WHITE PAPER

Measuring AI ROI Across Teams

A practical framework for proving the value of AI in engineering, sales, marketing and beyond

Den Jones | Founder & CEO, 909Cyber
November 2026

Executive summary

AI ROI is real but only provable when each team measures it against its own outcome metric, not against "hours saved". Engineering should measure delivery speed and quality; sales should measure pipeline and win rate; marketing should measure cost per qualified lead. Everyone shares one rule: capture a baseline before rollout and compare a pilot group against a control group.

Three things make AI ROI hard to prove:

- **Time saved is not value created.** An engineer who saves 4 hours a week only creates value if those hours turn into shipped work, fewer defects or lower headcount growth.
- **Perceived and measured gains diverge.** Developers and sellers often feel faster than the data shows; self reported surveys alone overstate the benefit.
- **Costs hide outside the license.** Review overhead, rework, security controls, training and token spend often exceed seat cost.

A common ROI framework

Every team uses the same formula; only the value drivers change.

> **AI ROI = (Value gained − Total cost of AI) ÷ Total cost of AI**

Value gained comes from four sources, in order of how convincing they are to a CFO:

1. **Revenue uplift:** more pipeline, higher win rate, faster time to market.
2. **Cost avoided:** approved, budgeted requisitions removed from the plan, contracted agency or contractor spend canceled, tickets deflected. Hypothetical hires that were never budgeted do not count.
3. **Risk reduced:** fewer production incidents, vulnerabilities caught earlier, fewer compliance findings.
4. **Capacity freed:** hours saved, only counted when you can show where those hours went.

Count deeper work, not just more work. Teams often spend AI gains raising the quality bar rather than raising output: test coverage moving from 60% to 85%, or compliance checks moving from a 10% sample to 100% automated coverage. Capture these as risk reduced, or velocity metrics will understate the value.

Total cost of AI includes more than seats:

Cost line	What to include
Seat licenses (fixed)	Per user subscriptions; predictable operating cost
Usage and agent compute (variable)	API tokens, agent runs, large context ingestion, RAG queries; can grow far faster than seat count, so set token caps or spend ceilings per team
Enablement	Training, prompt libraries, internal champions' time, knowledge base cleanup
Review overhead	Extra code review, content QA, fact checking AI output
Rework	Defects, reverted PRs, rewritten copy caused by AI output
Governance and security	DLP, data classification, vendor risk review, AI policy, model access controls; legal review of commercial terms (copyright and IP indemnification, no training on your data, zero data retention)
Tool sprawl and shelfware	Overlapping tools bought for the same team and unused seats; audit monthly active use and deprovision seats idle for 30 days. Credit retired point tools (copy, transcription, code review bots) as an offset
Integration	Engineering time to connect tools to CRM, repos, data

Measure in three tiers. Leading indicators (adoption, usage) show up in weeks; operational metrics (cycle time, rep activity) in one to two quarters; business outcomes (revenue, margin, headcount plan) in two to four quarters. Report all three so early adoption isn't mistaken for ROI, and slow outcomes aren't mistaken for failure.

Engineering

Engineering ROI is proven when AI raises delivery throughput without raising change failure rate; faster typing alone proves nothing. This is the team where perception and reality diverge most, so lean on system data over surveys.

> **Perception is not measurement.** In an early 2025 randomized trial by METR, 16 experienced open source developers took 19% longer to complete tasks with AI allowed, yet estimated afterwards that AI had cut their completion time by 20%. Tools have moved on since, so the lesson is not that AI slows developers down; it is that self reported speedups can point the wrong way, which is why DORA telemetry must sit beside any survey.
>
> **90% use AI; 24% trust it a lot.** Google's 2025 DORA report found about 90% of developers use AI and over 80% feel more productive, but only 24% trust its output a lot. It calls AI an amplifier that magnifies both strong and weak engineering practices.

Tier	Metric	How to measure	What good looks like
Delivery (DORA)	Deployment frequency	CI/CD logs	Up vs. baseline
Delivery (DORA)	Lead time for changes (commit to production)	Git + deploy timestamps	Down vs. baseline
Delivery (DORA)	Change failure rate	% of deploys causing a production failure or rollback (failed deploys ÷ total deploys)	Flat or down; up means AI is creating rework
Delivery (DORA)	Mean time to restore	Incident tooling	Flat or down
Flow	PR cycle time (open to merge)	Git analytics	Down, with review time watched separately
Flow	Review time per PR and PR size	Git analytics	Larger PRs and longer reviews signal AI bloat
Quality	Escaped defects, reverted PRs, code churn within 3 weeks	Issue tracker, git	Flat or down
Security	New vulnerabilities per KLOC, secrets committed, vulnerable dependencies	SAST/SCA, secret scanning	Flat or down despite more code
AI specific	Suggestion acceptance rate, % of merged code AI assisted, agent tasks completed end to end	Tool telemetry	Context only; never a target
Experience	Time in flow, toil hours, developer satisfaction (SPACE / DevEx survey)	Quarterly survey	Up; pair with system data
Business	Roadmap items delivered per quarter, engineering cost per feature, budgeted hires removed from plan	Planning tools, finance	The CFO number

