat.	B · EVIDENCE VENDOR
S41	Kohl, K. & Carro, L. 2026 <i>When Code Becomes Abundant: Redefining Software Engineering Around Orchestration and Verification</i> No quantitative results; position paper. doi.org/10.48550/arXiv.2602.04830	n/a Position paper	Conceptual, not empirical. MUST-CITE: this is peer-reviewed prior art for this report thesis. Accepted at ICSE 2026, Future of Software Engineering track. Failing to cite it is the fastest way for an academic reviewer to dismiss the report.	A · SYNTHESIS
S42	Alenezi, M. 2026 <i>Rethinking Software Engineering for Agentic AI Systems</i> No quantitative results; structured literature synthesis. doi.org/10.48550/arXiv.2604.10599	Relevant literature and emerging industry perspectives Structured literature synthesis	Conceptual synthesis, not empirical; no stated inclusion protocol. MUST-CITE as prior art for the thesis.	B · SYNTHESIS
S43	Davis, J.C., Amusuo, P.C., Singla, T., Cakar, B. & Davis, K.A. 2026 <i>Cheap Code, Costly Judgment: A Case Study on Governable Agentic Software Engineering</i> 12-week first-person case study; 88 field notes; 420 KLOC production code; 1.16 MLOC of tests, lints, documentation and tooling. doi.org/10.48550/arXiv.2607.01087	One expert engineer, one system (document accessibility remediation) First-person longitudinal case study	n=1 engineer, single system, self-observed. Cannot generalize. MUST-CITE: closest published work to this report proposed operating model, and shares the two-clause title construction that ruled out one of our candidate titles.	C · SIGNAL

ID	SOURCE	POPULATION & METHOD	KEY LIMITATION	TIER
S44	Various (HackerNoon; DEV.to; arXiv 2603.22106; A. Osmani / O'Reilly Radar; Pathmode; W. Vogels/AWS) 2026 <i>Prior use of terms this report uses: context debt, intent debt, cognitive debt, comprehension debt, intent engineering, verification debt</i> doi.org/10.48550/arXiv.2603.22106	— Terminology prior-art search	PRIOR-ART RECORD, not evidence for any claim. Context Debt, Agent Legibility and Intent Engineering are already in use by others; Evidence Engineering collides with Evidence-Based Software Engineering (Kitchenham et al., c.2004); Cost per...	☐ C · SYNTHESIS 
S45	METR 2026 <i>Uplift update: re-running the developer productivity study</i> 57 developers, 143 repositories, 800+ tasks. Returning participants: speedup of -18% (CI -38% to +9%). Newly recruited developers: speedup of -4% (CI -15% to +9%). Original study ran Feb-Jun 2025; the new... metr.org/blog/2026-02-24-uplift-update	Experienced open-source developers; returning participants plus a newly recruited cohort Randomized controlled trial, follow-up to S03	METR states that significant selection bias undermines the conclusion: developers increasingly declined to participate rather than work without AI assistance. Both confidence intervals cross zero, so neither result excludes no effect.	☒ A · EVIDENCE 

Notices and licence

Important notice

This publication contains general information only. Neither the publisher nor the author is, by means of this publication, rendering professional advice or services of any kind, including legal, regulatory, security, engineering or business advice. It is not a substitute for professional advice and should not be used as the sole basis for any decision that may affect your organisation, your systems or your employment.

No representations, warranties or undertakings, express or implied, are given as to the accuracy, completeness, currency or fitness for any particular purpose of the information here. It is provided **as is**. Sources were checked on the date stated; third-party material may have changed, been corrected, been superseded or been withdrawn since.

To the fullest extent permitted by law, neither the publisher nor the author accepts liability for any loss or damage arising from any person's reliance on this publication or any part of it, including the accompanying workbooks and any score, index or recommendation in them.

On the assessment instruments

The assessment instruments in the accompanying workbooks produce results from values **you** supply, weighted according to the author's stated judgment. They are not measurements, not certifications, not audits, not benchmarks, and **not comparable across organisations**. They ship deliberately empty. Do not use them as the basis for procurement, investment or personnel decisions.



TREE

Terms, licence, and the limits of what is claimed.

Forward-looking statements

Statements marked **FORECAST**, and any statement using words such as "will", "by 2030", "is likely to" or "should", are forward-looking. They are **not facts, findings or predictions**. They are conditional extrapolations from limited evidence, and they depend on assumptions that may not hold. No probability is assigned to any of them, and none should be inferred from ordering or from the space each receives. Where this report extrapolates from a measured trend, it does so as a directional signal only. Actual outcomes will differ, possibly materially.

