CHOOSING WITHOUT A WINNER: BAYESIAN FRONTIER INFERENCE FOR LIFECYCLE DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems are evaluated across an expanding range of dimensions, the practical question of which system to use has become increasingly difficult to answer from benchmark scores alone, since those scores compress uncertainty, obscure tradeoffs and often fail to distinguish changes in the systems themselves from changes in the regimes through which they are measured. We develop a Bayesian framework for learning multidimensional AI system frontiers that treats each evaluated configuration as a composite of model, harness and version, inference effort and other execution settings, jointly representing performance, cost, latency, task heterogeneity and measurement uncertainty while preserving the structure of evidence that would otherwise disappear into a scalar ranking. Separating changes in the available set of systems from shifts in task populations, evaluators and operational environments allows the temporal model to describe how the frontier itself evolves, which in turn makes it possible to connect that evolving frontier to organization specific losses, constraints, switching costs, fallback options and planning horizons when deciding whether a system should be adopted, retained, replaced or retired. Extending this decision framework with value of information analysis identifies which additional evaluations are most consequential, yielding a unified basis for choosing among AI systems and revisiting those choices as capabilities, circumstances and the evidence available to judge them continue to change.

1 INTRODUCTION

A team deciding whether to adopt a coding agent, keep the one it has or move to a newer release learns little from a leaderboard position. It needs to know which *configurations* are plausibly best for its workload once completion, cost and elapsed time are all weighed, how uncertain that judgement is, whether an apparent improvement reflects the system or a change in how it was measured and whether the gain justifies switching over its planning horizon. Benchmark scores lose most of this information, averaging over tasks of very different difficulty, reporting one coordinate of a multidimensional outcome and coming from harnesses, autograders and task populations that change between releases (Liang et al., 2023; White et al., 2025).

We argue that, on the public evidence available today, the honest output of an evaluation analysis is usually *not* a winner but a posterior distribution over a set of non-dominated configurations and, once a declared class of organisational losses is added, a set of admissible actions. *Choosing without a winner* means narrowing the choice, quantifying what remains uncertain and identifying which evidence could still change the decision, without manufacturing a ranking the evidence does not support. We make this operational with a staged Bayesian framework (Figure 1) whose unit of choice is a *service configuration* (model artifact, harness and version, inference effort and other execution settings). Source-specific likelihoods give joint posterior draws of each configuration's oriented outcome vector, whose random non-dominated set is the frontier. Frontier change is decomposed into arrival, mutable-service, workload and measurement terms, each activated only when a bridge identifies it. A decision layer then maps frontier draws and declared, provenance-labelled scenarios to admissible lifecycle actions, valuing an evaluation by whether it could change that set.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

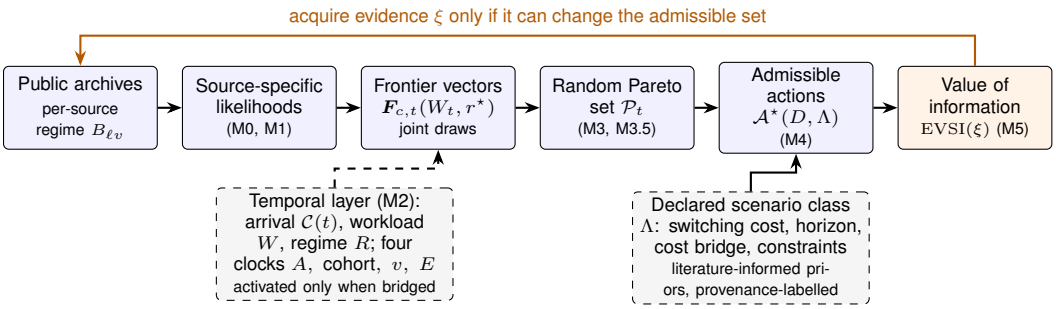


Figure 1: Framework overview. Source-specific likelihoods give joint posterior draws of frontier vectors $\boldsymbol{F}_{c,t}$, which define a random Pareto set. A declared scenario class Λ maps frontier draws to admissible actions, and value-of-information analysis asks which evidence could change them. The dashed temporal layer applies only where bridged and is here an audit, not a fitted model.

Applied to public archives for agentic software engineering and customer service, the framework contributes to inference and identification from public evaluation evidence, rather than to Pareto-set uncertainty or model selection as such, in five parts.

1. **A decision-facing definition of the frontier** (Section 3) as a posterior random set over service configurations, with separated clocks, a four-term decomposition of frontier change and a marginalisation contract that preserves the joint laws decisions need.
2. **Likelihoods that keep what public archives contain** (Section 4): published success totals and pass@k identify task-level success histograms exactly, permitting a Beta-Binomial heterogeneity model without trial data.
3. **Evidence that there is no winner and that scalar rankings discard probable frontier members** (Sections 5.1 to 5.3). On Terminal-Bench v4, 8 of 12 configurations are probable frontier members, and on τ^3-bench the one reconciled pair trades off on a new workload with probability 0.975. A linear weight grid nevertheless gives no region to members whose membership probability is at least 0.964.
4. **A temporal identification audit** (Section 5.4): across regime bridges, release panels and agent traces, public archives identify no temporal dynamics beyond candidate arrival.
5. **Admissible action sets and value of information, with exploratory evidence** (Sections 5.5 and 5.6): within a declared model of literature-informed switching costs, the horizon and the benchmark-to-organisation cost bridge drive admissibility more than benchmark performance, and evidence is valuable at decision boundaries, not where variance shrinks most.

All results are source-specific single-snapshot analyses, conditional on declared scenario classes, and none is a recommendation for a particular organisation (Section 6).

2 Related work

Noisy and shifting measurement. Model comparisons are often underpowered and underreported (Card et al., 2020; Dodge et al., 2019; Bowman & Dahl, 2021), and recent work quantifies noise and ranking uncertainty for language models (Heineman et al., 2025; Yauney et al., 2026; Neuhof & Benjamini, 2026) and critiques pass@k reporting (Hariri et al., 2026). Leaderboards suffer from selective reporting and construct problems (Singh et al., 2025; Bean et al., 2025; Truong et al., 2025), agent benchmarks under-document the configuration behind a score (Moghadasi & Ghaderi, 2026; Gao & Zhou, 2026; Luo & Peng, 2026), selecting among many entries calls for selection-aware inference (Andrews et al., 2024; Zrnic & Fithian, 2025; Roelofs et al., 2019; Blum & Hardt, 2015; Dwork et al., 2015) and latent-variable models pool across items (Lalor et al., 2016; Rodriguez et al., 2021; Vania et al., 2021; Zhou et al., 2026), as measurement guidance now recommends (Keller et al., 2026a;b). Benchmarks themselves move, refreshing tasks to limit contamination (White et al., 2025; Jain et al., 2025; Kiela et al., 2021; Sainz et al., 2023; Chen et al., 2025)

and saturating (Ott et al., 2022; Akhtar et al., 2026), while rankings shift with harnesses, judges and scoring rules (Alzahrani et al., 2024; Biderman et al., 2024; Zhang et al., 2026; Huang et al., 2026; Kalai et al., 2026; Zheng et al., 2023) (see also Han et al. (2024); Longjohn et al. (2024)). We instead infer, per source and without a cross-benchmark latent scale, a multidimensional frontier over deployable configurations, and add an explicit clock and regime decomposition with an audit of which changes current archives can separate.

Frontiers, decisions and adoption. Holistic evaluation reports many metrics (Liang et al., 2023; Liu et al., 2021), leaderboard utility depends on the user (Ethayarajh & Jurafsky, 2020), agent leaderboards now publish cost with accuracy (Kapoor et al., 2025; 2026) and routing methods trade cost against quality per query (Chen et al., 2024; Ong et al., 2025). Multi-objective search that selects among language models by policy optimisation (Monteiro et al., 2026) or searches Pareto-efficient generative-AI pipeline configurations (Conway et al., 2025) contrasts with our aim to infer an uncertain frontier and what public evidence identifies about it. Linear scalarisation recovers only supported Pareto points (Miettinen, 1999; Ehrgott, 2005), Bayesian methods quantify Pareto-front uncertainty (Binois et al., 2015) and econometric frontier estimation treats the frontier as the target (Liu et al., 2022). We bring these ideas to public AI archives and measure how often a weight grid misses probable members that are not supported points. Our decision layer follows statistical decision theory (Berger, 1985) and preposterior analysis (Raiffa & Schlaifer, 1961; Howard, 1966), including Monte Carlo EVSI (Jackson et al., 2022), while robust borrowing of external evidence (Chen & Ibrahim, 2000; Hobbs et al., 2011; Schmidli et al., 2014) and decision-model uncertainty reporting (Briggs et al., 2012) shape how we import priors. Within LLM evaluation, annotation-efficient model selection (Durmazkeser et al., 2026) is closest. Field studies of AI assistance for developers (Cui et al., 2026; Peng et al., 2023; Becker et al., 2025; Fan et al., 2026) and support agents (Brynjolfsson et al., 2025; Wang et al., 2023; 2026), like firm-level switching-cost estimates (Peukert, 2019), enter only as provenance-labelled scenario priors, never transported exactly.

