

Measuring Generative AI Outcomes



Scope: define, instrument, and report gen-AI workflow outcomes — not benchmark models



Built for programme owners, product teams, operations analysts, and evaluators



Anchor distinction: output quality (model behavior) vs. workflow outcomes (business, human)



Core question: compared with what would have happened without AI assistance?



Why now: ~95% of enterprise gen-AI pilots show no measurable P&L impact



The emerging standard lens: cost per verified outcome

Why Outcome Measurement Differs from Model Evaluation



Benchmarks answer "what can the model do?"; deployment asks "is this workflow fit, reliable, safe, worth scaling?"



Common failure modes: vanity metrics, cherry-picked cases, missing baselines, unverified time-saved claims



Decisions this evidence must support: go/no-go, scaling, redeployment, retrenchment, investment cases



Discipline the language: assistive vs. automated, human-in-the-loop, autonomy level, consequence tier

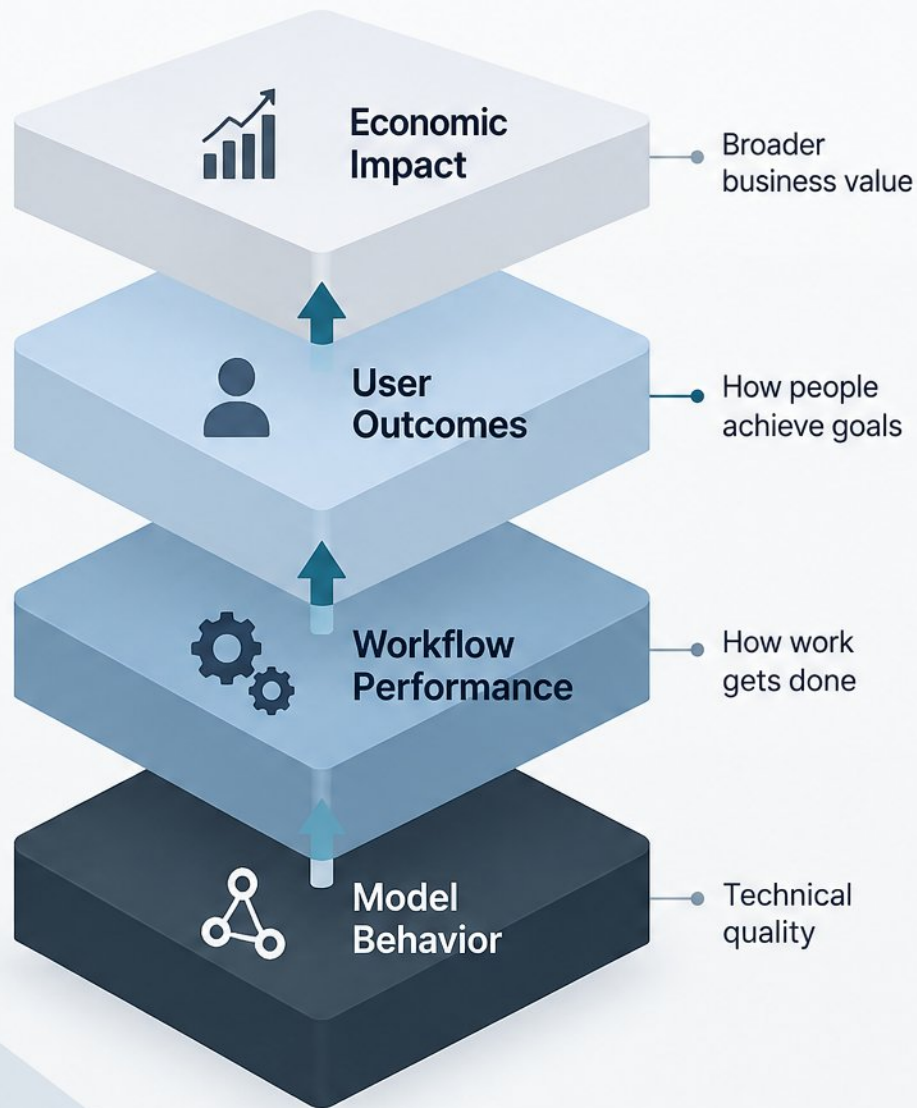


Credibility, not model capability, is the bottleneck — and finance challenges it first



