2026 RESEARCH REPORT · FOURTH ANNUAL EDITION

The State of Applied AI Teams

What 1,204 practitioners in 27 countries told us about building, evaluating, and governing AI systems in production — and why evaluation discipline, not model choice, now separates the leaders from everyone else.

1,204

TEAMS SURVEYED

27 countries, 11 industries

68%

SHIP AI CHANGES WEEKLY

up from 41 percent in 2025

2.9×

FEWER INCIDENTS

where merges are gated on evals

44%

LACK A ROLLBACK PLAN

essentially unchanged since 2024

EXECUTIVE SUMMARY

Four years in, this survey no longer measures whether teams use large models — nearly all of them do — but whether they have built the operating muscle to use them well. That distinction has become the sharpest predictor in our data: teams in the top quartile of practice maturity ship more often, roll back less, and report materially higher confidence in what they run. Four findings organize this year's report.

1

Evaluation discipline predicts outcomes

Teams that gate merges on an evaluation suite report 2.9× fewer model-related incidents and detect regressions eleven days sooner than teams relying on spot checks.

2

Adoption is table stakes

91 percent of respondents use model assistance daily. Access no longer differentiates anyone; workflow integration and review capacity do.

3

Tooling demand moved up the stack

Model APIs are commoditized. The unmet demand sits in evaluation infrastructure, dataset curation, and change review — the layer above the model.

4

Governance lags usage by two years

Fewer than a third of teams can say who approved the prompt now in production, a share unchanged since 2024.

ADOPTION IS TABLE STAKES

Adoption questions produced the least suspense of the survey. Ninety-one percent of respondents report daily use of model assistance for at least one core task, and 68 percent of teams now ship model-backed changes in a typical week.[a] The more interesting movement is in *who* does the work: application engineers now outnumber ML specialists in AI-touching roles, and the median team runs its models through the same delivery pipeline as the rest of its software.

[a] A "weekly shipper" merged at least one model-, prompt-, or retrieval-affecting change in a typical week of Q4 2025, by self-report.

Two in three teams ship a model-backed change weekly — up from two in five a year ago.

RESPONDENT ROLE	N	SHARE
Software engineers	401	33%
ML engineers	312	26%
Data scientists	187	16%
Product managers	142	12%
Research scientists	98	8%
Platform and other	64	5%

The sample tilts toward organizations that ship software continuously — see the methodology notes — but the direction of travel is consistent across sectors: AI work has left the lab and joined the release train.

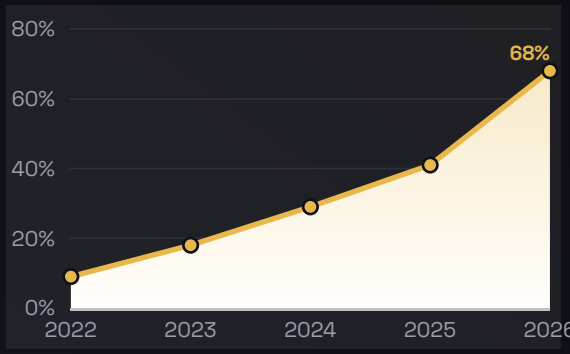
PRACTICE BEATS MODEL CHOICE

Practices, not model choice, split our respondents. Controlling for sector and team size, the model family in production explains almost none of the variance in reported outcomes; the presence of five specific engineering practices explains most of it. We asked about each practice directly and scored progress against the adoption benchmark set by the top-quartile cohort.

PRACTICE	USING	PROGRESS TO BENCHMARK
Prompt and config version control	73%	73 / 80
Offline evaluation suites	61%	61 / 75
Evaluation gates on every merge	52%	52 / 70
Canary or shadow deployment	48%	48 / 70
Human review of high-risk output	39%	39 / 60

The gap is no longer who uses the models. It is who can tell — quickly, cheaply, and before their users do — when the models are wrong.

OUTCOMES ARE REAL, AND CONTESTED



Weekly shippers, 2022–2026 (share of teams)