Where the dollars come from in engineering heavy companies: budgeted headcount removed from the plan (the biggest lever), faster time to market on revenue features, reduced contractor spend, legacy migrations and test coverage that were previously unaffordable, and lower incident cost.

Watch outs: acceptance rate and lines of code are vanity metrics; AI can shift the bottleneck from writing to reviewing; and junior and senior engineers see very different gains, so segment results by seniority and by task type (greenfield, maintenance, tests, docs). Track unassisted competency and onboarding ramp for junior engineers too; skill atrophy is invisible in quarterly metrics and shows up 12 to 24 months later in the talent pipeline.

Sales

Sales ROI is proven when AI turns admin time into selling time and that shows up as more qualified pipeline per rep. Salesforce's State of Sales report, 7th edition (2026), which surveyed 4,050 sellers in August and September 2025, found the average seller spends only 40% of their time selling, and high performers are 1.7x more likely than underperformers to use AI agents for prospecting.

Tier	Metric	How to measure	What good looks like
Activity	Time spent selling vs. admin	Calendar and CRM activity data, time study	Selling share up
Activity	Personalized touches per rep per day	Sequencer (e.g. Apollo, Outreach)	Up without reply rate falling
Pipeline	Reply and meeting booked rate	Sequencer, CRM	Up; the real test of AI written outreach
Pipeline	Qualified pipeline created per rep	CRM	Up vs. control group
Conversion	Stage conversion and win rate	CRM	Up, or flat at higher volume
Velocity	Sales cycle length	CRM	Down
Efficiency	New rep ramp time to quota	Enablement data	Down (AI call coaching, deal summaries)
Efficiency	CRM data completeness	CRM field audit	Up (auto logged calls and notes)
Business	Revenue per rep, cost of sales, quota attainment	Finance	The CFO number

Watch outs: more AI generated emails can lift volume while burning the market and damaging domain reputation, so track deliverability, spam complaints and unsubscribes beside volume. Attribution is noisy; use a pilot pod against a control pod over at least one full sales cycle. Enterprise B2B deals often run well past 90 days, so read activity and pipeline metrics at the day 90 checkpoint and judge win rate only once pilot deals have closed.

Marketing

Marketing ROI is proven when AI cuts cost per qualified lead or agency spend, not when it produces more content. MIT NANDA's 2025 cross functional study of enterprise GenAI (The Register) found 95% of organizations saw no measurable return. About half of AI budgets went to sales and marketing, yet back office use cases delivered stronger returns, including cutting agency and outsourcing costs. Marketing should target that hard cost line first.

Tier	Metric	How to measure	What good looks like
Velocity	Time from brief to published asset	Project tool timestamps	Down
Velocity	Assets, variants and campaigns launched per marketer	Content calendar	Up, with quality held
Quality	Engagement rate, time on page, organic ranking of AI assisted content	Web and SEO analytics	Flat or up vs. human only content
Testing	Number of A/B variants tested per campaign	Ad and email platforms	Up; more tests means faster learning
Funnel	MQL to SQL conversion, cost per MQL and SQL	Marketing automation, CRM	Conversion up, cost down
Efficiency	Agency, freelance and translation spend	Finance	Down; often the fastest hard dollar win
Efficiency	Customer acquisition cost (CAC), CAC payback	Finance	Down
Brand	Brand consistency score, legal and compliance rework	Review logs	No increase in rework

Watch outs: content volume is the most common vanity metric; generic AI copy can hurt search ranking and brand trust. Track AI search visibility too, since buyers now research through AI assistants as well as search engines.

Other functions

Support and security usually give the cleanest ROI because their work is already ticketed and timed.