3 Framework

3.1 Service configurations, regimes and clocks

Let m be an immutable model artifact (a provider model version). What an organisation deploys and a benchmark evaluates is a *service configuration* $c = (m, e, z, h, f)$, where e is inference effort, z the tool, scaffold or harness with its version, h the human-review rule and f the fallback rule. Since public evidence observes only some components, the rest stay missing rather than imputed, and a contrast is read as a model contrast only when every other observed component is identical.

Benchmarks are versioned measurement instruments. For family ℓ and pinned version v, the regime $B_{\ell v}$ fixes the task population, its domain structure, the harness, the scoring and aggregation rule and the autograder. Observations follow family $\rightarrow$ version $\rightarrow$ domain $\rightarrow$ item $\rightarrow$ trial. Over time we separate the workload state $W_{\ell vt}$ (which tasks, in which mixture) from the measurement state $R_{\ell vt}$ (autograder, harness, metric, scoring), cross-classifying evaluator provenance with the benchmark hierarchy. Pooling across versions requires a bridge (shared items, configurations evaluated under both regimes or a validated anchor), and no latent scale is assumed across families.

Frontier claims involve distinct clocks that archives routinely conflate: the availability time A_c of a configuration, the problem or cohort time of tasks, the benchmark version v and the evaluation time E at which outputs were produced and scored, the reporting time being auxiliary. Evaluation and reporting times never substitute for A (a model evaluated late is not a model released late), and with $\mathcal{C}(t) = \{c : A_c \leq t\}$ the frontier can move for four reasons, which equation 1 lists as a taxonomy for identification rather than as arithmetic on sets,

$$\Delta \text{ frontier} = \underbrace{\text{arrival}}_{\mathcal{C}(t) \text{ grows}} + \underbrace{\text{mutable-service change}}_{S_{c,t}} + \underbrace{\text{workload change}}_{W_{\ell vt}} + \underbrace{\text{measurement change}}_{R_{\ell vt}}. \quad (1)$$

An immutable artifact is stationary by default, so a hosted-service state $S_{c,t}$ is introduced only when repeated evaluations of one nominal configuration under unchanged semantics identify it, and the workload and measurement terms only when bridges identify them. Archive facts motivate the separation. In the HELM (Liang et al., 2023) history the panel grows from 22 to 68 models over 16 snapshots while all 3,320 repeated model–scenario transitions are numerically unchanged, so the

frontier moves by arrival alone. In LiveBench a task refresh shifts common cells by -10.49 pp but documented-unaffected tasks by $+0.91$ pp, so a measurement change can pass for a system change.

3.2 THE FRONTIER AS A POSTERIOR RANDOM SET

For configuration c, reference workload W_t and declared reference regime $r^\star$, the *frontier vector* is the oriented posterior functional

$$\boldsymbol{F}_{c,t}(W_t, r^\star) = \mathbb{E}_{w\sim W_t}[h(\theta_m, S_{c,t}, w, r^\star, \boldsymbol{\theta})]\,, \tag{2}$$

with coordinates oriented so that larger is better (completion, quality, $-$cost, $-$duration, $-$review burden, ...). The frontier is the non-dominated set

$$\mathcal{P}_t = \mathrm{ND}\{\boldsymbol{F}_{c,t} : c \in \mathcal{C}(t)\}, \qquad c \in \mathcal{P}_t \iff \nexists\, c' \in \mathcal{C}(t) : \boldsymbol{F}_{c',t} \succ \boldsymbol{F}_{c,t}, \tag{3}$$

where $\succ$ means weakly better in every coordinate and strictly better in at least one. Because $\boldsymbol{F}$ is random, so is $\mathcal{P}_t$, summarised by membership probabilities $P(c \in \mathcal{P}_t \mid D_t)$, by pairwise dominance and constraint-satisfaction probabilities and by exclusion risk, each computed per posterior draw. A coordinatewise envelope is not a frontier, since its best coordinates may come from different configurations, and membership is not a function of marginals: if a beats b on coordinate $k \le d$ with probability p_k, then $P(b \in \mathcal{P})$ can lie anywhere in $[1 - \min_k p_k, \min\{1, d - \sum_k p_k\}]$, and plug-in membership from posterior means can fall outside it. Hence a **marginalisation contract**: nuisance variables may be integrated only if the joint law of all frontier vectors remains available, so Pareto and dominance events are never computed from marginal intervals or posterior means. Downstream, expected losses are $\mathbb{E}[L(a, \boldsymbol{F}) \mid D]$ rather than $L(a, \mathbb{E}[\boldsymbol{F} \mid D])$, chance constraints $P(g_k(a, Y) > 0 \mid D) \le \epsilon_k$ use joint tails and organisational state stays explicit. Value-of-information calculations retain every state the evidence can update, and hitting times need joint trajectories: $P(T_{\text{exit}} > t) = P(c \in \mathcal{P}_s\ \forall s \le t \mid D)$ is not a product of marginal membership probabilities.

3.3 DECISION LAYER: ADMISSIBLE LIFECYCLE ACTIONS

Let s_t be the organisation's service state (incumbent, contract and capacity, fallback). Lifecycle actions are $a \in \{\text{adopt}(c), \text{acquire}(c), \text{maintain}, \text{switch}(c'), \text{retire}\}$. Evidence actions ξ are separate. Were losses commensurable over horizon H with discount γ, the criterion would be

$$J(a) = \mathbb{E}[L(a) \mid D], \qquad L(a) = \textstyle\sum_{h=0}^{H} \gamma^h\, \ell(a, \boldsymbol{F}_{t+h}, s_t) + C^{\text{sw}}(a, s_t), \tag{4}$$

with per-period loss ℓ, one-off transition cost C^{sw} (switching, contract and exit), non-monetisable requirements as chance constraints and a Bayes-adaptive decision process as the long-run target. In general ℓ would bundle service, inference, review, correction, harm and uncompleted-work costs, of which public benchmarks identify only completion, per-task cost and duration, our three case-study coordinates. As even these lack an identified exchange rate, we do not minimise equation 4. For a declared class Λ of scenarios (literature-informed priors on switching cost, horizon, cost bridges, transport shocks and constraints) we instead report the *admissible action correspondence*

$$\mathcal{A}^\star(D, \lambda) = \mathrm{ND}\{\boldsymbol{x}(a; \boldsymbol{F}, \lambda) : a \in \mathcal{A}_{\text{feas}}(\lambda)\}, \qquad \lambda \in \Lambda, \tag{5}$$

computed per posterior draw, where $\boldsymbol{x}(a; \boldsymbol{F}, \lambda)$ is the oriented action-level vector (completion, $-$cost including amortised switching cost, $-$duration). Readouts are probabilities such as $P(\mathcal{A}^\star = \{\text{switch}\} \mid D, \lambda)$ and their variation over Λ. Since switching and exit costs attach to actions, an incumbent outside the instantaneous frontier can remain admissible (Farrell & Klemperer, 2007).

3.4 VALUE OF INFORMATION

For an evidence action ξ with outcome D_ξ and a scalar loss L from a declared class, the expected value of sample information (Raiffa & Schlaifer, 1961; Howard, 1966; Ades et al., 2004; Jackson et al., 2022) is

$$\mathrm{EVSI}(\xi) = \min_a \mathbb{E}[L(a) \mid D] - \mathbb{E}_{D_\xi \mid D}\Big[\min_a \mathbb{E}[L(a) \mid D, D_\xi]\Big]. \tag{6}$$

We estimate its paired form, which replaces the first term by the post-evidence loss of the current optimum and so is exactly zero when no replicate changes the action, whereas equation 6 computed by reweighting posterior draws adds a mean-zero Monte Carlo drift. For set-valued decisions we also report the probability that evidence changes $\mathcal{A}^\star$ (applied in Section 5.6).

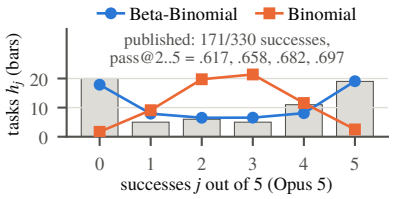


Figure 2: **One histogram fits** the aggregates of Opus 5.

Table 1: **One likelihood per source** (Section B).

