weave

REPORT

Weave AI Impact Report: Q2 2026

Published August 2026 | Reporting period: April to June 2026
Router savings and codebase expertise extend into July.
System-measured benchmarks from 1,400+ engineering organizations. No surveys.

Table of contents

Executive summary

18 months into the agentic era, the question has flipped from "is AI making engineers faster?" to "where did the speed go?" This quarter's system data from 1,470 organizations and 21,409 engineers gives an unusually direct answer: in the median organization, output per engineer rose **1.8x** in the three quarters from Q3 2025 to Q2 2026, and nearly every metric leaders use to steer that gain is now measuring the wrong thing.

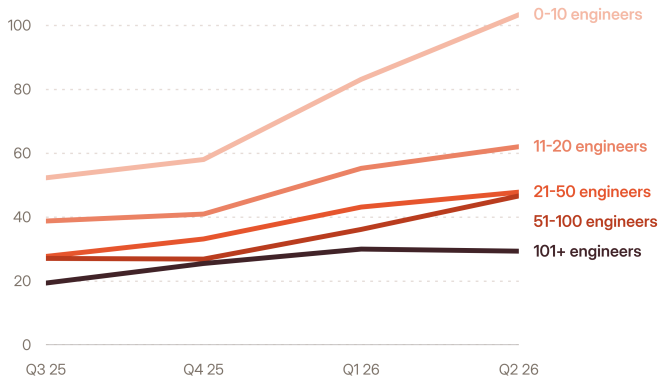
WEAVE ENGINEERING INDEX

Output per engineer rose 1.8x in three quarters

Median monthly output per engineer by team size, Q3 2025 to Q2 2026. Every bucket's median rose, from 2.0x in the smallest organizations to 1.5x in the largest.

22.9%

of the cohort shipped less per engineer at the end of the window than at the start



Source: Weave platform data, Q3 2025 - Q2 2026; constant cohort of 607 orgs active in all four quarters, each pinned to its Q2 2026 size bucket. weaveos.com

Four themes define the quarter:

1. Value and volume have decoupled

Volume and value now move on separate curves. Since January 2025 merged PR volume is up 6x and the median merged PR grew from 36 lines to 109, while the value each PR carries rose 59%: three times the code for half again the value. The input side inflated further still, with tokens consumed per unit of shipped output up 14x since January. Platform-wide, AI's share of merged output value crossed the halfway mark in June at 52%, up from 41% in April. Every volume metric (lines, tokens, PRs, adoption) overstates progress; only complexity-weighted output survives as a ruler.

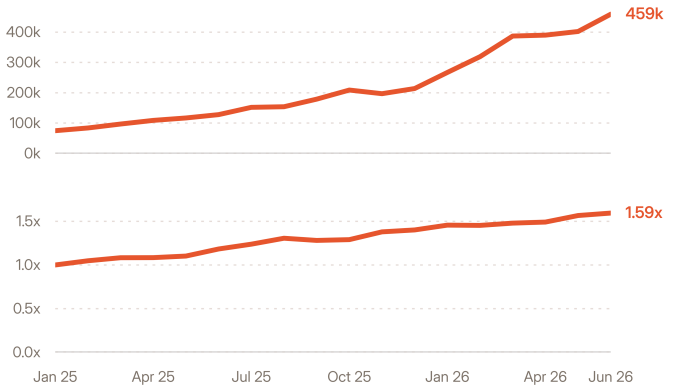
LEGACY METRIC INDEX

PR volume is up 6x, and each PR carries 59% more value

Merged PRs per month across the platform, above, against output value per PR indexed to Jan 2025, below. Both panels are zero-based and share the x-axis.

9.8x

more output value a month than Jan 2025, the two rates compounded



Source: Weave platform data, Jan 2025 - Jun 2026. Output value per PR is indexed to Jan 2025 at 1.00x.

weaveos.com

2. The waste is not where you think

Of every $100 in AI session spend, only $26 ends in a clean ship: a session that completes its work and whose code merges. Another $29 merges without a completion signal, $20 completes and never merges, $17 carries no classified outcome, and $8 ends abandoned or blocked. Fewer than 4 in 10 agent sessions with a classified outcome complete. Of the tool calls that fail, the largest classified bucket at 41% is a file, path, or resource the agent expected and did not find, though an unclassified bucket of the same size sits beside it. The highest-leverage fix we can measure is boring: sessions that invoke a codified skill add 5.4x more lines.

WEAVE AI IMPACT INDEX · AGENTS

$20 of every $100 in session spend completes unmerged

AI session spend allocated by what happened to the session and its code, Q2 2026.

$26

of every $100 is unambiguously shipped, completed and merged

	PER $100
Shipped: completed and merged	$26
Merged, no completion signal	$29
Completed, never merged	$20
No classified outcome	$17
Abandoned or blocked	$8

Source: Weave session telemetry joined to merged PRs, Q2 2026; 21,907 sessions, $458k. Segments rounded individually; may not sum to totals.

weaveos.com

1.8x
growth in output per engineer at the median organization over nine months, constant cohort; 22.9% of it shipped less

6x vs 59%
growth in merged PR volume against growth in the value each PR carries, since January 2025

$26
of every $100 of session spend ends in a completed session whose code merges

3. The quality panic is aimed at the wrong layer

The industry expected AI volume to break the quality gates. It hasn't: revert rate fell below 1% while monthly merged PR volume reached 432,000, and the median wait for review is at least 2.5x shorter than its 2025 peak. AI-signed review events grew 16x between Q4 2025 and Q2 2026, though they are still 1.1% of all reviews. The real structural risk is concentration: the top 10% of engineers produce roughly 40% of output at every organization size, and in 100+ engineer organizations 14% of the team holds half the codebase expertise.

4. Spend became governable, but most orgs haven't started

Routing is the AI cost lever that requires no behavior change from engineers. In July 2026, the most recent complete month, the Weave Router saved 58% against list price, and open-weight models generated 61% of those savings on 31% of routed traffic. Almost nobody is doing it yet: fewer than 1% of engineers on the platform routed a request in July, the Weave Router's second month of general availability. Read that as a starting line rather than a weakness. The lever is two months old, and routing is only now starting to look like a normal line item instead of an experiment. The spending gap between the top 1% of engineers by output ($34/day) and the bottom half ($2/day) is not where the waste sits; ungoverned defaults are.

The bottom line for leaders

The gain is real, and it is table stakes; your competitors got it too. What separates organizations now is conversion: measuring value instead of volume, aiming enablement at the bottom half of the distribution, engineering agent context instead of prompt phrasing, and routing every request like the purchasing decision it is. The rest of this report is the benchmark grid, segmented by organization size, so you can locate yourself against peers rather than averages.