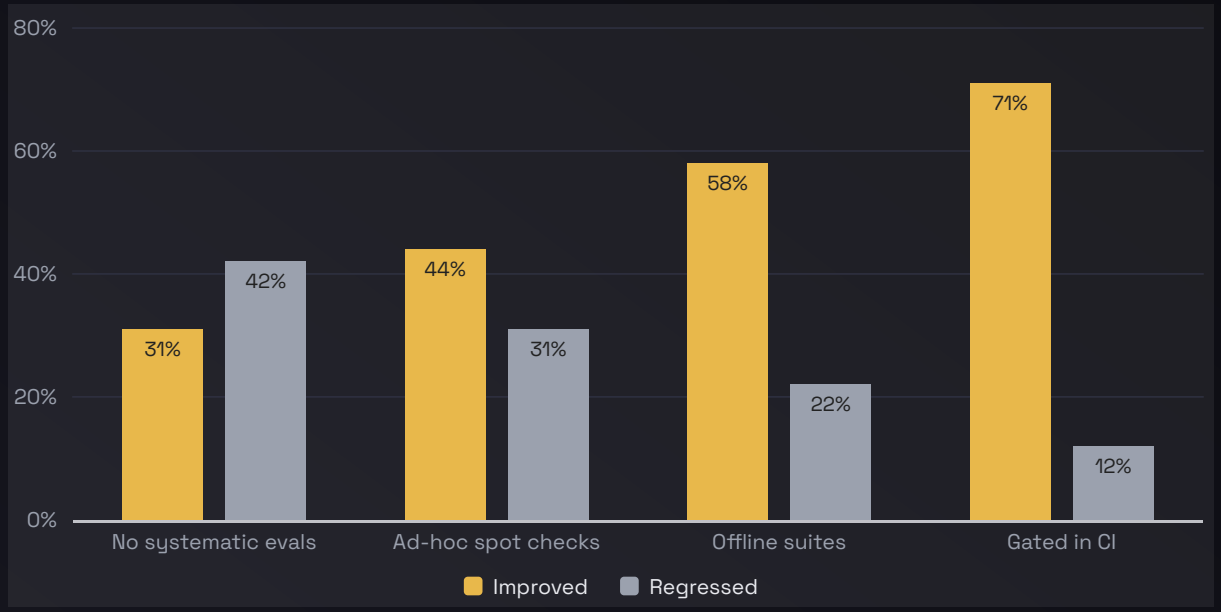
Outcomes divide cleanly along the practice line drawn above. High-maturity teams report shorter cycle times, fewer late-stage surprises, and — most strikingly — a different relationship with quality: they find their own regressions before their users do.

The trend line is steep. In 2022, fewer than one team in ten shipped model-backed changes weekly; this year it is two in three. Every point of that curve arrived with the same complaint: confidence tooling trails shipping tooling. When we asked what respondents would buy tomorrow, evaluation infrastructure topped the list for the third consecutive year.[b]

[b] Respondents chose from twelve investment areas; evaluation infrastructure was ranked first or second by 47 percent of teams.

Reported quality outcomes by evaluation maturity (share of teams)

Evaluation maturity	Improved	Regressed
No systematic evals	31	42
Ad-hoc spot checks	44	31
Offline suites	58	22
Gated in CI	71	12



Quality outcomes by evaluation maturity

Read the last row twice: teams that gate merges on evaluations are not only safer, they are **more likely to report that model changes improved product quality at all**. Discipline does not slow these teams down; it is what allows them to move.

GOVERNANCE LAGS USAGE

Governance is the report's uncomfortable finding. Usage has doubled in two years; the share of teams with a documented approval path for prompt and model changes has not moved. Sixty-two percent of respondents told us that a material model change could reach production without any named person signing off on its behavior.

> Regulated-industry respondents are the exception: documented approval is near-universal in finance and healthcare — but respondents there describe it as retrospective paperwork rather than a gate, completed after the change ships.

The pattern is familiar from earlier platform shifts: capability first, accountability later. The teams closing the gap are treating governance as an engineering problem — approval encoded in the merge gate itself, next to the evaluation suite — rather than as a committee that meets on Thursdays.

Sixty-two percent of teams told us a material model change could reach production without any named person signing off on its behavior.

METHODOLOGY

Sampling ran from 6 January to 20 February 2026: 1,204 complete responses from teams that ship AI-backed functionality, recruited through the Signalworks practitioner panel and community channels in eight languages. We weight responses by region and organization size to a population model of software teams, and report shares with a ±3.2-point margin at 95 percent confidence.[c]

[c] Margins are computed on the weighted sample with design-effect correction; sector-level splits carry wider intervals and are flagged where reported.

Fielded 6 January – 20 February 2026; median completion time 14 minutes; 22 percent of invitees completed.

SURVEY WINDOW

Weights are post-stratification ratios computed per region and organization-size cell, so that a cell's share of the weighted sample matches its share of the population model:

```
def weight(r, p, s):
    """Post-stratification weight."""
    c = (r.region, r.org_size)
    return p.share(c) / s.share(c)
```

Unweighted figures appear in the appendix tables; no headline finding changes direction unweighted.

WEIGHTING

Maturity scores are the unweighted mean of five practice indicators; quartile cuts use the weighted distribution. Comparisons here are cross-sectional — we make no causal claim; the gradient survives every robustness cut we ran.

The full instrument, codebook, and anonymized microdata are published in the report repository.

REPRODUCIBILITY

© 2026 SIGNALWORKS RESEARCH · FOURTH ANNUAL EDITION · CC BY 4.0