Source	Observed	Likelihood	Identifies
Terminal-Bench v4 (M0b/M3)	success totals, pass@2–5	Beta-Binomial (7) on recovered histogram	completion, task ICC
τ^3-bench pair (M1/M3)	12 cell differences, bootstrap SEs	Normal, factor integrated out (8)	mean, workload covariance
Terminal-Bench v4 native trials (M3.5)	trial rows (reward, cost, duration)	nested Bayesian bootstrap	joint quality, cost and duration
LiveBench releases (M2)	common and unaffected deltas	bivariate Normal bridge	drift c_t, revision r_t if $\pi_t > 0$

4 Inference from public archives

Public archives report different things (aggregates, per-cell summaries with bootstrap errors or, occasionally, trial records), so we use source-specific likelihoods matched to what each source reports (exact for counts and histograms, a Gaussian approximation for bootstrap summaries) and never pool sources without a bridge (Table 1, with full specifications in Section B).

Task-level histograms from pass@k. Terminal-Bench v4 publishes, per configuration, total successes over $T = 66$ tasks with $n = 5$ trials and unbiased pass@k (Chen et al., 2021) for $k = 2, \dots, 5$, but not which trial belongs to which task. With h_j tasks having j successes, the integers $\binom{n}{k}T(1 - \text{pass@}k) = \sum_j h_j \binom{n-j}{k}$ form a unitriangular system in $\boldsymbol{h}$, so four-decimal rounding loses nothing while $\binom{n}{k}T < 10^4$ (at most 660) and $\boldsymbol{h}$ is identified exactly for all 13 configurations, although this anonymous multiset cannot identify named task effects or cross-configuration covariance. We fit the overdispersed Beta-Binomial model (Skellam, 1948)

$$Y_{cw} \sim \text{BetaBinomial}\big(n, p_c\kappa_c, (1-p_c)\kappa_c\big), \qquad \kappa_c \sim \text{LogNormal}(\log 5, 1), \tag{7}$$

with uniform p_c and intraclass correlation $\rho_c = 1/(\kappa_c + 1)$, against a Binomial$(n, p_c)$ baseline whose curve in Figure 2 misses the mass at 0 and 5 successes. Where Hariri et al. (2026) replace pass@k with posterior summaries computed from trial-level outcomes, the histogram lets us fit such a model when only the published aggregates survive.

Correlated quality, cost and duration. For τ^3-bench the multidimensional panel is one reconciled model pair, observed as differences $\hat{\boldsymbol{\delta}}_i \in \mathbb{R}^3$ in (quality, −cost, −duration) over 12 domain-by-complexity cells with bootstrap standard errors $\boldsymbol{s}_i$. A one-factor workload effect adds dependence to coordinatewise random effects, and since sampling factor scores creates a funnel (Betancourt & Girolami, 2015) we integrate it out exactly (Section B),

$$\hat{\boldsymbol{\delta}}_i \sim \mathcal{N}\big(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\tau}^2 + \boldsymbol{s}_i^2) + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top\big), \tag{8}$$

whose law without $\boldsymbol{s}_i$ gives dominance and tradeoff probabilities for an unseen workload.

Regime bridge, nested bootstrap and diagnostics. When a release revises a documented fraction π_t of tasks, deltas on unaffected common tasks identify a common component c_t and deltas on all common tasks add $\pi_t r_t$. Native trial panels get a nested Bayesian bootstrap (Rubin, 1981) whose task weights are shared across configurations and coordinates. MCMC fits are promoted only with 8 chains, rank-normalised $\hat{R} < 1.01$, bulk and tail ESS ≥ 100 per chain and zero divergences (Hoffman & Gelman, 2014; Vehtari et al., 2021), with adequacy judged separately by predictive checks and held-out log predictive density (LPD) (Gelman et al., 1996; Vehtari et al., 2017).

5 Case studies

Each case study uses one source and one estimand, pooling nothing across benchmarks (provenance in Section A). Every fit meets the sampling standard of Section 4 except the exploratory LiveBench and LiveCodeBench fits of Section 5.4 and a τ^3-bench stress test (Table 4).

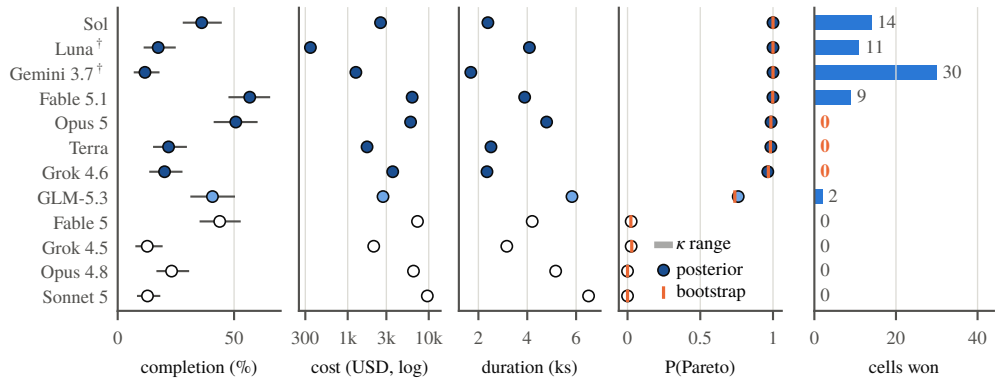


Figure 3: **No Terminal-Bench v4 configuration is best on every coordinate.** The 12 configurations with valid cost, sorted and filled by Pareto-membership probability (dark $\geq$ 0.95, light 0.05–0.95, hollow $<$ 0.05): completion mean and 95% interval, reported cost and mean duration, and membership with task-bootstrap estimates (orange) and the range over κ-prior medians 2, 5, 20 (grey). Membership is conditional on the reported cost and duration, which public summaries give without uncertainty, and for Grok 4.6 and Opus 5 it hinges on small gaps in them (Section E). A native trial-level panel of four configurations, with all three coordinates jointly uncertain and paired by task, keeps the same members (Terra 0.998, Sol $>$ 0.999, Opus 5 0.990; Sonnet 5 0, Section E). †: member by construction (lowest cost or duration). **Right:** cells of the 66-point weight grid won under linear scalarisation (Section 5.3), orange for probable members winning none.

5.1 A STATIC FRONTIER WITHOUT A WINNER: TERMINAL-BENCH V4

Does a leaderboard snapshot, read with its uncertainty, name a winner? The pinned Terminal-Bench v4 snapshot (Merrill et al., 2026) evaluates 13 service configurations on the same 66 tasks with five trials each, in four exact harness, version and effort blocks plus two singletons (Table 5). Only completion in the frontier vector (completion, $-$reported cost, $-$reported mean duration) carries posterior uncertainty (Section 4), duration being whole-trial wall-clock time, not provider latency, and one configuration whose reported zero cost is missing is excluded, leaving 12. The Binomial baseline fails the posterior predictive checks that the Beta-Binomial model passes under every tested prior (Figure 6 and Section C), so treating trials as independent would overstate precision.

In Figure 3 completion, cost and duration rank the configurations differently, and 8 of 12 configurations, drawn from every harness block and both singletons, are probable frontier members, whereas the remainder are almost surely dominated. Neither a task-level bootstrap nor the κ-prior sweep moves these probabilities appreciably. They couple configurations independently, since the anonymous histograms carry no task pairing, but positive dependence, plausible as all share the 66 tasks, only pushes each towards 0 or 1, keeping the same 8 of 12 probable members at every tested correlation, and seven stay at or above 0.80 under every coupling (Section C). Within blocks (Figure 4), most model contrasts are resolved, Opus 5 exceeding Opus 4.8 by $+27.6$ pp [15.5, 39.1], but others, such as Fable 5 against Opus 5, are not (all contrasts in Table 6). These are observational contrasts, and the snapshot identifies no effort or harness effect. A decision maker asking "which configuration is best?" of this snapshot therefore receives a set, not an element.

5.2 A ROBUST MULTI-COORDINATE TRADEOFF: τ^3-BENCH

Does a tradeoff survive when all coordinates and the workload are uncertain? On τ^3-bench (Yao et al., 2025; Barres et al., 2026; Shi et al., 2026) (text mode) only one model pair, Claude Opus 4.5 and GPT-5.2 at high effort, has trajectories reconciling with the published pass$\hat{}k$ statistics across the airline, retail and telecom domains. Quality is unresolved, the posterior mean difference (Claude minus GPT) being $+0.82$ pp [-2.64, 4.18], while Claude is more expensive and faster in all 12 domain $\times$ complexity cells (Figure 7 and Section C), and for a new workload drawn from the fitted heterogeneity the pair trades off with probability 0.975, almost unchanged under independent random effects. Dropping this heterogeneity degrades held-out prediction on every holdout (Figure 4).