<1%
revert rate in June 2026, on 432,000 merged PRs after bot and noise exclusions

58%
saved against list price by the Weave Router in July, the most recent complete month

61%
of those savings came from open-weight models, on 31% of routed traffic

Introduction: measuring what actually shipped

Welcome to the Weave AI Impact Report for Q2 2026: benchmarks for how engineering organizations really use AI, and what they really get for it, measured from system data across 1,470 organizations and 21,409 engineers.

Most industry reports on AI impact lean heavily on surveys. Surveys measure how engineering *feels*, and feelings matter, but they lag reality and they anchor on the loudest voices. What they cannot tell you is whether the code that shipped last quarter carried more value than the code that shipped a year ago. Nothing in these pages is self-reported. Every number is measured from the systems where engineering work happens: pull requests, review events, agent sessions, token streams, and spend.

The measurements are organized around a concept we call **output**: a complexity-weighted measure of shipped code value. Output is not lines of code. A thousand lines of generated boilerplate and a ten-line fix to a race condition are wildly different units of work, and any metric that treats them equally will mislead you, a problem that gets worse when AI can produce those thousand lines in seconds. Output weighs each merged change by the complexity and scope of what it actually does, which makes it comparable across teams and across time.

The report is built on four indices:

- **Part 1: Weave AI Impact Index.** How AI is changing speed, quality, cost, and (new this quarter) how agents actually behave in production workflows.
- **Part 2: Weave Engineering Index.** Output-based benchmarks for how much teams ship, what kind of work it is, and how concentrated it has become.
- **Part 3: Legacy Metric Index.** The metrics everyone still asks about (PR counts, lines of code, ramp time), and what they hide.
- **Part 4: Prompt Routing Index.** What intelligent model routing saves, and where the savings come from.

PART 1

Weave AI Impact Index

Speed | Quality | Cost | Agents

The AI Impact Index measures the four dimensions of AI's effect on engineering: how much AI contributes and how widely (Speed), what happens to the code (Quality), what a unit of value costs (Cost), and how autonomous workflows actually behave in production (Agents). Every metric is a system measurement; nothing on these pages is self-reported.

SPEED	QUALITY	COST	AGENTS
AI share of output	Revert rate	AI cost per output	Spend by session outcome
AI adoption rate	Review turnaround	Token inflation	Session completion
AI spend per engineer	Review rounds	Output per human cost	Agent autonomy
Skill leverage	AI reviewer share		Failure taxonomy

AI-assisted PRs crossed half of June's shipped output value

AI's share of shipped output value has crossed the halfway mark. In June 2026, 52% of all merged output value came from AI-assisted PRs, up from 41% in April; the median organization went from 21% to 37% over the same three months. The trajectory is steep and, so far, shows no plateau: once AI-assisted code flows through reviews, shared libraries, and collaborative workflows, it permeates the codebase quickly.

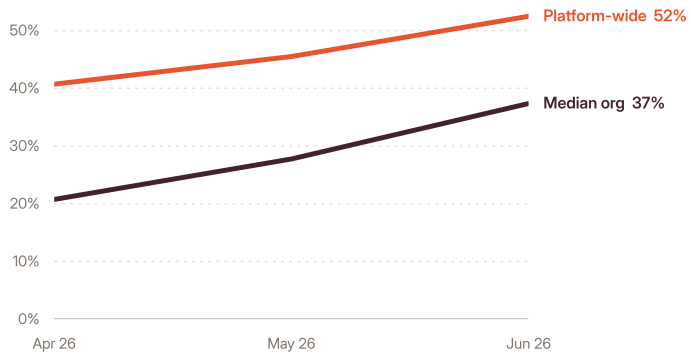
WEAVE AI IMPACT INDEX · SPEED

The median org sits 15 points behind the platform

AI-assisted share of merged output value, monthly. Platform-wide pooled share against the median organization.

17

point rise at the median org since April, against 12 platform-wide



Source: Weave platform data, Apr to Jun 2026; 1,235 organizations in June.

weaveos.com

Read that as a share of *value*, not of volume, because the two diverge sharply. A whole-PR measure like this one marks a PR as AI-assisted if a single AI contributor touched it, and it is the conservative reading: counted line by line, AI's share of Q2's merged code runs far higher still, and higher again or lower by half depending on whether generated files sit in the denominator. Part 3 takes that apart. The reason this report leads on value is that value admits only one denominator.

Headcount adoption is close to saturated: 76% of active engineers platform-wide used AI in the quarter, 16,245 of 21,409. The remaining adoption gap is depth rather than headcount, and it is not where enablement budgets usually go. Engineers below the median output percentile ship 40% of their output with AI, while the band between the median and the 99th percentile clusters near 49%. That nine-point gap sits at the median, not at the top of the distribution, so enablement money goes furthest below the median rather than on power users who have already converged.

For Weave users: Dashboard > AI > AI output percentage.

See this for my team →

Sessions that invoke a skill add 5.4x more lines

Agent sessions that invoke a skill (a codified, reusable procedure for a recurring task) add 5.4x more lines than sessions without one (744 average lines added vs 138) and complete more often (41.2% vs 35.9%). Skills are what turning "prompting" into "process" looks like: they carry the context an agent would otherwise have to be told every session.

41.2% vs 35.9%

session completion rate with a codified skill against without, on 13,673 and 65,845 Q2 sessions

744 vs 138

average lines added per session, a volume measure the paragraph below discounts

Three caveats, because this is the strongest recommendation in the report. Lines added is a volume metric, and the rest of this report argues volume metrics flatter AI, so read the 5.4x as throughput and not as value. Skills get built for high-volume workflows, so selection explains some unknown part of the gap. And these are raw group means with no adjustment for tool, task, engineer, or organization; the median session without a skill adds zero lines, which is how skewed the underlying distribution is. The completion-rate gap is the more load-bearing of the two findings.

For Weave users: Dashboard > AI > Skill usage.

See this for my team →

Volume up, reverts down

If AI were flooding codebases with broken changes, revert rate is where it would show. Instead, the share of merged PRs that get reverted has drifted *down*, from roughly 1.2% through 2025 to 0.98% in June 2026, while the monthly count of merged non-ignored PRs, the denominator behind that rate, climbed nearly 6x to 432,000. One caveat belongs here rather than in the methodology: revert attribution lags the merge, so recent months are floors and not finals.

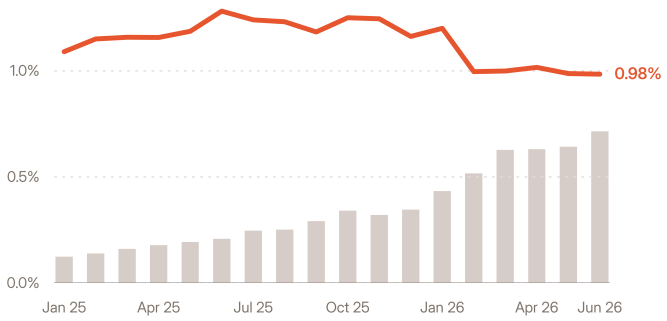
WEAVE AI IMPACT INDEX · QUALITY

Revert rate is falling while PR volume surges

Monthly share of merged PRs that get reverted, as a line, against total merged PRs across the platform, as bars.