Freeze the socio-technical configuration before reporting any number

The Outcome Stack: From Model Behavior to Business Impact



- Four layers: technical quality → workflow performance → user outcomes → economic impact
- Logic model: inputs → activities → outputs → outcomes → impact; impact is distal and confounded
- Measurement breaks at handoffs, human override, off-platform work, and hidden rework

Logic Model



- Practical rule: measure at every layer, report the layer that answers the decision
- Arrows carry diagnostic value: if correctness doesn't predict efficiency, fix the workflow
- Keep execution, task, and business success as separate verdicts to avoid inflated claims

Framing Outcomes and Designing the Counterfactual

BETTER QUESTIONS
CLEARER EVIDENCE



Write the outcome statement pre-deployment: population, action, metric, threshold, timeframe, aggregation



Counterfactuals ranked: randomized holdout, staggered rollout, diff-in-diff, matched comparison, synthetic control, pre/post



Before/after alone is weak: seasonality, policy changes, learning curves, novelty bias



Diff-in-diff essentials: parallel trends, placebo tests, small-panel inference, multi-wave contamination



Design rollout shape deliberately: timing variation and ring/geography splits preserve identification



Match design to decision: pilot evidence and scaling evidence need different precision

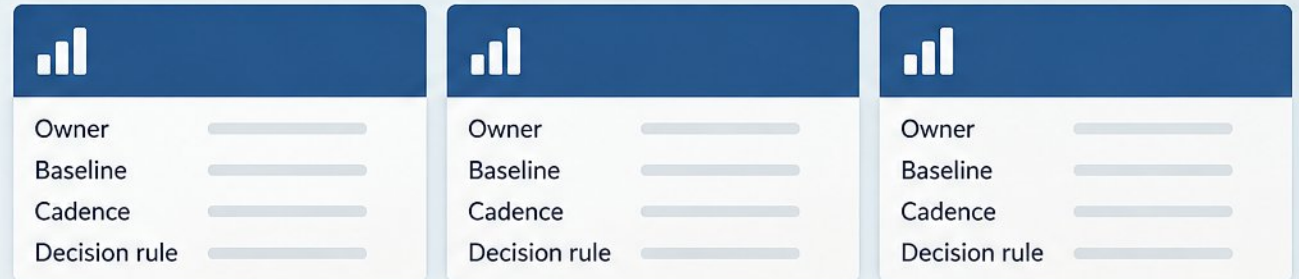


SAME CONDITIONS. A FAIRER COMPARISON.

Metrics Architecture: A Tiered Scorecard

- Six metric families: adoption, quality, human effort, efficiency, value, risk
- Adoption is context, never the headline outcome
- Guardrails share the page with speed: rework, escalations, complaints
- Generous quality set: groundedness, completeness, task fit, citation precision
- Cap the count: three for executives, six org-wide, three per team
- Every metric card needs an owner, baseline, cadence, decision rule

Executive Scorecard (3 metrics)



Organization Scorecard (6 metrics)



Team Scorecard (3 metrics)



Work Units, Cost per Outcome, and Unit Economics

FROM A SINGLE WORK UNIT TO REAL COST

DEFINE | UNDERSTAND | OPTIMIZE

- Define one countable work unit before scaling: resolved ticket, durable pull request, reconciliation
- Cost per successful outcome = $(\text{tokens} + \text{orchestration} + \text{retrieval} + \text{tools} + \text{retries} + \text{review} + \text{build}) \div \text{success rate}$
- Failed and abandoned tasks still burn tokens and reviewer time — happy-path costing understates cost 30%+
- Reviewer minutes and success rate dominate the model line, not inference price
- Cost per task = $F/N + v$; find break-even volume N^* versus the human alternative
- Tiered routing and prompt caching beat cheap retries: stronger models often cost less per completed task