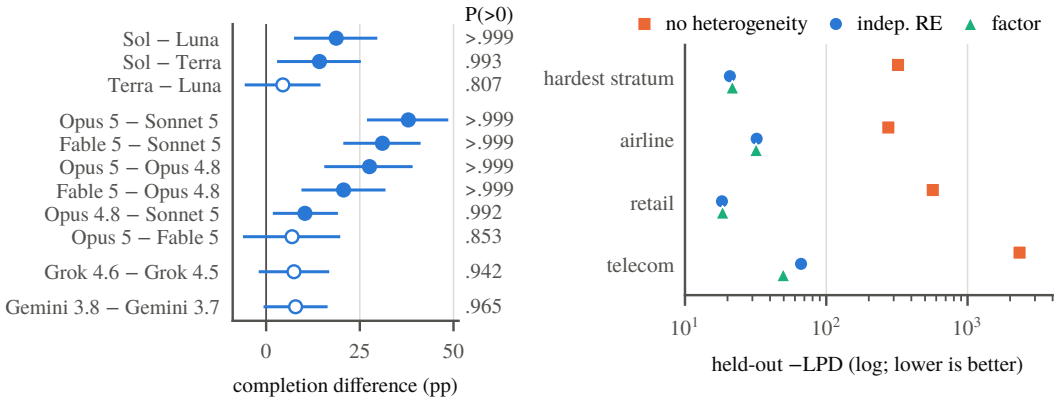


Figure 4: **Left: most within-harness contrasts are resolved.** Terminal-Bench v4 completion differences within identical harness, version and effort, oriented to a positive posterior mean, with 95% intervals (open if covering zero) and $P(> 0)$. **Right: workload heterogeneity earns its place.** τ^3-bench held-out negative LPD with the hardest complexity stratum or one domain held out, without heterogeneity, with independent random effects and with the exact-marginal factor.

The exact-marginal factor model adds materially to independent random effects only on telecom, and the interval of its positive cost–duration dependence includes zero, so we treat that dependence as exploratory. The choice between the two thus turns on cost against speed.

5.3 Scalar rankings discard probable frontier members

A common response to a broad frontier is to fix weights and rank. For each weight vector on the 66-point simplex grid with step 0.1 over min–max normalised (completion, −cost, −duration) we record the configuration maximising the weighted score. Linear scalarisation can select only supported points, on the convex hull of the frontier, and in Terminal-Bench v4 this bites (Figure 3, right): at the posterior means Opus 5, GPT-5.6 Terra and Grok 4.6 win none of the 66 cells and none of the 501,501 weights of a grid with step 0.001, so this is no grid artefact. Per posterior draw they are frontier members with probability at least 0.964 but supported points, winning some weight of a grid with step 0.01, with probability only 0.227, 0.210 and 0.004. The native panel of Figure 3, jointly uncertain in all three coordinates, shows the complementary failure: three configurations are near-certain members, yet any single weight vector endorses only one of them, even without error-flagged trials (Section E). Either way, a fixed weighting hides probable frontier members.

5.4 What public archives identify over time

The decomposition in equation 1 is useful only if data can separate its terms, so we audited every public source with a plausible temporal signal, retaining a temporal component only if, as pre-declared, it improves chronologically held-out prediction under a model that passes its predictive checks. Table 2 sorts the outcome by term. Only candidate arrival is observable: reconstructing the static frontier as entrants appear in a historical airline-domain panel of τ-bench-family submissions, the share of available configurations on the frontier falls from 1.0 to 0.267 and then 0.250 after 1, 7 and 8 entrants (two coordinates) as the candidate set grows, although no incumbent changes. Entrant-quality trends, mutable-service drift and measurement-regime dynamics are not identifiable for want of what Table 2 lists, and simulated recovery shows fitted trend and innovation scales to be prior-dominated. This negative identifiability result is not evidence that the frontier is static.

5.5 Lifecycle decisions under declared scenario classes

When is moving to a newer release the only admissible action? We map Terminal-Bench v4 frontier draws to the actions continue and switch for three version-family pairs fixed before any decision analysis (Opus 4.8→Opus 5, Fable 5→Fable 5.1, Grok 4.5→Grok 4.6). Fallback is feasible but unscored because no benchmark identifies its outcome, so admissibility here is among the scored

Table 2: **Public archives identify arrival and nothing else over time.** Terms of equation 1, sources audited, what they lack and the predeclared outcome (per source in Table 13).

Term	Sources audited	Missing	Outcome
Arrival: candidate set $\mathcal{C}(t)$	HELM snapshots; τ-bench airline panel	nothing (dated entries suffice)	**identified**
Arrival: entrant-quality trend	LiveCodeBench; τ^3-bench events; SWE-bench Verified; τ^2 banking (T0–T2)	within-regime weeks ($\leq$5 occupied weeks vs. floors of 7 (T2) and 9 (T3)) and a stable task mix	not identified (failed checks, confounding, prior-dominated recovery)
Mutable service $S_{c,t}$	τ^2 event panel; τ^3/LCB-Pro/AstaBench dates	stable identities (0 of 102 events complete) and authenticated execution clocks	blocked
Workload, measurement W, R	LiveBench, 10 adjacent releases	a bridge informative about revision effects r_t	inconclusive; dynamic regime not earned

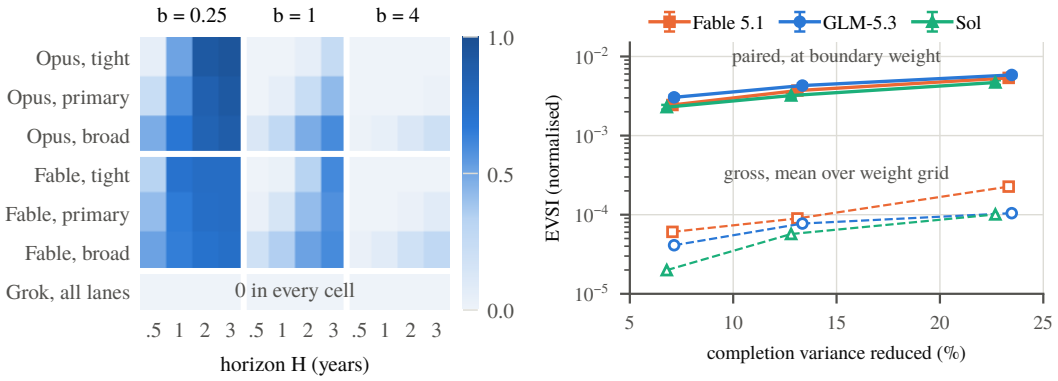


Figure 5: **Left: horizon and cost bridge decide whether switching is the only admissible action.** Its probability by pair and switching-cost lane (rows) and by bridge b and horizon H (columns), with workload 1, $\rho = 0$, a 7,200-s deadline and fallback unscored (Table 9). **Right: evidence is valuable only near decision boundaries.** Paired EVSI at the boundary weight (solid), with ± 2 MC SE bars drawn but hidden behind the markers as the SE is only 3.3–3.6% of each estimate, and grid-mean gross EVSI (dashed) against completion-variance reduction for 5-, 10- and 20-task panels.

actions only. With incumbent i, successor j, reported benchmark cost c, workload multiplier w and horizon H (years), the cost coordinates are

$$x_{\text{cost}}(\text{continue}) = -Hc_i w, \qquad x_{\text{cost}}(\text{switch}) = -\big(r\, b\, c_i w + Hc_j w\big), \tag{9}$$

where r is the one-off switching cost as a fraction of annual incumbent spend and b bridges benchmark cost to organisational spend, so that on cost alone switching is preferable exactly when $rb < H(1 - c_j/c_i)$, a condition from which w cancels. The prior for r is log-normal with mean 0.33 from firm-level IT-outsourcing estimates (Peukert, 2019), in tight, primary and broad lanes. We take $b \in \{0.25, 1, 4\}$ and $H \in \{0.5, 1, 2, 3\}$ and transport completion with a correlated logit shock, and 3,600 or 7,200-second deadlines define feasibility. These 864 scenarios form Λ.

Figure 5 maps the probability that switching is the *only* admissible action. It grows with the horizon and collapses as the cost bridge rises, so that for Fable 5→5.1 under the primary lane at one year it is 0.656, 0.146 and 0.001 at $b = 0.25, 1, 4$, and the Opus pair behaves alike. Grok 4.6, by contrast, costs 1.715 times as much as Grok 4.5, so improved completion never makes switching the only admissible action. The 3,600-second deadline leaves no feasible action for the Opus and Fable pairs, and the workload and fallback axes change nothing, as the specification guarantees (Section E). Within this scenario model the decision thus turns on the horizon and cost bridge more than on benchmark performance, a conditional result rather than evidence about organisations in general, and no scenario class yields a unique action for every draw.