1 in 102

merged PRs reverted in June 2026, from 1 in 92



Bars are merged PRs per month, 75k to 432k, on their own unlabelled scale. The line is the revert rate, and it fell from 1.09% to 0.98% across the same period.

Source: Weave platform data, Jan 2025 - Jun 2026. Denominator: merged PRs excluding bot and noise-flagged.

weaveos.com

What that buys a leader is a narrower worry. The volume argument against AI-assisted code now needs evidence beyond throughput, because throughput went up roughly 6x and the failure signal went down. It does not settle the harder questions on the pages that follow: who is doing the reviewing, and how much of the codebase any one person still understands.

For Weave users: Dashboard > Quality > Revert rate.

See this for my team →

The human gate got at least 2.5x faster

Survey-based reports say review turnaround is getting worse. Our source-control data says the opposite: the median wait from review request to review fell from a 2025 peak of about 2.5 hours a quarter to under an hour in Q2 2026, across 829,000 reviews. That wait excludes weekends but not nights, so it is elapsed time with the weekend taken out rather than a business-hours figure. We quote quarters rather than months deliberately. The monthly series moves by tenths of an hour between pulls in both directions, which is enough to swing a month-to-month multiple by a third; quarters pool enough reviews to hold still, and even they moved enough between two pulls that 2.5x is published as a floor and not as a measurement. As with revert rate, read the latest quarter as a floor too: turnaround can only be measured on reviews that have already happened.

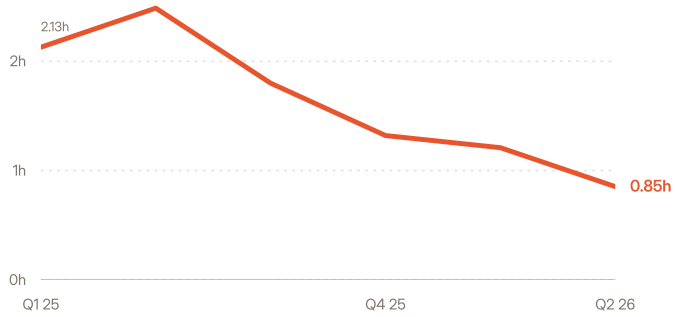
WEAVE AI IMPACT INDEX · QUALITY

The wait for review is at least 2.5x shorter than its 2025 peak

Median wait from review request to review, by quarter, across the platform. Weekends are excluded; nights and other non-working hours are not.

14.6h

the p75 wait in Q2 2026, against a 51 minute median



The p75 wait is stated rather than plotted: it fell from a peak of 20.2 hours to 14.6 where the median fell from 2.5 to 0.9, so the improvement barely reached the reviews that were already slow.

Source: Weave source-control data, Q1 2025 - Q2 2026. Only reviews with an explicit review request are counted, and only reviews that have happened, so the latest quarter is a floor. weaveos.com

A second trend runs alongside it, and we are not claiming it explains the first. AI-signed review events grew 16x between Q4 2025 and Q2 2026, from 0.07% to 1.1% of all reviews. There is no annual multiple to quote because the Q3 2025 base was zero. And at 1.1%, this is a first-pass practice starting to appear in the data, well short of a share that could move a platform-wide median. Whatever compressed the wait, the arithmetic says it was mostly humans.

For Weave users: Dashboard > Process > Code review turnaround.

See this for my team →

What a unit of output costs, by tool

Spend is the dimension where 2026's second half will be decided: budgets approved on faith in 2025 are now being asked for receipts. The first benchmark is what a unit of AI-assisted output actually costs. The median organization pays $7.82 per unit through Claude Code and $11.77 through Cursor.

TOOL	25TH PERCENTILE	MEDIAN	75TH PERCENTILE
Claude Code	**$1.94**	**$7.82**	**$15.15**
Cursor	**$6.07**	**$11.77**	**$23.22**

Org-level median of tool spend divided by the output attributed to that tool, Q2 2026, across organizations carrying both spend and attributable output for that tool. Cursor spend covers IDE and Agent output. The two ranges overlap across most of their span, so read this as two wide distributions rather than a ranking.

These are workflow prices, not token prices: the gap between tools reflects session completion rates and how much generated work actually merges (see the Agents section), not just model pricing.

Where the cheapest units come from

The same price varies far more across engineers than across tools. The highest-output percentile pays a median $0.80 of AI spend per unit of output shipped; the bottom half pay $10.45, 13x more per unit, on far smaller volumes. This cut draws its percentile lines against the full output population ranked on page 13 and then keeps the engineers with observed spend, so its top band is a subset of that population's top 1% and a different set from the spend cohort on that page.

COHORT	MEDIAN AI SPEND PER UNIT OF OUTPUT
Top 1%	**$0.80**
Bottom 50%	**$10.45**

Output bands drawn on the full population of engineers with Q2 output, then restricted to those with observed AI spend. Only the two ends of the series are published, so the middle bands are not shown. Ranked by output, so a low cost per unit at the top is partly built into the cut.

For Weave users: Dashboard > AI > AI cost per output.

See this for my team →

Token inflation: 14x more tokens per unit of output

Tokens consumed per unit of shipped output rose from 0.31M in January 2026 to 4.4M in June, peaking at 4.6M in May: a 14x inflation across the first half of 2026. Agents are thinking longer, retrying more, and burning more context per unit of shipped value, far faster than output itself is growing.

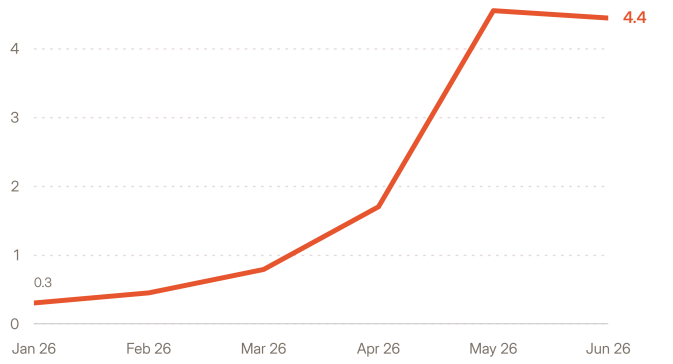
WEAVE AI IMPACT INDEX · COST

Token inflation: 14x more tokens per unit of output

Platform-wide millions of tokens consumed per unit of shipped output, monthly.

5.6x

the rise from March to June alone



Source: Weave platform data, Jan - Jun 2026 (cost telemetry coverage begins Jan 2026).

weaveos.com

Falling per-token prices have masked this inflation; they will not mask it forever. Cost per output is the number to govern. June's reading is 2% below May's, which is one month and well inside the noise of a series that went from 0.8M to 4.6M in the three months before it. We are not calling that a turn, and we would not credit routing for it either: the router covered a small fraction of the platform in June, on traffic Part 4 excludes as unreliable.