Measurement Methods and Evidence Quality

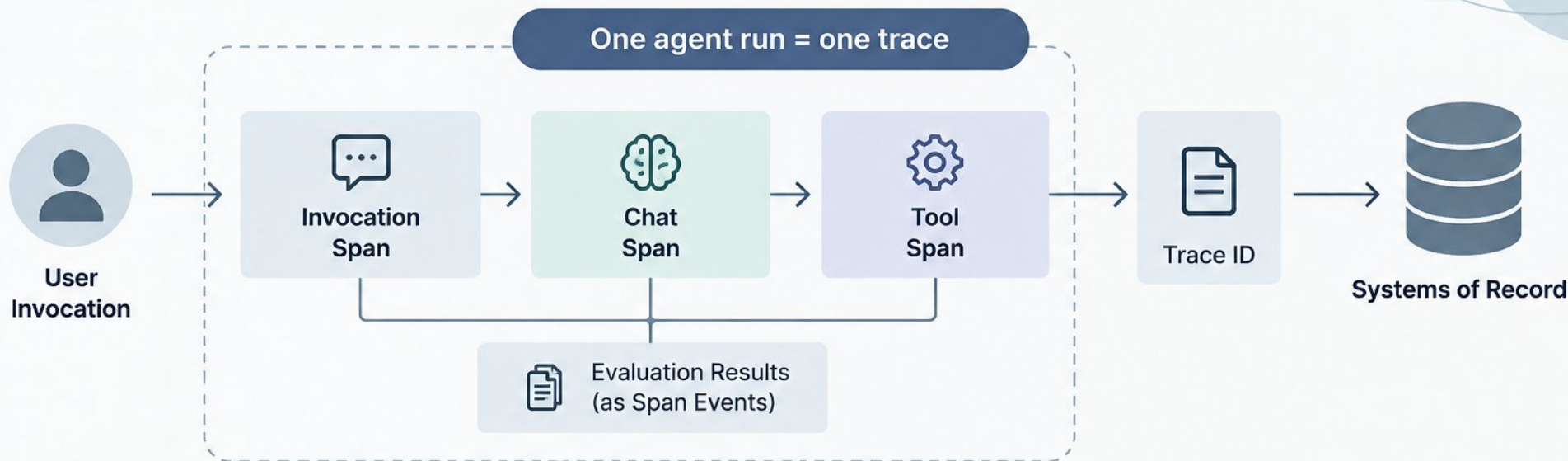
Choose the right methods. Understand the limits. Report a fuller picture.

BETTER
MEASUREMENT
LEADS TO
MORE RELIABLE
DECISIONS

A/B tests	Holdouts	Shadow mode	Human audits	Calibrated LLM judges	Surveys	Diary study
						
Evidence strength High	Evidence strength High	Evidence strength Medium	Evidence strength Medium	Evidence strength Medium	Evidence strength Variable	Evidence strength Variable
Known biases Context dependent	Known biases Selection effects	Known biases Distribution shift	Known biases Distractibility	Known biases Judge variation	Known biases Self-report bias	Known biases Recall bias

- Method menu: A/B tests, holdouts, shadow mode, human audits, calibrated LLM judges, surveys
- Judge choice is a measurement decision — exact-match agreement overstates discriminative ability
- Position bias persists: top judges flip verdicts in over 10% of reversed-order cases
- Report consistency, position flip rate, accuracy, and trustworthy verdict rate together
- Correlated juries add little: within-family error correlation often exceeds 0.9
- Keep judges honest: clear rubrics, blind review, inter-rater targets, recalibration

Instrumentation and Data Readiness



Minimum viable logging: input context, model version, output, human edit, decision