5.6 Value of information concentrates at decision boundaries

Which further Terminal-Bench v4 evaluations could change a decision? For each of the twelve configurations and panels of 5, 10 or 20 further tasks (five trials each), we simulate outcomes from

the posterior predictive, update the posterior by importance reweighting (Section B) and recompute the scalarised decision at every weight cell. Gross EVSI is sharply localised (Figure 5, showing the three configurations with the largest maximum at 20 tasks, GLM-5.3, Fable 5.1 and GPT-5.6 Sol, all beside preference-region boundaries of Section 5.3): its maximum, 0.0056 normalised units (GLM-5.3, 20-task panel, weights $(0.7, 0.3, 0)$), occurs on a boundary between preference regions, where the probability that the evidence changes the action reaches 0.48, while the grid mean is far smaller. The same 20-task evidence reduces posterior completion variance by 23–24% whatever the weights, so uncertainty reduction is a poor proxy for decision value (grid mean and EVPI in Section E). At that boundary weight the paired estimator gives GLM-5.3 0.0058 (MC SE 0.0002), clearly above zero and with the three displayed configurations within a few standard errors of one another. For the lifecycle pairs of Section 5.5, by contrast, an estimator that fixed the no-evidence risk showed exactly the Monte Carlo drift the paired estimator removes (Section E). We do not run the paired estimator for those pairs and make no EVSI ranking claim there.

6 Discussion, limitations and conclusion

The case studies support four source-specific claims: frontiers are broad, linear scalarisation can discard very probable frontier members, within our declared scenario model admissibility turns on organisational inputs more than on benchmark performance and, more tentatively, sample information is valuable at decision boundaries, not where variance shrinks most. We make no organisational recommendation, propose no universal cost model, assume no scale shared across benchmarks and choose no scalar winner. Literature enters only through declared scenario lanes, EVSI is in normalised units and not money, no temporal model is fitted because public chronology identifies none (Section 5.4), effort and harness effects, like fallback and retire outcomes, remain unidentified and audits are self-reviewed (Section G).

Recommendations for benchmark builders. These limits mostly reflect what leaderboards publish, and we urge builders to adopt a reporting standard that removes each of our workarounds.

- **Trial-level outcomes.** Publish every task $\times$ trial outcome under task identifiers aligned across configurations, not only accuracy or pass@k (Hariri et al., 2026), since anonymous histograms recovered from pass@k hide task effects and cross-configuration covariance.
- **Complete aggregates.** Where only aggregates are released, state n, T and enough decimals that the published statistics round back to the exact counts, so that the histogram stays uniquely recoverable.
- **Per-trial cost and duration.** Report both with their dispersion, with token counts and the price schedule at run time, rather than one total, because Terminal-Bench v4 cost and duration entered as fixed values, a reported zero cost was in fact missing and forced an exclusion (Section 5.1), and re-pricing feeds the cost bridge that dominated the lifecycle analysis (Section 5.5).
- **The full configuration as identity.** Name the model snapshot, reasoning effort or thinking budget, harness and agent version, temperature and top-p, turn and token limits, timeouts, tool set, retries and seeds, and cross some models with several efforts and harnesses, because harness and effort change together across the Terminal-Bench v4 blocks, leaving neither effect identified (Section 5.1), and 0 of 102 τ^2 events carry a complete identity (Table 2).
- **Repeated jobs.** Run each configuration in several independent jobs over time, since one job confounds job, service and window effects (Section E) and leaves service change unidentified.
- **Clocks and task versions.** Record authenticated execution timestamps apart from release and reporting dates and version every task revision with the fraction it changed, because the temporal audit lacked such clocks (Table 2) and the regime bridge needs that fraction (Section 4).
- **Error flags.** Separate infrastructure errors from task failures in every trial record, which made the error sensitivity analysis of the native panel possible (Section 5.3).
- **Reconciled trajectories.** Release trajectories that reconcile with the headline statistics, since on τ^3-bench only one pair's did and our multidimensional evidence rests on it (Section 5.2).

ETHICS STATEMENT

This work analyses public benchmark records and published aggregate results from peer-reviewed and preprint studies and involves no human subjects, no new data collection from people and no personal data. The main ethical risk is misuse of the decision outputs, since posterior frontier probabilities or admissible action sets could be read as procurement advice for a particular organisation or as vendor rankings. We mitigate this risk by reporting set-valued outputs under declared scenario classes, by labelling every imported prior with its provenance and transport assumptions and by stating explicitly (Section 6) which claims the evidence does not support. Configuration names are those used by the public leaderboards, and no result should be read as a statement about a provider beyond the specific configuration and snapshot evaluated.

REPRODUCIBILITY STATEMENT

Section 4 and Section B give every likelihood and prior, Section A lists each data source with its pinned upstream revision and Table 4 reports sampler diagnostics under the standard of Section 4. All MCMC fits run on CPU in a pinned container image, with the chain count of each fit listed in Table 4 and fixed seeds recorded in per-run manifests.

AI USE STATEMENT

We used large language models for literature research to identify relevant prior publications, for assistance in data analysis (especially figure plotting) and for stylistic improvement of the text. The authors take full responsibility for all content.

REFERENCES

A. E. Ades, G. Lu, and K. Claxton. Expected value of sample information calculations in medical decision modeling. *Medical Decision Making*, 24(2):207–227, 2004. doi: 10.1177/0272989X04263162.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, Yiyang Nan, Jyoutir Raj, Yu Fan, Shubham Singh, Subramanyam Sahoo, Eliya Habba, Usman Gohar, Siddhesh Pawar, Robert Scholz, Arjun Subramonian, Jingwei Ni, Mykel Kochenderfer, Sanmi Koyejo, Mrinmaya Sachan, Stella Biderman, Zeerak Talat, Avijit Ghosh, and Irene Solaiman. When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv:2602.16763, 2026.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, 2024. doi: 10.18653/v1/2024.acl-long.744.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *Quarterly Journal of Economics*, 139(1):305–358, 2024. doi: 10.1093/qje/qjad043.

A. Bacchelli and C. Bird. Expectations, outcomes, and challenges of modern code review. In *Proceedings of the International Conference on Software Engineering (ICSE)*, pp. 712–721, 2013. doi: 10.1109/ICSE.2013.6606617.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. In *International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=OC2z7iSQKa.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma

Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott Hale, Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity. Technical report, METR, 2025. URL https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985. ISBN 0-387-96098-8.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434, 2017.

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. In Satyanshu K. Upadhyay, Umesh Singh, Dipak K. Dey, and Appaia Loganathan (eds.), *Current Trends in Bayesian Methodology with Applications*, pp. 79–101. CRC Press, Boca Raton, FL, 2015. doi: 10.1201/b18502-11.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. arXiv:2405.14782, 2024.

Mickaël Binois, David Ginsbourger, and Olivier Roustant. Quantifying uncertainty on Pareto fronts with Gaussian process conditional simulations. *European Journal of Operational Research*, 243 (2):386–394, 2015. doi: 10.1016/j.ejor.2014.07.032.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014, 2015. URL https://proceedings.mlr.press/v37/blum15.html.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021. doi: 10.18653/v1/2021.naacl-main.385.

Andrew H. Briggs, Milton C. Weinstein, Elisabeth A. L. Fenwick, Jonathan Karnon, Mark J. Sculpher, and A. David Paltiel. Model parameter estimation and uncertainty: A report of the ISPOR-SMDM modeling good research practices task force-6. *Value in Health*, 15(6):835–842, 2012. doi: 10.1016/j.jval.2012.04.014.

Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020. doi: 10.18653/v1/2020.emnlp-main.745.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

M.-H. Chen and J. G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000. doi: 10.1214/ss/1009212673.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10080–10098, 2025.

Alexander Conway, Debadeepta Dey, Stefan Hackmann, Matthew Hausknecht, Michael Douglas Schmidt, Mark Lewis Steadman, and Nick Volynets. syftr: Pareto-optimal generative AI. In *Proceedings of the Fourth International Conference on Automated Machine Learning*, volume 293 of *Proceedings of Machine Learning Research*, pp. 1–33, 2025. URL `https://proceedings.mlr.press/v293/conway25a.html`.

Zheyuan (Kevin) Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. *Management Science*, 2026. doi: 10.1287/mnsc.2025.00535.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi: 10.18653/v1/D19-1224.

Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. Large language model selection with limited annotations. arXiv:2605.24981, 2026.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pp. 2350–2358, 2015.

Matthias Ehrgott. *Multicriteria Optimization*. Springer, 2nd edition, 2005. ISBN 3-540-21398-8.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, 2020. doi: 10.18653/v1/2020.emnlp-main.393.

Guangrui Fan, Dandan Liu, Lihu Pan, and Rui Zhang. When help hurts: Verification load and fatigue with AI coding assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*, 2026. doi: 10.1145/3772318.3791176.

Joseph Farrell and Paul Klemperer. Coordination and lock-in: Competition with switching costs and network effects. In *Handbook of Industrial Organization*, volume 3, chapter 31, pp. 1967–2072. Elsevier, 2007. doi: 10.1016/S1573-448X(06)03031-7.

Shanshan Gao and Liyi Zhou. Can agent benchmarks support their scores? evidence-supported bounds for interactive-agent evaluation. arXiv:2605.10448, 2026.

Andrew Gelman, Xiao-Li Meng, and Hal S. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807, 1996.