Independence, affiliation and trademarks

This report is **independent**. It was not commissioned, sponsored, funded, reviewed or approved by any organisation named in it. No organisation named here participated in the research, reviewed it before publication, or endorsed it. No payment, product access or other benefit was received from any organisation named in this report in connection with its preparation.

Organisations, products and publications are named for identification, citation and commentary only. This is nominative use and implies no relationship or endorsement in either direction. All trademarks are the property of their respective owners.



SCATTERLINES

Terms, licence, and the limits of what is claimed.

Licences of cited works. Several sources carry licences more restrictive than this report's own. The DORA 2025 report is licensed CC BY-NC-SA 4.0; OWASP GenAI material is CC BY-SA 4.0; the World Economic Forum's Future of Jobs report is all rights reserved. **Nothing here adapts or reproduces any of them.** Figures and findings are quoted for commentary and attributed, which those licences permit; where a chart plots a source's published figures it is redrawn from the numbers, which are facts and not subject to copyright. US Government works (NIST, the Bureau of Labor Statistics) are reproduced freely.

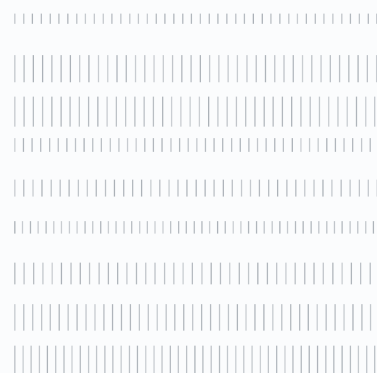
NIST does not endorse this report or any product mentioned in it. **OWASP** does not endorse this report. **JetBrains Mono** is a trademark of JetBrains s.r.o.; the typeface is used under the SIL Open Font License.

How this report was produced, including the use of AI

This report was written with AI assistance, and a report about the trustworthiness of AI-produced work should say so plainly.

AI was used for drafting and redrafting prose from the author's outlines, for summarising and comparing sources the author had selected, and for producing tables, figures and workbook structures from the author's specifications. AI was **not** used to select the thesis, choose the sources, decide what the evidence means, or decide what to publish. **No quantitative claim in this report was accepted on the basis of AI output.**

AI-assisted drafting has well-known failure modes: confident assertion without support, plausible figures that overstate a source's precision, and fabricated internal consistency. The verification protocol exists specifically to catch them, and it caught several in earlier drafts. The author is solely responsible for every claim, every framework, every number and every error. AI systems are tools and are not credited as authors.



BANDS

Terms, licence, and the limits of what is claimed.

The author's views, and a declaration of interest

The views expressed are the author's own. They do not represent the views or positions of any employer, past employer, client or partner. This report was prepared independently, in the author's own time, using no employer resources, no client data and no confidential information from any source.

Declaration of interest. The author is employed by **GlobalLogic**, a Hitachi Group company, which holds commercial partnerships with several of the technology vendors whose published research is cited here. GlobalLogic did not commission, fund, review or approve this report, had no sight of it before publication, and this work was done outside the author's duties there. The author received no payment, product access, early access or other benefit from any organisation named in this report in connection with this work, and holds no financial interest in any of them beyond ordinary employment. The author earns royalties on a book, *The Value Cube*, which argues a position this report's evidence supports; it is advertised on the closing page. This report stamps a commercial interest on twelve of its twenty-six sources and names what each one sells, so it would be incoherent to be vaguer about the author's own.

This declaration is made because a reader is entitled to know it, not because any of it influenced the argument. Where the report relies on vendor-authored evidence, it says so on the figure itself.



KNOTS

Terms, licence, and the limits of what is claimed.

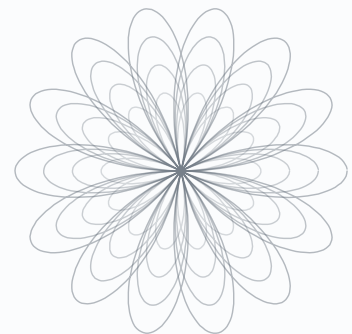
The survey

The survey is **anonymous**. It asks for no email address, no name and no account, and the database has no column and no table that could hold one. What is recorded is your eight answers, the two-letter country code the host derives from the connection, and how you arrived at the page.

No IP address, no cookie and no tracking of any kind is stored. So that one person cannot answer fifty times, the network address is combined with a secret salt and hashed, and only that hash is kept; it identifies nobody.