A note on measurement

Part of this rise may be our instrument improving, not agent behavior changing. Token telemetry coverage expanded across the same window, so some of what looks like growth is coverage catching up with usage that was always there. We will revisit it in the end-of-year report, when a full year of stable coverage makes the underlying trend separable from the coverage effect.

For Weave users: Dashboard > AI > Token usage.

See this for my team →

AI spend concentrates where output concentrates

Among engineers with observed AI spend, the top 1% by output spend a median of $34 per day, 17x the bottom half's $2, while shipping 84x the monthly output. Heavy spend and heavy shipping travel together, which is the opposite of the pattern a cost-control instinct expects.

COHORT	MEDIAN AI SPEND PER DAY	MEDIAN MONTHLY OUTPUT
Top 1%	**$34**	**729**
Next 9%	**$23**	**191**
Next 40%	**$7**	**51**
Bottom 50%	**$2**	**9**

Engineers with observed AI spend, ranked by Q2 output into four disjoint bands. Spend is a daily median and output a monthly median, so the two columns are not a ratio. The ordering is true by construction.

Output per dollar of engineer cost

How to read output per dollar. It is this report's one cost ruler: units of shipped output bought per dollar spent, where a unit of output is roughly an hour of work. The cost figure on page 11 is the same ratio the other way up, what a unit costs in AI spend. The column below is scaled to $1,000 of payroll, because per single dollar every figure in it is a decimal: 53 means $1,000 of payroll buys 53 units, roughly 53 hours of work.

Human cost dwarfs AI cost and varies far more. This cut ranks every engineer with Q2 output, a wider population than the spend cohort above, which is why the same percentile label covers a different set of people. At a fully-loaded $220k a year, the top 1% converts payroll into output at 53 units per $1,000 and the bottom half at 0.3, a spread of over 170x. The flat salary assumption does real work at the bottom, where the median engineer may be a part-timer or a manager who still merges code. Even the heaviest AI users spend about 5% of payroll on AI.

COHORT	MEDIAN MONTHLY OUTPUT	OUTPUT PER DOLLAR, PER $1,000 OF PAYROLL
Top 1%	**978.5**	**53.4**
Next 9%	**221.1**	**12.1**
Next 40%	**50.3**	**2.7**
Bottom 50%	**5.6**	**0.3**

Every engineer with Q2 output at a fully-loaded $220k a year, or $18.33k a month. The ordering is true by construction; the spread is the finding.

That is an association, not a return on investment; capping the heaviest spenders aims at the wrong end. The governable problem is the long tail.

THE ONE NUMBER TO TAKE TO YOUR CFO

Only $26 of every $100 of session spend ships cleanly

Agents are no longer a feature of coding tools; they are the workflow. Weave observes agent sessions end to end (every prompt, tool call, and correction, and whether the code ever reached production), so $100 of session spend can be followed all the way to its destination. No survey can see this. The split covers the 21,907 Q2 sessions carrying both trace and cost telemetry, $458k of spend, and it is charted in the executive summary.

The full allocation: $26 completes its work and merges, $29 merges without a completion signal, $20 completes and never merges, $17 carries no classified outcome at all, and $8 ends abandoned or blocked. Only the first of those five is unambiguously money well spent, and only the last is unambiguously money wasted; the $75 in between is the part no dashboard reports.

$26

completes its work and merges: the clean ship

$54

merges in total, counting sessions whose code landed without a clean completion signal

$46

never merges: completed-but-unshipped, abandoned, blocked, or unclassified work

The number to sit with is the $20 that completes its work and then never merges: exploration, duplicated attempts, work that lost a race with another change, prototypes nobody promoted. Some of that is healthy. All of it is invisible if you only track spend and completion rates, which is what most AI dashboards give you.

For Weave users: Dashboard > AI > Session outcomes.

See this for my team →

Fewer than 4 in 10 classified sessions complete

Completion rates by tool: Claude Code 38%, Cursor 37%, with abandonment running 35-43%. Both are shares of sessions that carry a classified outcome. In the spend cut on the previous page, 41% of sessions carried none, so the completion share across all sessions is lower than these bars suggest. Treat them as workflow benchmarks rather than tool rankings.

TOOL	SESSIONS	COMPLETED	ABANDONED	BLOCKED
Claude Code	**71,236**	**38%**	**35%**	**6%**
Cursor	**8,282**	**37%**	**43%**	**7%**

Q2 2026 sessions carrying a classified outcome. The three shares do not sum to 100%, because sessions in other classified states are not shown. Claude Code carries many times more sessions than Cursor here, so the rows are not evenly evidenced.

Whether the unfinished majority matters depends on what those sessions cost. An abandoned exploratory session is cheap, and abandoning it is often the right call. The expensive case is the one on the previous page: the $20 in every $100 that finishes its work and still never ships. That, rather than the completion rate, is the number that belongs in a per-seat ROI computation.

Autonomy runs high across the same population. Pooled across 79,518 sessions, 93% complete without a single correction prompt or interrupted tool call, so engineers are not micromanaging agents mid-flight. Against that, only 3.5% of sessions succeed on the first prompt and the average session takes 3.7 prompts. The steering has moved from interrupting the agent to re-prompting it, and iteration is the correction mechanism now. Both numbers improve the same way: context (skills, docs, conventions) and cheap verification (fast tests), so the agent gets it right earlier. We measured autonomy by organization size too, but three of its four buckets rest on two or three organizations, so we are not publishing that cut.

For Weave users: Dashboard > AI > Session completion rate · Agent autonomy.

See this for my team →

What agents actually trip over

Across 3.3 million tool calls, 81,000 failed. The largest *classified* failure, at 41% of all failures, is *not found*: a file, path, or resource the agent expected that wasn't there. It is statistically tied with a bucket of the same size, also 41%, that our taxonomy labels "other" and cannot yet characterize. Timeouts, the failure mode everyone engineers around, are 3%.

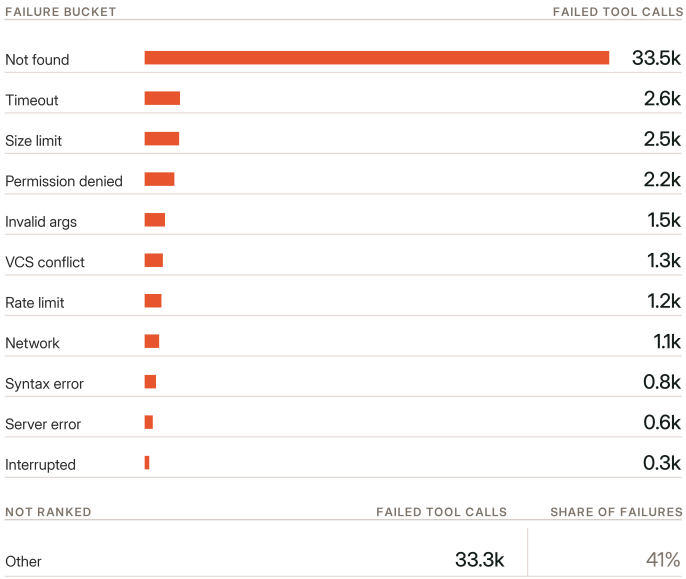
WEAVE AI IMPACT INDEX · AGENTS

Two buckets tie at 41%: 'not found' and 'other'

Failure taxonomy across 3.3M agent tool calls (81k failures), Q2 2026.