Maggie Hamill and Katerina Goseva-Popstojanova. Analyzing and predicting effort associated with finding and fixing software faults. *Information and Software Technology*, 87:1–18, 2017. doi: 10.1016/j.infsof.2017.01.002.

Elise Han, Chengpiao Huang, and Kaizheng Wang. Model assessment and selection under temporal distribution shift. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 17374–17392, 2024.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't pass@k: A Bayesian framework for large language model evaluation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PTXi3Ef4sT.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23 (1):53–60, 1981. doi: 10.1137/1023004.

Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information. *Biometrics*, 67(3):1047–1056, 2011. doi: 10.1111/j.1541-0420.2011.01564.x.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Hui Huang, Xuanxin Wu, Muyun Yang, and Yuki Arase. Reasoning model is superior LLM-judge, yet suffers from biases. In *Proceedings of the Workshop on Evaluating Evaluations (EvalEval)*, 2026.

Christopher H. Jackson, Gianluca Baio, Anna Heath, Mark Strong, Nicky J. Welton, and Edward C. F. Wilson. Value of information analysis in models to inform health policy. *Annual Review of Statistics and Its Application*, 9:95–118, 2022. doi: 10.1146/annurev-statistics-040120-010730.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051, 2026. doi: 10.1038/s41586-026-10549-w.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=Zy4uFzMviZ.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *International Conference on Learning Representations (ICLR)*, 2026.

Andrew Keller, Kweku Kwegyir-Aggrey, Ryan Steed, Anita Rao, Julia Sharp, and Stevie Bergman. Expanding the AI evaluation toolbox with statistical models. Technical Report NIST AI 800-3, National Institute of Standards and Technology, 2026a.

Andrew Keller, Ryan Steed, Tony Wang, Stevie Bergman, and Peter Cihon. Practices for automated benchmark evaluations of language models. Technical Report NIST AI 800-2 (Initial Public Draft), National Institute of Standards and Technology, Center for AI Standards and Innovation, 2026b.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657, 2016. doi: 10.18653/v1/D16-1062.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 280–289, 2021. doi: 10.18653/v1/2021.acl-demo.34.

Xiaobin Liu, Thomas Tao Yang, and Yichong Zhang. Quasi-Bayesian inference for production frontiers. *Journal of Business & Economic Statistics*, 40(3):1334–1345, 2022. doi: 10.1080/07350015.2021.1927745.

Rachel Longjohn, Markelle Kelly, Sameer Singh, and Padhraic Smyth. Benchmark data repositories for better benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Jingkun Luo and Da-Tian Peng. Success is not self-explanatory: Auditing success provenance in agent evaluation. arXiv:2607.24054, 2026.

Nicolas A. Menzies. An efficient estimator for the expected value of sample information. *Medical Decision Making*, 36(3):308–320, 2016. doi: 10.1177/0272989X15583495.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Jenia Jitsev, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, Zizhao Chen, Yue Liu, Robert Zhang, Leon Liangyu Chen, Anurag Kashyap, Jan-Lucas Uslu, Jeffrey Li, Jianbo Wu, Minghao Yan, Song Bian, Vedang Sharma, Ke Sun, Steven Dillmann, Akshay Anand, Andrew Lanpouthakoun, Bardia Koopah, Changran Hu, Etash Guha, Gabriel H. S. Dreiman, Jiacheng Zhu, Karl Krauth, Li Zhong, Niklas Muennighoff, Robert Amanfu, Shangyin Tan, Shreyas Pimpalgaonkar, Tushar Aggarwal, Xiangning Lin, Xin Lan, Xuandong Zhao, Yiqing Liang, Yuanli Wang, Zilong Wang,

Changzhi Zhou, David Heineman, Hange Liu, Harsh Trivedi, John Yang, Junhong Lin, Manish Shetty, Michael Yang, Nabil Omi, Negin Raoof, Shanda Li, Terry Yue Zhuo, Wuwei Lin, Yiwei Dai, Yuxin Wang, Wenhao Chai, Shang Zhou, Dariush Wahdany, Ziyu She, Jiaming Hu, Zhikang Dong, Yuxuan Zhu, Sasha Cui, Ahson Saiyed, Arinbjörn Kolbeinsson, Jesse Hu, Christopher Michael Rytting, Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=a7Qa4CcHak.

Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, 1999. ISBN 0-7923-8278-1.

Mahdi Naser Moghadasi and Faezeh Ghaderi. What twelve LLM agent benchmark papers disclose about themselves: A pilot audit and an open scoring schema. arXiv:2605.21404, 2026.

Marcio Monteiro, Weichen Li, Puyu Wang, Marius Kloft, and Sophie Fellenz. Landmark-guided policy optimization for multi-objective language model selection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62547.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. arXiv:2607.16259, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

OpenAI. Introducing SWE-bench verified, 2024. URL https://openai.com/index/introducing-swe-bench-verified/.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13:6793, 2022. doi: 10.1038/s41467-022-34591-0.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv:2302.06590, 2023.

Christian Peukert. Determinants and heterogeneity of switching costs in IT outsourcing: Estimates from firm-level data. *European Journal of Information Systems*, 28(3):291–317, 2019. doi: 10.1080/0960085X.2018.1529374.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA, 1961.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, 2021. doi: 10.18653/v1/2021.acl-long.346.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 9175–9185, 2019.

Donald B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. doi: 10.1214/aos/1176345338.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014. doi: 10.1111/biom.12242.

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. τ-Knowledge: Evaluating conversational agents over unstructured knowledge. arXiv:2603.04370, 2026.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=4Ae8edNqm0.

J. G. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, 10(2):257–261, 1948. doi: 10.1111/j.2517-6161.1948.tb00014.x.

A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2):84–88, 1992. doi: 10.1080/00031305.1992.10475856.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Ben Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in AI benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1141–1158, 2021. doi: 10.18653/v1/2021.acl-long.92.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. doi: 10.1007/s11222-016-9696-4.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221.

Lingli Wang, Ni Huang, Yili Hong, Luning Liu, Xunhua Guo, and Guoqing Chen. Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management*, 32:1002–1018, 2023. doi: 10.1111/poms.13953.

Yiwei Wang, Chuan Zhu, Tianjun Feng, Lauren Xiaoyuan Lu, and Bingxin Jia. Agentic AI and human-in-the-loop interventions: Field experimental evidence from Alibaba's customer service operations. arXiv:2605.14830, 2026. arXiv preprint, version 2.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=cReExMQLiK.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. In *International Conference on Learning Representations (ICLR)*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E. Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang Chen, Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, and José Hernández-Orallo. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026. doi: 10.1038/s41586-026-10303-2.

Tijana Zrnic and William Fithian. A flexible defense against the winner's curse. *The Annals of Statistics*, 53(6):2516–2535, 2025. doi: 10.1214/25-AOS2553.

A Data provenance

Table 3 lists every source together with its pinned upstream revision, its unit of observation and the role it plays in the analysis. Result bundles are identified by content hashes in the supplementary material, and the slot registry in `values.tex` of the manuscript source maps each reported number back to the bundle from which it comes.

Table 3: Data sources. "Role" gives the section in which the source is used, and no two sources are pooled.

Source	Pin	Unit of observation	Role
Terminal-Bench v4 (v4.0.0) (Merrill et al., 2026)	`harbor-framework/terminal-bench` @ `8310ca48`, digest `39d9f44b...`	13 configurations × 66 tasks × 5 trials, with totals and pass@2–5 (anonymous histograms)	Sections 5.1, 5.3, 5.5 and 5.6
Terminal-Bench v4 native trials	4 selected jobs, input SHA-256 `e5f90e4d...`	4 configurations × 66 tasks × 5 repeats = 1,320 trial rows (reward, cost, duration)	Section 5.3
τ^3-bench v1.0.1, text (Barres et al., 2026; Shi et al., 2026)	`sierra-research/tau2-bench` v1.0.1 @ `fc0055dc`	one reconciled pair in 12 domain × complexity cells (278 tasks × 4 trials)	Section 5.2
LiveBench (White et al., 2025)	adjacent public releases	10 release transitions with common and documented-unaffected task deltas	Section 5.4
LiveCodeBench (Jain et al., 2025)	`livecodebench.github.io` @ `f186ab50`	model × problem cells, pinned historical matrix	Section 5.4
SWE-bench Verified (Jimenez et al., 2024; OpenAI, 2024)	`SWE-bench/experiments` @ `40f164d5`	5 dated entries, 2,017 task rows (166 common tasks)	Section 5.4
HELM history (Liang et al., 2023)	public snapshots	16 dated snapshots	Section 3.1
τ-bench family events (Yao et al., 2025; Barres et al., 2026)	`sierra-research/tau2-bench` v1.0.0 @ `17e07b1d`, v1.0.1 @ `fc0055dc`	event- and trial-level traces, with text and voice strata kept separate	Section 5.4
Literature priors	see Table 8	published estimates and counts	Section 5.5

B Model specifications

Terminal-Bench v4 completion (M0b/M3). For configuration c with histogram $\boldsymbol{h}_c$, the likelihood is $\prod_{j=0}^{5} \mathrm{BetaBinomial}(j \mid 5, p_c\kappa_c, (1-p_c)\kappa_c)^{h_{cj}}$ with $p_c \sim \mathrm{Beta}(1,1)$ and $\kappa_c \sim \mathrm{LogNormal}(\log 5, 1)$, and prior sensitivity is assessed with κ medians of 2 and 20. The baseline replaces the Beta-Binomial with $\mathrm{Binomial}(5, p_c)$. Posterior predictive checks compare replicated and observed histogram cells, pass@k profiles and task-count variances, and the frontier vectors are $(p_c, -\mathrm{cost}_c, -\mathrm{duration}_c)$, with reported cost and duration held fixed because their task-level distributions are unpublished, and with a reported zero cost treated as missing.