Track identity and cohort: users, teams, task families, risk classes



One agent run equals one trace: invocation, chat, and tool spans linked



Privacy by design: prompts and outputs off by default; sample on outcomes



Attach evaluation results as span events with trace IDs for audit



Join interaction data to systems of record to attribute reopens, reverts, refunds

Cost, Productivity, and Value Realization

- TCO is 3–5x the vendor line; inference is just 18–22%
- Data foundation and integration exceed 40% of TCO
- Full stack: inference, tooling, integration, evaluation, governance, change, maintenance, review
- Net value = gross time saved – review, rework, verification, incident cost
- Repurposed time $\neq$ headcount reduction; boards discount saved hours never reaching the P&L
- Value types: savings, avoidance, capacity, revenue, risk reduction, optionality
- Grade each figure: Verified, Modeled, or Assumed; credit verified at lower bound



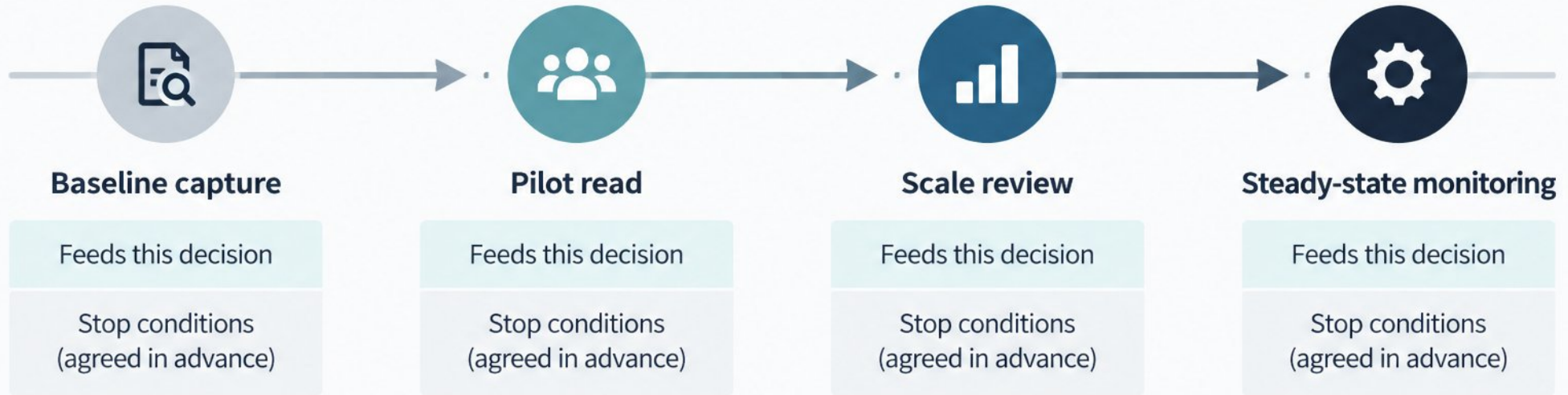
Human Factors, Trust, and Workforce Outcomes

- Measure reviewer experience directly: cognitive load, confidence, usefulness, frustration, context switching
- Higher confidence in AI links to less critical thinking; self-confidence links to more
- Effort shifts from execution to oversight; incoherent interaction hurts output quality most
- Track over-reliance, under-use, abandonment, override rate, and time-to-verify
- Skill effects over time: deskilling, upskilling, dependence
- Who gains time, who absorbs review burden; guard against surveillance creep



Reporting, Review Cadence, and Decision Rituals

A disciplined rhythm from evidence to confident decisions



- Executives need decisions and ranges, operators need drivers, cohorts, traces
- Rhythm: baseline, pilot read, scale review, then steady-state monitoring
- Decision memos with evidence beat live dashboards for consequential calls
- Pre-date decision rules and escalation triggers before launch
- Separate "not proven yet" from "proven not to work"
- Pre-agreed criteria, joint sign-off, and a real no-go option



Drift, Governance, and Regulatory Context

Detect. Govern. Comply. Keep generative AI reliable and responsible.