2.5%

of 3.3M agent tool calls failed



Source: Weave agent telemetry, Q2 2026. 'Other' is unclassified: we cannot yet say what those failures were.

weaveos.com

The honest reading is narrow. Of the failures we can name, the largest by a wide margin is an agent navigating a repository it cannot see clearly: stale paths, wrong working directories, missing context. That makes repository hygiene and context scaffolding a cheap place to start. It says nothing about how they compare against a model upgrade, and it leaves two fifths of all failures unexplained until we classify them.

For Weave users: Dashboard > AI > Agent failure heatmap.

See this for my team →

PART 2

Weave Engineering Index

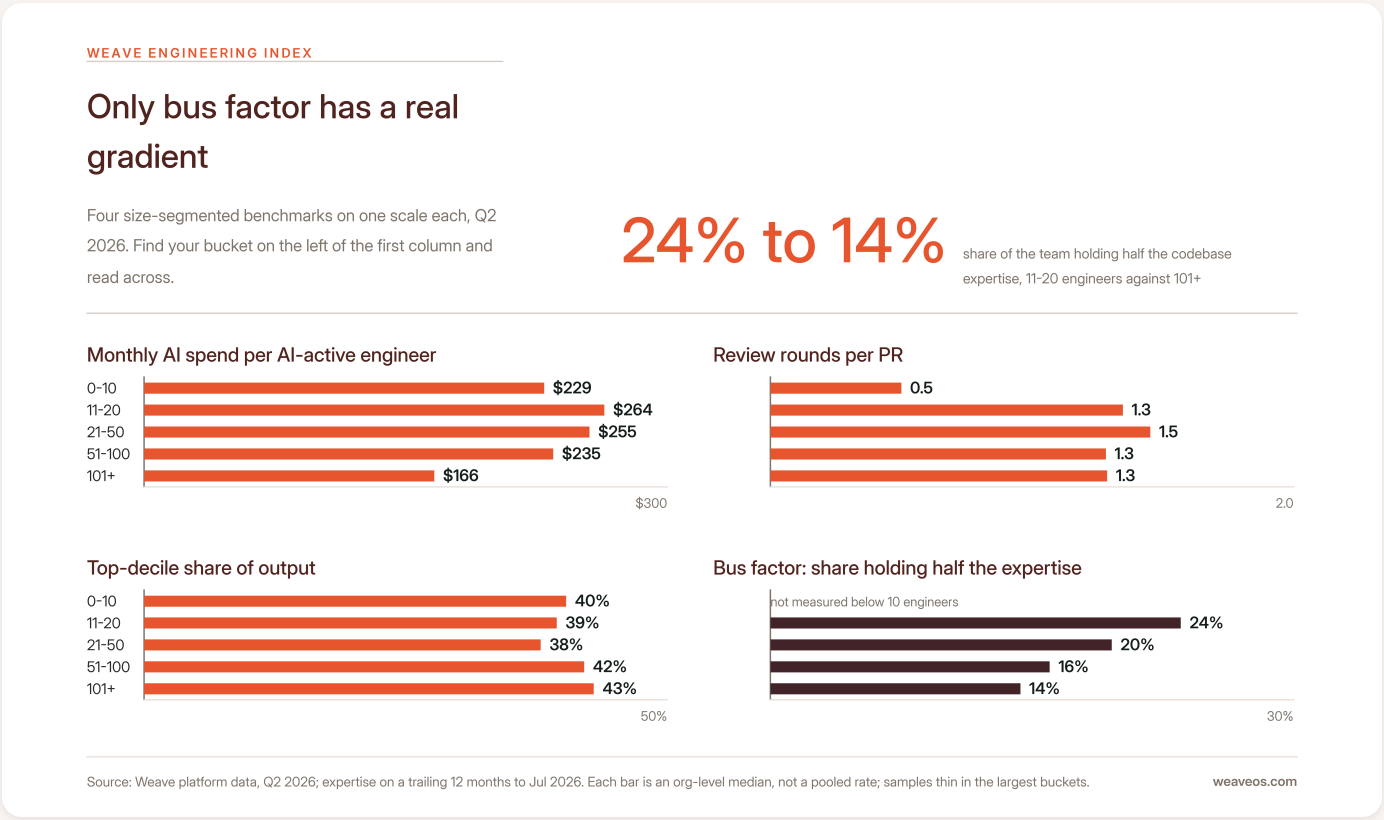
Output | Work mix | Concentration | Delivery | PE spotlight

The Engineering Index answers the questions leaders actually ask in planning meetings: how much should a team our size ship, what mix of work is normal, and how exposed are we to our own top performers? All benchmarks are org-level medians by team size; find your peer group, not the average. It opens with a benchmark grid that also carries three metrics from Part 1, so you can locate your size once and read every size-segmented benchmark in one place.

OUTPUT	WORK MIX	CONCENTRATION	DELIVERY
Output per engineer by size Output trend, constant cohort	Feature vs KTLO vs Bug output shares	Top-10% output share Bus factor (expertise)	Deploy cadence (directional)

Most benchmarks barely move with team size

Find your bucket on the left and read across. Above ten engineers, three of these four lines are close to flat: spend per AI-active engineer peaks at $264 at 11-20 and only falls at 101+, to $166, review rounds run 1.3 to 1.5, and the top decile produces 38% to 43% of output at every size. Only bus factor has a real gradient, and it inverts when you convert it to people: 24% of an 11-20 engineer team holds half the codebase expertise against 14% above 100, so half the expertise in a 15-person team sits with fewer than four people, while a 150-person organization needs 21 to walk out the door. The small team carries the live continuity risk; the large one is where enablement doubles as succession planning.



How to read this grid

Each panel has its own zero-based scale, printed at the right of its baseline, because the four metrics are in different units, and organization counts per bar are in the key. Two readings in the 0-10 bucket are artifacts rather than findings: 0.5 review rounds reflects teams that merge without a recorded review, and the top decile rounds up to at least one engineer, so there the top 10% is really the top 10% to 20%. Bus factor omits the bucket entirely, because the measure is not meaningful below ten engineers. Concentration and review rounds are Q2 2026 cross-sections with no historical series.

For Weave users: Dashboard > Output > Contributor focus · Dashboard > AI > AI adoption rate. **See this for my team →**

How much should your team be shipping?