τ^3-bench (M1/M3). Cell-level differences $\hat{\delta}_{id}$ in (quality in pp, $-$cost in USD, $-$duration in s) are rescaled by coordinate-specific constants, and their standard errors are set to $s_{id} = (\text{hi} - \text{lo})/3.92$, where hi and lo are the endpoints of 95% task-cluster percentile bootstrap intervals that our own pipeline computed from the reconciled task rows, separately for each coordinate. The Gaussian summary likelihood is therefore an approximation, although the task-level standard errors lie between 0.96 and 1.15 times s_{id}. Because each coordinate is resampled on its own, the likelihood also omits the within-cell sampling correlations between coordinates, which the task rows put, for the unoriented differences, at a median of -0.54 (range -0.72 to 0.38) for cost and duration, -0.08 (-0.71 to 0.25) for quality and cost and 0.01 (-0.79 to 0.44) for quality and duration. The diagonal likelihood can thus attribute part of this sampling dependence to workload covariance, which is one more reason why we treat the factor dependence as exploratory. The coordinatewise random-effects model is $\hat{\delta}_{id} \sim \mathcal{N}(\mu_d, s_{id}^2 + \tau_d^2)$, to which the factor model of equation 8 adds the loadings $\boldsymbol{\lambda}$. Its likelihood is evaluated with the matrix determinant lemma and the Woodbury identity (Henderson & Searle, 1981), a positive cost loading fixes the sign, and new-workload draws $\boldsymbol{\delta}_{\text{new}} \sim \mathcal{N}(\boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\tau}^2) + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top)$ give the dominance and tradeoff probabilities for an unseen workload. The priors are $\mu_d \sim \mathcal{N}(0, 3^2)$ on the rescaled coordinates and $\tau_d \sim \text{HalfNormal}(\sigma_\tau)$, with $\sigma_\tau \in \{0.5, 1, 2\}$ for sensitivity. In the factor model, $\lambda_{\text{cost}} \sim \text{HalfNormal}(1)$ and the remaining loadings $\sim \mathcal{N}(0, 1)$. The baseline sets $\tau = 0$, and the holdouts remove either the highest complexity stratum or one domain at a time.

Measurement-regime bridge (M2). Given observed deltas $(y_t^{\text{un}}, y_t^{\text{all}})$, standard errors $(s_t^{\text{un}}, s_t^{\text{all}})$ and changed fraction π_t, integrating out $c_t \sim \mathcal{N}(\mu_c, \sigma_c^2)$ and $r_t \sim \mathcal{N}(\mu_r, \sigma_r^2)$ yields a bivariate normal with variances $\sigma_c^2 + (s_t^{\text{un}})^2$ and $\sigma_c^2 + \pi_t^2\sigma_r^2 + (s_t^{\text{all}})^2$ and covariance σ_c^2. The priors, in pp, are $\mu_c \sim \mathcal{N}(0, 2^2)$, $\sigma_c \sim \text{HalfNormal}(3)$, $\mu_r \sim \mathcal{N}(0, 10^2)$, $\sigma_r \sim \text{HalfNormal}(15)$. The baseline fixes $r_t = 0$, and r_t is unidentified when $\pi_t = 0$.

Nested Bayesian bootstrap (M3.5 native panel). Task weights $\boldsymbol{w} \sim \text{Dirichlet}(\mathbf{1}_T)$ are shared across configurations and coordinates, and repeat weights $\boldsymbol{v}_{ct} \sim \text{Dirichlet}(\mathbf{1}_R)$ are shared across coordinates within a cell, so that coordinate k of configuration c is $F_{ck} = \sum_t w_t \sum_r v_{ctr} y_{ctrk}$, an established construction (Rubin, 1981) that we run with 32,000 independent draws and validate by exact simplex and aggregation checks. Under the primary rule quality treats missing rewards as failures, with a scored-only rule as a sensitivity analysis, whereas cost and duration use observed masks. Normalisation for the weight grid uses fixed bounds, namely the minimum and maximum across configurations of the equal-task empirical means, so that the grid does not move with the draw.

Lifecycle scenarios (M4b). For each posterior draw of (p_i, p_j) and scenario λ, completion is transported as $\operatorname{logit} \tilde{p} = \operatorname{logit} p + \eta$ with $(\eta_i, \eta_j) = 0.5\,(z_i, z_j)$, where z_i and z_j are standard normal with correlation $\rho \in \{-0.25, 0, 0.25\}$, so that each shock has standard deviation 0.5 on the logit scale. Costs follow equation 9 with $\log r \sim \mathcal{N}(\log 0.33 - \sigma_r^2/2, \sigma_r^2)$ and $\sigma_r \in \{0.4, 0.7, 1.2\}$ (tight, primary, broad), which fixes the mean of r at 0.33 in every lane while its median falls from 0.31 to 0.26 and 0.16. An action is feasible if its duration meets the deadline, and fallback, although feasible, is unidentified and therefore never enters as a scored alternative. The admissible set is then the non-dominated set of feasible action vectors in each draw.

Value of information (M5a, M5). For a candidate panel of K additional tasks for configuration c, we simulate predictive outcomes $y \sim p(y \mid D)$, reweight the retained posterior draws by the likelihood of the simulated outcomes (Menzies, 2016), the weighting step of sampling importance resampling (Smith & Gelfand, 1992), with weights $w_d(y) \propto \prod_{i=1}^{K} \Pr(Y_i = y_i \mid p_{c,d}, \kappa_{c,d})$, and recompute the decision. The paired estimator is $\widehat{\text{EVSI}} = R^{-1} \sum_{r=1}^{R} \big[\hat{\mathbb{E}}\{L(a_0) \mid D, y_r\} - \min_a \hat{\mathbb{E}}\{L(a) \mid D, y_r\}\big]$, whose Monte Carlo standard error is the replicate standard deviation divided by $\sqrt{R}$. For Terminal-Bench v4 (M5a) we use $N = 16{,}000$ retained posterior draws and $R = 2{,}000$ replicates for each configuration and panel size. Replaying these replicates with their original seeds, which reproduces the paired estimates exactly, shows that the effective sample size $1/\sum_d w_d(y)^2$ of the normalised weights is at least 534 in every replicate for the three displayed configurations, with 1% quantiles of at least 2,953 and medians of at least 11,451 (at least 90 across all twelve configurations).

Diagnostic standard. All MCMC fits use NUTS (Hoffman & Gelman, 2014) on CPU. A fit is promoted only with 8 chains, rank-normalised $\hat{R} < 1.01$, bulk and tail ESS ≥ 100 per chain (Vehtari et al., 2021) and zero divergences, and only after its energy (BFMI) (Betancourt, 2017), trace and rank plots have been inspected. Scientific adequacy is judged separately, by posterior predictive checks (Gelman et al., 1996), held-out LPD (Vehtari et al., 2017) and prior sensitivity, and Table 4 lists the configuration of each fit.

C DIAGNOSTICS

Table 4: Sampler and adequacy diagnostics for every fit reported in Section 5. The production standard is 8 chains with $\hat{R} < 1.01$, bulk and tail ESS ≥ 800 and zero divergences. The Terminal-Bench v4 (M0b and M3), M4b, production τ^3-bench and M3.5 DVFI fits are 8-chain runs that meet it, whereas the others, the τ^3-bench stress test among them, are 2- and 4-chain diagnostic runs that do not.