Drift Classes and Governance

Each drift signal needs a threshold, an owner, and a consequence.

Drift class	Threshold	Owner	Consequence
 Input	_____	_____	_____
 Label	_____	_____	_____
 Concept	_____	_____	_____
 Pipeline	_____	_____	_____
 Vendor-model	_____	_____	_____
 Usage	_____	_____	_____

Regulatory Context

Key obligations for generative AI.

-  EU AI Act: Article 12 logs, Articles 9/15 testing, 72/73 monitoring
-  Annex III high-risk duties apply from 2 December 2027
-  One evidence set can serve both NIST AI RMF and AI Act needs



Six drift classes: input, label, concept, pipeline, vendor-model, usage

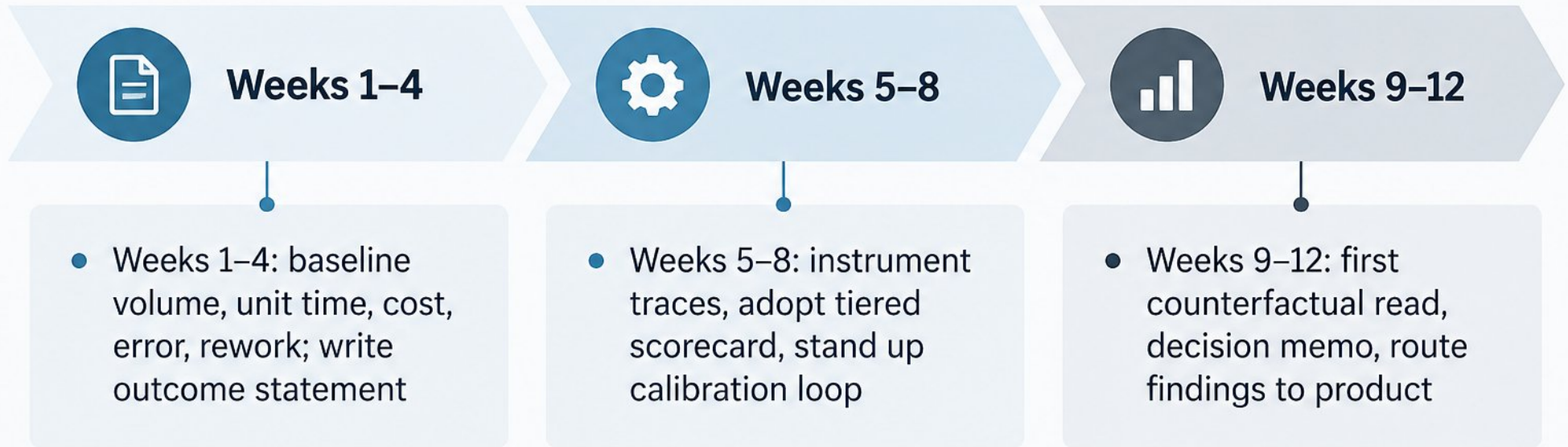


Green infrastructure metrics can hide degraded generative output



Every drift signal needs a threshold, an owner, and a consequence

Building a 90-Day Measurement Plan



Maturity path: baseline and ad-hoc → defined metrics with owners → integrated gates



Scale where outcome density improves; redesign where review burden rises



Leave behind: outcome template, metric cards, evidence rubric, trace schema, gate memo

Key Takeaways and Decision Checklist



Measure every layer; report the layer the decision needs. Adoption is never the headline.



No baseline, no impact claim — state the counterfactual and disclose what it cannot rule out.



Pair every speed metric with quality and rework on the same view.



Use cost per verified outcome, net of retries, review, and rework.



Grade each figure Verified, Modeled, or Assumed; credit value conservatively.



Instrument early, sample on outcomes, set stop conditions before launch.



Measurement is a decision input, not a proof obligation.

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/5e84668e/public