The median 11-20 engineer organization ships 62 units of output per engineer per month, the peak of the curve above the smallest teams. From there the benchmark falls with scale: 44 for 21-50, 44 for 51-100, and 29 for 101+. The p90 organizations run 1.75x to 5.9x their size peers' median, and in four of the five buckets that within-bucket spread is wider than the distance between any two size medians. The 51-100 bucket is the exception, and only barely. Coordination costs are real, but they are not destiny.

TEAM SIZE	MEDIAN ORG	P90 ORG
0-10 engineers	**55**	**327**
11-20 engineers	**62**	**206**
21-50 engineers	**44**	**100**
51-100 engineers	**44**	**76**
101+ engineers	**29**	**74**

Q2 2026, org-level medians. The median and p90 columns are on the same units but not the same scale of story: the p90 column runs from 74 to 327 where the median runs from 29 to 62, which is why the two were never plotted on one axis. Sample size varies by bucket and is thinnest in the two largest.

Output per engineer rose 1.8x in three quarters

The trend charted in the executive summary is the headline of this report: in the median organization, output per engineer rose 1.8x over the nine months from Q3 2025 to Q2 2026, measured in a constant cohort of 607 organizations active in all four quarters to eliminate mix-shift. Every size bucket's median rose with it, by 2.0x in the smallest organizations (52 to 103 monthly units) and 1.5x in the largest (19 to 29).

Read the 1.8x as the middle of a wide distribution rather than as something that happened to everyone: 22.9% of the cohort shipped less per engineer at the end of the window than at the start. And for the organizations that did gain, it is table stakes, because the gain happened across the whole platform and your competitors got it too. The differentiation questions are the ones this index turns to next: what the new capacity is spent on, and how concentrated it is.

Deploy cadence, directional

Among organizations with CI deploy telemetry connected, median successful deploys per week run from about 16 in the smallest bucket to about 67 at 51-100 engineers, and the ladder is not monotonic: 101+ organizations sit at 43. CI deploy telemetry is a newer integration, so samples may be small and connected orgs skew toward stronger delivery practices. We publish it because DORA-fluent readers will ask, but the output benchmarks above are the better yardstick.

For Weave users: Dashboard > Output > Benchmarks · Dashboard > Process > Deploy frequency. **See this for my team →**

Feature vs KTLO vs bug work

The median 0-10 engineer organization spends 71% of its output on new features; the median 101+ organization spends 60%. The difference splits between two destinations, not one: keep-the-lights-on work rises from 14% to 24% of output across those bands, and bug work from 10% to 17%. This is the output-measured version of what the industry calls the innovation ratio, and unlike the industry's survey-based version, it is moving.

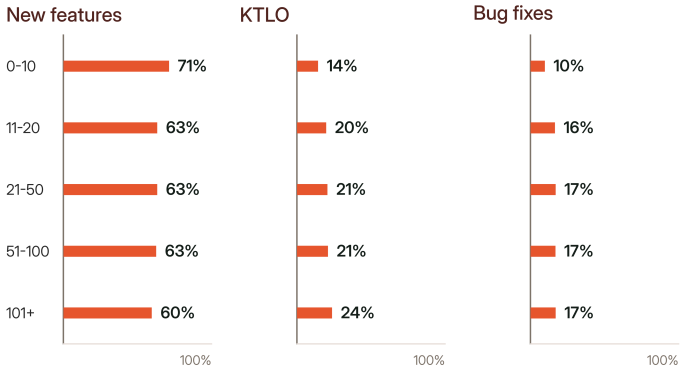
WEAVE ENGINEERING INDEX

Feature work shrinks as organizations grow

Median share of output by work type and organization size, Q2 2026. Three independent medians on one 0 to 100% scale, not three slices of a total.

24%

of output at 101+ engineers is keeping the lights on, against 14% at 0 to 10



Source: Weave PR classifications, Q2 2026. Each share is an independent org-level median, not a share of one total.

weaveos.com

Platform-wide, the feature share of output rose from 54% in Q3 2025 to 63% in Q2 2026 while the KTLO share fell from 35% to 20% and bug work rose from 11% to 17%. The rise in output per engineer is not being absorbed proportionally; freed capacity is tilting toward feature work. The organizations that gain most are the ones doing that deliberately, rather than letting maintenance expand to fill the new capacity.

For Weave users: Dashboard > Output > Output by type.

See this for my team →

Spotlight: PE-backed companies have made AI the default

Private-equity portfolio companies come closest to a top-down AI mandate applied at scale, so we cut the core index for them separately: 17 organizations and 2,883 engineers, each one a company we could confirm is majority-owned by a sponsor. In the median cohort organization 92% of active engineers use AI, and 52% of merged output is AI-attributed — more than half of what these companies ship. Median AI spend is $96 per AI-active engineer per month.

WEAVE ENGINEERING INDEX · SPOTLIGHT

In PE-backed companies, AI writes half the merged output

PE-backed portfolio companies with 50+ active engineers, Q2 2026: 17 organizations, 2,883 engineers. No comparison group; every figure describes the cohort.

52%

of merged output in the median cohort organization is AI-attributed

MEASURE	MEDIAN COHORT ORGANIZATION
AI adoption	92%
AI share of merged output	52%
AI spend per AI-active engineer	$96
Output per engineer / month	23
Revert rate	0.63%

Each row is in its own unit, so nothing here compares down the column. The revert rate is pooled across the cohort; every other row is a median of organization-level values. Spend is over the 16 members with cost telemetry.

Source: Weave platform data, Q2 2026. 17 PE-backed organizations, 2,883 engineers; the spend median is over the 16 carrying cost telemetry. weaveos.com

Output per engineer runs at a median of 23 units a month, and the pooled revert rate is 0.63% across 138,993 merged PRs. Neither figure is set against a comparison group: this cut publishes what the cohort does, not how it ranks. One member carries no AI cost telemetry at all, so the spend median is over 16 organizations rather than 17; every other figure is over all 17.

How to read this cohort

17 organizations is a real sample rather than a handful, but small enough that the spread around each median is wide: AI share alone runs from 31% at the lower quartile to 74% at the upper. Four of the five figures are medians of organization-level values and the revert rate is pooled across the cohort, and the five rows are in five different units, so no row reads against another. Membership was classified by hand from public ownership records, with ambiguous and minority-stake cases left out, so the cohort is if anything undercounted.

PART 3

Legacy Metric Index

PR counts | Lines of code | Time to 10th PR

These are the metrics everyone still asks about. We publish them for exactly one reason: to show, with each one, what it hides. Every chart in this part pairs the legacy metric with its output-based counterweight: the number that tells you whether the activity carried value.