Results are public and update as responses arrive, so nothing is held back for an edition. The anonymised answers are published under CC BY 4.0 for anyone to re-analyse.

Nothing collected is personal data, so there is no list to leave and nothing to request, correct or delete.



ROSE

Terms, licence, and the limits of what is claimed.

Help settle the questions this report could not

I read 26 sources to write this, of which fifteen report a measurement or survey and the rest are standards, position papers or documentation. Not one of them measured how much AI-generated code a developer actually reads before approving it, or how many organisations have written down what an agent must prove before it may act.

The previous chapter makes five predictions, each with the condition that would prove it wrong. The questions below exist to test exactly those conditions. Eight questions, about nine minutes, no marketing.

01 TESTS: ADOPTION SATURATION

In the last 30 days, how often did you delegate a task to an AI coding agent?

Never · Monthly · Weekly · Daily · Several times a day

02 TESTS: REVIEW RATIONED

Of the AI-generated code you accepted last month, how much did you read line by line?

All · Most · A risk-based sample · A quick skim · None

03 TESTS: REVIEW RATIONED

Does your organisation route code review by risk class, or attempt to review everything?

Written risk-routing policy · Informal convention · Review everything · Don't know

04 TESTS: REVIEW RATIONED

Roughly what share of changes merged in the last month had no human review at all?

None · Under 5% · 5-20% · 20-50% · Over 50% · We don't track this

WHERE

sanjeev-
azad.pages.dev/surveys/where-
the-gain-goes

IT IS ANONYMOUS

No email, no name, no account. The database has no column and no table that could hold one. Answers, a country code and how you arrived, nothing else.

RESULTS ARE PUBLIC AND LIVE

sanjeev-
azad.pages.dev/surveys/where-
the-gain-goes/results

Nobody waits for an edition to see the findings. The response count and channel mix sit at the top, because those tell you how much weight the rest deserves.

The cost: later respondents can see earlier answers, so the sample may be anchored. Responses are dated so anyone can test for drift.

HOW IT WILL BE GRADED

The survey does not close. It is pre-registered at **Tier B at best**, never A: a self-selected online sample is a large survey, and this report puts large surveys at B whoever runs them. Below 400 responses, or with the sample mostly from one channel, it stays on the results page and is not quoted in an edition.



THE SURVEY

05 TESTS: DELIVERY GAP WIDENS

Compared with 12 months ago, how has the time from change written to change in production moved?

Much faster · Faster · Unchanged · Slower · Much slower · We don't measure it

06 TESTS: EVIDENCE COMPLETENESS

For a typical change in your main repository, which of these are produced automatically and linked to the change?

Acceptance criteria · Test results · Security scan · Policy check · Provenance of who or what authored it · Approval record · Rollback record · Post-release outcome · None of these

07 TESTS: AUTHORITY POLICY

Is there a written policy tying how much an agent may do to what must be proven first?

Documented and enforced in tooling · Documented, not enforced · Informal · None · Don't know

08 TESTS: THE JOB CHANGES

Over the last 12 months, how has your own time split moved across writing code, reviewing it, specifying work, and accounting for it?

More / same / less, for each

|||||

WHY THESE EIGHT

Each one tests a condition that would prove one of the previous chapter's predictions wrong. A survey that does not test its own forecasts produces data and predictions that never meet.

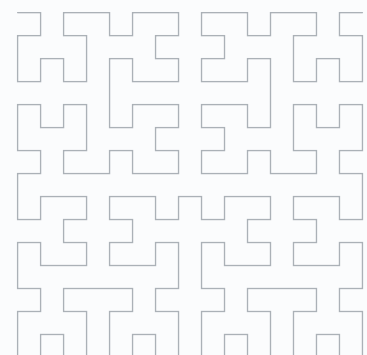
|||||

ON QUESTION 02

Only one published source attempts this as a behaviour rather than an attitude, and it asks whether developers verify at all rather than how much they read [S39]. This is the question this report most wants answered.

ON QUESTION 06

This produces an evidence-completeness distribution, a number that does not currently exist anywhere.



About this report

Licence

© 2026 Sanjeev Azad. Published independently by the author at sanjeev-azad.pages.dev/where-the-gain-goes, part of Innovate with Sanjeev.

This report and the accompanying workbooks are licensed under a **Creative Commons Attribution 4.0 International Licence (CC BY 4.0)**. You are free to share and adapt this material for any purpose, including commercially, provided you give appropriate credit, link to the licence, and **indicate if changes were made**.