Function	Primary ROI metrics	Guardrail metric
Customer support	Ticket deflection rate, first contact resolution, average handle time, cost per ticket	CSAT, escalation rate, reopen rate
Security	Mean time to detect and respond, alerts triaged per analyst, time to complete questionnaires and audits, vulnerability remediation time	False negative rate, AI caused data exposure incidents
IT / service desk	Tickets auto resolved, access request turnaround	Account takeover attempts via help desk social engineering
Finance	Days to close, invoices processed per FTE, audit prep hours	Error and restatement rate
Legal	Contract review turnaround, outside counsel spend	Missed clause rate on sampled reviews
HR / recruiting	Time to fill, recruiter capacity, candidate screening time	Quality of hire, fraudulent candidate rate

AI also creates new risk that belongs in the ROI case: shadow AI, data leakage into public models, deepfake candidates and AI assisted social engineering of the service desk. A security team that measures these risks can show AI governance spend as value protected, not overhead.

Pitfalls to avoid

- **Counting hours saved as dollars.** Multiply hours by salary only if the hours were redeployed or a budgeted hire was removed from the plan; otherwise call it capacity.
- **No baseline.** Without 60 to 90 days of pre rollout data, every result is an anecdote.
- **Surveys alone.** Pair self reported gains with system data; in the METR trial, developers estimated AI cut their task time by 20% when it actually added 19%.
- **Control group leakage.** If the control group uses unsanctioned personal AI accounts, the measured uplift shrinks or disappears.
- **Paying for shelfware.** Overlapping tools and idle seats quietly erode ROI; consolidate vendors and reclaim unused licenses monthly.
- **Uncapped agent spend.** Token and agent compute costs are variable; without per team caps, one runaway workflow can erase the margin gain.
- **Junior skill atrophy.** Staff who lean on AI before building judgment ramp slower later; measure unassisted competency, not just output.
- **Making telemetry a target.** Acceptance rate, prompts sent or content volume become gamed the moment they are goals.
- **Ignoring the downstream cost.** Faster code that fails more often, or more emails that hurt deliverability, is negative ROI.
- **One metric for the whole company.** A single "AI productivity score" hides wins in one team and losses in another.
- **Measuring too early.** Expect a dip during the learning curve; judge outcomes after two quarters, not two weeks.

Measurement plan: 90 day checkpoint, 180 day decision

Leadership gets a checkpoint at day 90 on leading and operational metrics. The scale, adjust or stop decision on business outcomes comes at day 180, once the learning curve has passed.

Days 0 to 30: baseline and design

- Pick 3 to 5 metrics per team from the tables above: one outcome, two operational, one guardrail.
- Pull 60 to 90 days of historical data for each.
- Choose a pilot group and a comparable control group in each team.
- Tally the full cost of AI using the cost table.
- Keep the control group clean: survey or monitor for personal AI accounts (shadow AI), or block them for the pilot.
- Clean the knowledge sources AI will draw on (docs, CRM records, code comments, knowledge base); stale or conflicting data drives hallucinations, review overhead and rework.
- Agree kill thresholds for each guardrail metric before launch (see Example kill criteria).

Days 31 to 60: pilot and instrument

- Roll out tools with enablement, not just licenses.
- Turn on tool telemetry and a short monthly pulse survey.
- Review guardrail metrics weekly (change failure rate, deliverability, CSAT).

Days 61 to 90: checkpoint

- Compare pilot vs. control vs. baseline on leading and operational metrics.
- Fix adoption gaps; stop early only if a guardrail breaches its agreed kill threshold.
- Keep the control group in place for the outcome read.

Days 91 to 180: outcome decision

- Compare business outcomes (revenue, cost, risk) pilot vs. control; extend for teams whose sales cycle runs longer.
- Convert operational gains to dollars only where redeployment or a budgeted cost removed is shown.
- Decide per team: scale, adjust or stop. Re-measure every quarter after that.
- Sign off jointly: the department head (flow and quality), finance (hard savings and requisition offsets) and the CISO or IT leader (data security, access and shadow AI posture).

Example kill criteria

Set these before the pilot starts so a red flag cannot be explained away as adoption friction. Calibrate to your own baseline.

Team	Exit trigger
Engineering	Change failure rate up more than 15% over baseline
Sales	Spam complaint rate above 0.3%, or bounce rate breaching your domain safety limit
Support	CSAT down more than 0.5 points on a 5 point scale, or more than 5 percentage points if tracked as a %
Security	Any confirmed exposure of sensitive data through an AI tool

Sources

- METR: Measuring the Impact of Early 2025 AI on Experienced Open Source Developer Productivity
- Google: How are developers using AI? Inside our 2025 DORA report
- Salesforce: State of Sales report, 7th edition (2026)
- The Register: MIT NANDA report, 95% of organizations see zero GenAI return