PR counts are exploding, and so is what each one carries

Merged PR volume across the platform hit 459,000 per month in June 2026, up 6x since January 2025 (a figure that conflates platform growth with per-team volume; the per-engineer benchmarks in Part 2 are the ruler). That count is every merged PR. The quality pages count 432,000 for the same month, because the revert rate drops bot-authored and noise-flagged PRs from its denominator. The value index is the surprise: output per PR rose 59% over the same period, so each PR carries more complexity-weighted value. The two rates are plotted against each other in the executive summary, and because they sit on the same underlying work they compound: the platform merges 9.8x the monthly output value it did in January 2025.

The catch is what each PR now contains. On this all-merged-PR series the median PR grew from 36 lines to 109 since January 2025: lines are growing faster than the value they carry, the textbook signature of generated verbosity. PR count remains the most gameable metric in engineering, so if your team's PR count doubled and someone calls it a win, the only honest follow-up is: did output double too?

AI writes two-thirds of merged lines

AI now writes 67% of merged lines, counting only the lines a person or an agent authored. Generated files are excluded and they are the largest category of the three: on a denominator that includes them, AI's share is 36%. Which figure is right depends entirely on what you think a line is worth, which is the argument against steering by lines at all. Any AI target expressed in them invites bloat.

67%

of merged lines were written by AI in Q2 2026, excluding generated code

98

median lines in a merged PR in Q2 2026, up from 54 in December 2025, on the LOC attribution series

For Weave users: Dashboard > Output > PRs per engineer · Dashboard > AI > AI code percentage. **See this for my team →**

New engineers reach ten merged PRs in 26 days, down from 37

The one legacy metric that has improved: median days from a new engineer's first to tenth merged PR fell from 37 for the Q3 2024 cohort to 26 for Q1 2026, with the p75 falling from 119 days to 53. Two things belong alongside that. Almost all of the gain landed early, since the median was already 29 days by Q1 2025 and has moved three days in the four quarters since. And the measure only counts engineers who actually reached ten PRs, so the most recent cohorts, which have had months rather than years to get there, are biased fast.

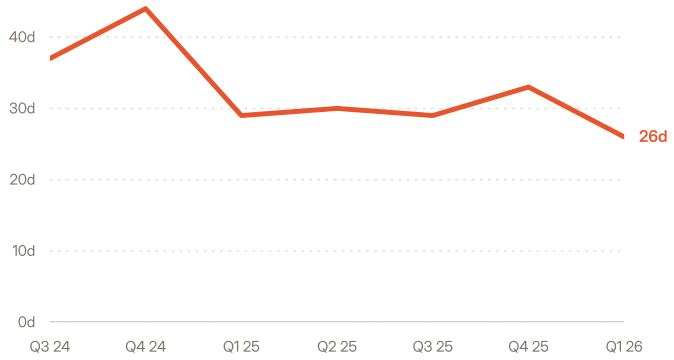
LEGACY METRIC INDEX

From 37 days to 26, most of it by early 2025

Median days from first to tenth merged PR, by onboarding cohort quarter.

3

days gained across the four quarters since Q1 2025



Source: Weave platform data, cohorts Q3 2024 - Q1 2026. Counts only engineers who reached ten PRs, biasing recent cohorts fast. weaveos.com

Agents are a plausible contributor, and the pitch is easy to believe: they answer codebase questions without social cost and produce acceptable first PRs quickly. The timing does not support it strongly, though, because most of the improvement predates the agent surge the rest of this report measures. Either way, ramping *activity* faster is not the same as ramping *judgment*. Pair fast-ramp metrics with review-rounds and bug-attribution data for new joiners before declaring onboarding solved.

For Weave users: Dashboard > Output > Time to nth PR.

See this for my team →

PART 4

Prompt Routing Index

Router savings | Open-weight models

Every prompt an engineering organization sends is a purchasing decision made in milliseconds, usually by default. The Prompt Routing Index measures what happens when those decisions are made deliberately, routing each request to the cheapest model that can actually do the job.

The Weave Router saved 58% against list price in July

The Weave Router prices every request two ways: what it would have cost at the requested model's list price (what a passthrough router charges), and what it actually cost after routing. The gap is the savings. In July 2026, the most recent complete month, 222,946 requests ran through the router, and the routed bill came to 42% of what those same requests would have cost at list. The router saved 58% against the list-price baseline.

58%
saved against the list-price baseline in July, the most recent complete month

42%
of the list-price bill is what those same requests actually cost after routing

222,946
requests routed in July, the month these rates are measured on

The mechanics matter more than the totals: most requests that ask for a frontier model do not need one. The single largest July arbitrage was rerouting Opus-class requests to DeepSeek V4 Flash, 16,854 requests served at 97% below the price of the model that was asked for. Routing changes the bill without changing how anyone works: nobody re-prompts, nobody retrains.

A savings rate is an outcome, not a discount

Inside July the weekly savings rate ranged from 29% to 84%. The rate is set by what engineers ask for in a given week, not by a fixed markdown, so treat 58% as one month's realized result rather than a number to budget against. The routing itself did improve: in June, 49% of reroutes landed on a model costing more than the one requested; in July, 25% did.

Open-weight models punch above their traffic share

Open-weight models served 31% of routed requests in July but generated 61% of the month's savings, roughly twice the saving per request of the routed fleet as a whole. Each routed-to-open-weight request saves disproportionately, because the price gap between frontier list prices and open-weight serving costs is the widest arbitrage in the market. You do not need to bet your stack on open models to benefit from them; you need a router that knows which third of your requests they can handle.

31%
of July's 222,946 routed requests went to open-weight models

61%
of the month's savings came from those requests

2x
the fleet-average saving per request, for requests that land on an open-weight model

For Weave users: Dashboard > Router > Cost savings · Model mix.

See this for my team →

Takeaways

Four themes run through this quarter's data. Here is what to do with each, by role.

1. Value and volume have decoupled: measure value

Merged PR volume is up 6x since January 2025 and the median merged PR grew from 36 lines to 109, while the value each PR carries rose 59%; tokens per unit of output inflated 14x since January. Every volume metric now overstates progress.

- **Engineering leaders:** re-base any AI OKR expressed in lines, PRs, or adoption percentage onto output. This report's size-segmented benchmarks are the starting grid.
- **Platform teams:** instrument cost per output as a first-class SLO next to latency.
- **Finance:** the number that justifies (or kills) next year's AI budget is $/output and its trend, not spend, not seats, not tokens.

2. The waste is in sessions that finish and ship nothing

Only $26 of every $100 of session spend ends in a completed session whose code merges; $20 completes and never merges, and another $26 ends abandoned, blocked, or unclassified. Fewer than 4 in 10 sessions with a classified outcome complete, and the largest classified failure bucket, at 41%, is a file or path the agent could not find.

- **Engineering leaders:** treat merge-rate-of-agent-work as the core agent KPI, not completion rate.
- **Platform teams:** context scaffolding is the cheapest reliability upgrade available: repo hygiene, fast verification loops, and codified skills. Skill sessions add 5.4x more lines and complete 41.2% of the time against 35.9%. Weight the completion gap more heavily; lines added is a volume measure and carries every caveat this report attaches to volume measures.