Excluded: the report title and visual identity, which are reserved; and third-party material identified as such. Quotations and data from cited sources remain the property of their owners and are used under their own terms.

If you adapt the assessment instruments, changing the tiers, dimensions or levels, please say so clearly and do not present the result as this report's model.

EVIDENCE BASE

26 sources. 12 are authored by organisations that sell AI development or verification tooling, and are flagged as such.

TYPEFACES

Manrope 4.504 and JetBrains Mono 2.304, both SIL Open Font License 1.1. Neither carries a Reserved Font Name.

ARTWORK

Every plate is generated, not licensed: a deterministic construction seeded from the page it sits on. No stock photography is used anywhere.

How to cite, and how this gets corrected

Citation

Sanjeev Azad (2026). *Where the Gain Goes* (Edition 1.0). Innovate with Sanjeev. sanjeev-azad.pages.dev/where-the-gain-goes

When citing a statistic, **please cite the underlying source rather than this report**. This report is a secondary source for those figures, and citing it as the origin would misattribute them. Full details are in the ledger.

Corrections

This report will contain errors. When they are found they will be corrected and the correction recorded publicly, in a dated, append-only errata from which nothing is removed. Superseded versions stay available, so anyone who cited an earlier one can still show what it said.

If a cited source is retracted in a way that undermines a conclusion, **the conclusion changes**.

VERSIONING

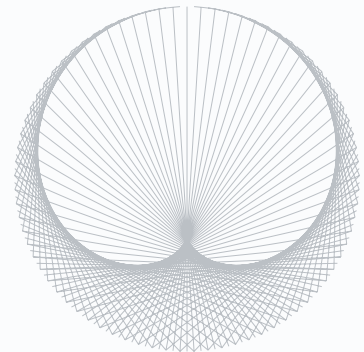
PATCH typography and links. MINOR a corrected fact, logged with the old and new text. MAJOR a conclusion changes.

STANDING INVITATION

If you have primary data that contradicts anything here, send it. It will be published in the errata and incorporated, with credit.

CONTACT

sanjeev.azad@gmail.com
Corrections and questions. The survey holds no personal data, so there is nothing to request or delete. Answered within 30 days.



One rule, applied again and again, returning past where it started but never to the same point. Editions accumulate the same way.

Where this argument continues

This report is about one measurement: what happens to engineering output between the developer who produced it and the organisation meant to benefit. It ends where the harder question begins, which is how an enterprise decides what a trustworthy outcome is worth in the first place.

THE BOOK

The Value Cube

Optimize for value per trusted outcome, not cost per token.

A different way of thinking about enterprise AI. Available on Notion Press and Amazon in India, the US and the UK.

That line and this report's argument are the same idea at different altitudes. This report stays inside engineering and asks what an agent must prove before it may act. The book asks the question a level up: what makes an outcome worth trusting, and what is that worth to the business paying for it.

Declaration. The book is sold, and I earn from it. It argues a position this report's evidence supports, which is a commercial interest in that evidence and is disclosed here for the same reason every other one in this report is.

If you found the argument here useful, the disagreements are more useful still. Both the survey and the errata exist to collect them.

THE BOOK



the-value-cube.github.io

THIS REPORT, AND THE SURVEY



[/where-the-gain-goes](https://where-the-gain-goes)



[/survey](https://survey)

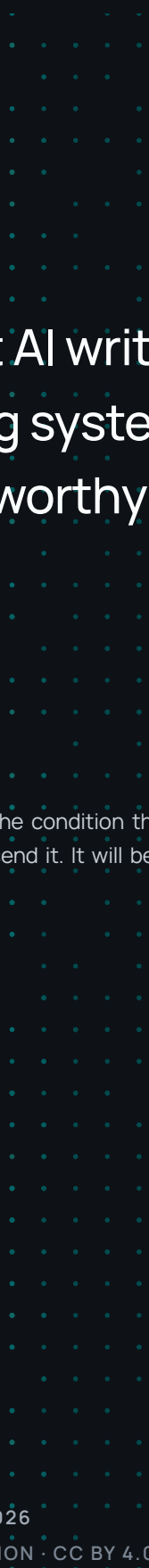
ELSEWHERE

sanjeev-azad.pages.dev

linkedin.com/in/sanjeevazad

x.com/AzadSanjeev

github.com/innovate-with-sanjeev



The future is not that AI writes the
code. It is engineering systems that
turn intent into trustworthy evidence.

|||||

Every prediction in this report is printed with the condition that would prove it wrong. If you have data that contradicts one, send it. It will be published in the errata, and in the next edition, with credit.