3. Quality fears are aimed at the wrong layer

Revert rate fell below 1% while PR volume surged 6x, and the median wait for review is at least 2.5x shorter than its 2025 peak. AI-signed review events grew 16x between Q4 2025 and Q2 2026 and are still 1.1% of all reviews, which is a separate development rather than an explanation for the first two. The risk that scales is concentration: top-10% engineers carry roughly 40% of output at every size, and 14% of large-org engineers hold half the expertise.

- **Engineering leaders:** your bus factor is now a delivery risk metric; review it like one.
- **Enablement:** spend below the median: the bottom half runs 40% AI share against ~49% for the band between the median and the 99th percentile. That is where the next step up comes from.

4. Spend is governable now: govern it

Routing saved 58% against list price in July, the most recent complete month, with zero engineer behavior change; open-weight models generated 61% of those savings on 31% of traffic.

- **Engineering leaders:** make new-model adoption a routed rollout, not a stampede.
- **Finance:** demand the two-price view (list price vs routed price) on every AI invoice.

The bottom line

If you do one thing with this report, change your denominator. An organization still steering on lines, PR counts, or adoption percentage in H2 2026 will read a gain that happened to everyone as a win it earned, will fund whichever tool generates the most volume rather than the most value, and will find out which it was about a year late. That is the cost of doing nothing here, and it is paid quietly. Every benchmark in these pages is available for your own organization, against your own size peers.

Methodology

Sample

This report draws on system data from organizations using the Weave platform: 1,470 organizations and 21,409 engineers with merged output in the quarter, spanning startups under five engineers to organizations with several hundred. The primary reporting window is Q2 2026 (April 1 to June 30, 2026); monthly trend views extend back as far as January 2025 and the ramp-time cohorts to Q3 2024, and two cuts extend forward into July 2026: router savings, whose published month is July, and codebase expertise, which is measured on a trailing 12 months to July. Both are labeled where they appear. Engineer-level percentile analyses use two populations: 19,972 engineers with shipped output in Q2 for the output-per-cost cuts, and the 8,264 of those with observed AI spend for the spend cuts. Where the percentile line is drawn matters as much as which population it is drawn on, so the same top 1% label covers a different set of people in each cut. The spend and output cuts sit side by side on page 13 for that reason. The cost-per-output cut on page 11 is a third set again: it draws its percentile lines against the output population and then keeps only the engineers with observed spend, so its bands are narrower than the label implies. Throughout the report we give percentile bands rather than cohort headcounts.

What we measure, and what we don't

Every metric in this report is a system metric, collected from source control, review systems, CI pipelines, agent session telemetry, token streams, and billing data. This report contains no survey data. Self-reported measures capture sentiment that system data cannot; we cite external survey research where relevant, but no Weave number is self-reported.

Output

Output is Weave's complexity-weighted measure of shipped code value. Each merged pull request receives an estimate based on the scope and complexity of the change, not its line count. Output is explicitly not lines of code: the two are stated separately wherever both appear. Bot authors, reverts-of-reverts, and PRs flagged as noise are excluded; generated and vendored code is filtered before estimation.

Attribution

- **AI attribution:** PRs are attributed to AI tools via editor and agent telemetry joined to commits. "AI-assisted output" means output on PRs with a detected AI contributor. AI authorship is never inferred by inspecting the code itself; if telemetry is not connected, the work counts as human-authored.
- **Session-to-production:** agent sessions link to merged PRs through commit and branch matching; this is how session spend allocates across shipped, merged-without-completion, completed-never-merged, abandoned, and blocked.

Organization size buckets

Benchmarks segment by active engineer count into one scheme used throughout the report: 0-10, 11-20, 21-50, 51-100, and 101+. Bus factor omits 0-10, because the measure is not meaningful below ten engineers, and its row in the benchmark grid is marked accordingly. Benchmarks are medians of org-level values, so large organizations do not dominate pooled numbers.

PE-backed cohort

For the Part 2 spotlight, each organization's primary code host identifies its company; PE majority ownership was judged by hand against public records, and ambiguous cases and minority stakes were excluded. The cohort is the 17 such organizations with 50+ active engineers, 2,883 engineers, and has no comparison group. One member has no AI cost telemetry, so the spend median is over 16 of them.

Router baseline

Router savings compare each request's actual routed cost against the same request priced at the requested model's list price, the cost a passthrough router would have incurred. That is conservative: it credits no savings when the router forwards to the requested model. Published figures cover July 2026, the most recent complete month. Two earlier months are excluded: a small May pilot on unrepresentative traffic, and June (110,312 requests at a 22.8% rate), when the routing model still sent half of its reroutes to a model more expensive than the one requested. We publish savings rates and request volumes but not the router cohort's size or its spend, because those are commercially sensitive. Weekly variance inside July is discussed on page 26.

Limitations

- The sample skews toward organizations that measure engineering, likely ahead of the industry on AI adoption. Treat benchmarks as "measured-org" benchmarks.
- AI attribution requires connected editor/agent telemetry, and small teams are least likely to have it set up: a quarter of organizations under ten engineers register zero AI adoption, mostly missing telemetry rather than absent AI. Read every size-cut AI-share figure as a floor in the smallest bucket; no size-segmented adoption benchmark is published, for the same reason.
- Cost coverage varies: API-metered tools report exact spend, seat-licensed tools are estimated from usage, and low-fidelity rows are excluded from $/output medians.
- Revert attribution lags merges and recent-quarter rates are floors, not finals; review turnaround is a floor too, since it counts only reviews that have happened. It counts only reviews carrying an explicit review-request event, so teams that review without formally requesting a reviewer are absent rather than fast. The wait excludes weekends only; nights are still in it.
- Deploy telemetry coverage is sparse; the deploy cut is directional and labeled as such.
- Agent session outcomes and tool-call failures both carry a large unclassified bucket, around 41% in each; completion and failure shares are stated over classified cases.

About Weave

Weave is the engineering analytics platform for the AI era. By measuring complexity-weighted output (not lines, not PR counts) across source control, review systems, agent sessions, and spend, Weave shows engineering leaders what AI is actually delivering, what it costs per unit of value, and where the next gains are hiding.

Weave connects to GitHub, GitLab, Bitbucket, Linear, Jira, Slack, and every major AI coding tool (Claude Code, Cursor, Codex, Copilot, Devin, and more) and includes the Weave Router, the intelligent model-routing layer measured in Part 4 of this report.

The Weave research team

This report was produced by the Weave research team using the same production pipelines that power the Weave platform. Questions about methodology, requests for custom benchmark cuts, or press inquiries: **research@workweave.ai**.

Benchmark your organization

Every benchmark in this report is available live, for your own organization, segmented against your size peers, including the AI Impact Index, output benchmarks, and router savings. Connect your tools and see your first benchmarks the same day at **weaveos.com**.

Engineering analytics for the AI era.
Measure output, not activity.

weaveos.com