Fit	Runtime	Chains×(warm+draw)	$\hat{R}$	Bulk ESS	Tail ESS	Div.	BFMI	Adequacy
Terminal-Bench v4 Beta-Binomial (M0b)	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.00	24,000	11,000	0	0.931	PPC pass
Terminal-Bench v4 Binomial baseline	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.00	19,000	12,000	0	1.001	PPC fail
Terminal-Bench v4 frontier (M3), κ medians 2/5/20	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.00	24,000	11,000	0	0.931	meets standard
τ^3 random effects (M1, 3 prior scales)	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.002	6,388	3,551	0	0.861	PPC, holdout pass
τ^3 augmented factor (stress test)	PyMC	$4 \times (5\text{k} + 5\text{k})$	1.032	179	679	188	0.354	sampler fail
τ^3 exact-marginal factor / joint frontier (M3)	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.002	3,699	3,068	0	0.763	meets standard
τ^3 holdout refits (12 fits)	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.002	3,520	2,926	0	0.741	meets standard
LiveBench bridge (M2)	PyMC	$4 \times (2\text{k} + 2\text{k})$	1.002	2,477	1,977	0	0.892	inconclusive
LiveBench dynamic state (M2b)	PyMC	$2 \times (2\text{k} + 2\text{k})$	1.003	729	–	0	0.777	not earned
LiveCodeBench R0 / R1-T	JAX HMC	$2 \times (2\text{k} + 2\text{k})$	1.022 / 1.010	250 / 239	390 / 432	0	0.669 / 0.719	PPC fail
LiveCodeBench R1-S / R1-E	JAX HMC	$2 \times (2\text{k} + 2\text{k})$	1.012 / 1.037	276 / 113	421 / 164	0	0.737 / 0.808	PPC fail
Terminal-Bench v4 for M3.5 (DVFI)	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.00	26,000	11,000	0	0.973	meets standard
Terminal-Bench v4 for M4b	PyMC	$8 \times (2\text{k} + 2\text{k})$	1.000	24,000	11,000	0	0.902	meets standard
Native panel bootstrap (v8)	NumPy	32,000 iid draws	n/a (exact simplex checks pass)					exploratory

Table 5: Terminal-Bench v4 static frontier: harness block, posterior completion (mean and 95% interval), reported cost, mean duration and draw-wise Pareto-membership probability for the 12 configurations with valid cost.

Configuration	Block	Completion % [95% CI]	Cost (USD)	Duration (s)	P(Pareto)
GPT-5.6 Sol	Codex/max	36.1 [27.9, 44.7]	2,542	2,388	1.000
GPT-5.6 Luna	Codex/max	17.4 [11.1, 24.9]	347	4,088	1.000
GPT-5.6 Terra	Codex/max	21.9 [15.2, 29.7]	1,734	2,515	0.985
Opus 5	Claude Code/max	50.7 [41.2, 60.1]	5,969	4,793	0.987
Fable 5	Claude Code/max	43.8 [35.2, 52.9]	7,265	4,203	0.025
Sonnet 5	Claude Code/max	12.8 [8.3, 18.3]	9,604	6,510	0.000
Opus 4.8	Claude Code/max	23.2 [16.6, 30.7]	6,481	5,163	0.000
Grok 4.6	Grok Build/none	20.1 [13.6, 27.8]	3,592	2,352	0.964
Grok 4.5	Grok Build/none	12.7 [7.6, 19.3]	2,094	3,162	0.025
Gemini 3.7 Flash	mini-SWE-agent/high	11.7 [6.9, 18.0]	1,262	1,689	1.000
Fable 5.1	singleton	56.7 [47.6, 65.5]	6,244	3,894	0.999
GLM-5.3	singleton	40.7 [31.2, 50.4]	2,728	5,831	0.761

Terminal-Bench v4 static frontier, further detail. The four exact harness, version and effort blocks are Codex/max with three GPT-5.6 variants, Claude Code/max with four Claude models, Grok Build/none with two Grok models and mini-SWE-agent/high with two Gemini Flash models, and the two singletons are Fable 5.1 and GLM-5.3 (Table 5). The Binomial baseline covers only 24.2% (22/91) of the audited histogram statistics and 0% of the task-count variances, the Beta-Binomial model 100% of both under every tested prior (Figure 6). The Beta-Binomial fits imply intraclass correlations from 0.24 (Sonnet 5) to 0.61 (GLM-5.3), and posterior completion runs from 11.7% [6.9, 18.0] for Gemini 3.7 Flash to 56.7% [47.6, 65.5] for Fable 5.1, with Opus 5 at 50.7% [41.2, 60.1] (all configurations in Table 5). The almost surely dominated configurations are Fable 5 (0.025), Grok 4.5 (0.025), Sonnet 5 (0) and Opus 4.8 (0), and the members by construction are GPT-5.6 Luna, lowest cost, and Gemini 3.7 Flash, lowest duration, so the informative probabilities are those of the others, such as Opus 5 at 0.987, GPT-5.6 Terra at 0.985, Grok 4.6 at 0.964 and GLM-5.3 at 0.761. A 5,000-resample task-level bootstrap reproduces these probabilities to within 0.023, and the κ-prior sweep moves them by at most 0.009. Within blocks, GPT-5.6 Sol exceeds Terra by $+14.2$ pp [2.9, 25.3], $P(> 0) = 0.993$, while Fable 5 against Opus 5 is unresolved at -6.9 pp [-19.8, 6.2], and Table 6 lists every within-harness contrast.

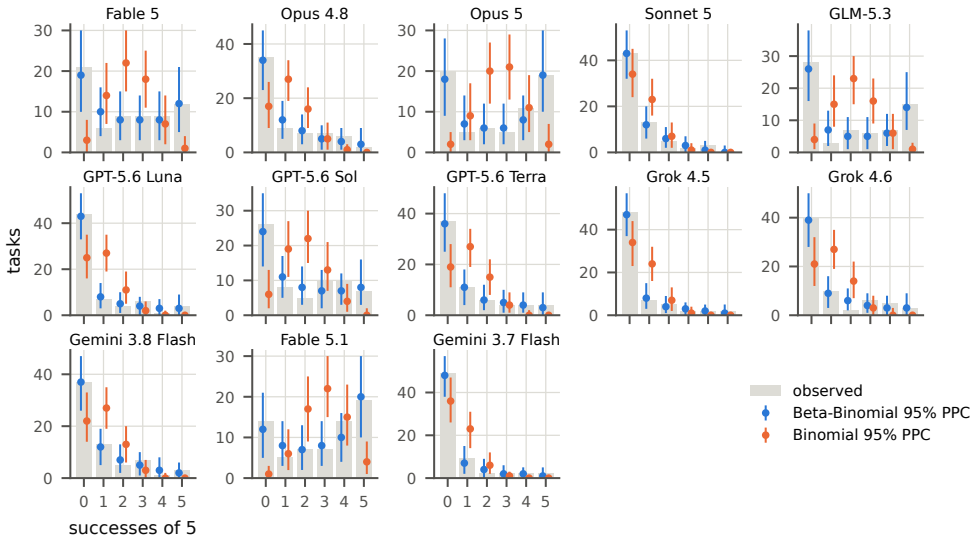


Figure 6: Terminal-Bench v4 posterior predictive check: observed task-success histograms (bars, 66 tasks with five trials each) against 95% posterior predictive intervals under the Beta-Binomial (blue) and Binomial (orange) models, 1,000 draws each. The Binomial model misses the U-shaped mass at 0 and 5 successes.

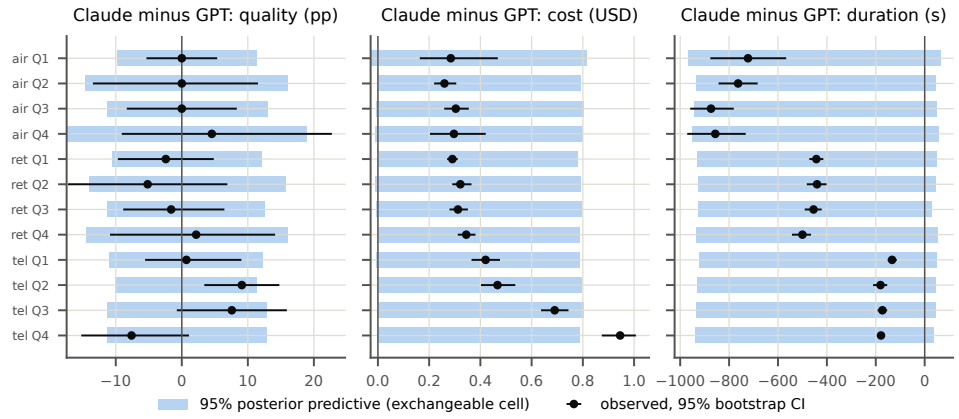


Figure 7: τ^3-bench per-cell differences (Claude Opus 4.5 minus GPT-5.2, 12 domain $\times$ complexity cells) with 95% bootstrap intervals, against the 95% posterior predictive interval for an exchangeable cell under the 8-chain random-effects model.

Dependence across configurations. The M3 posterior factorises over configurations, each having its own prior and histogram likelihood, so the membership probabilities of Section 5.1 couple the completion probabilities p_c independently. The anonymous histograms carry no task pairing that could identify another coupling, although membership still depends on it. All configurations attempt the same 66 tasks, so a sample that happens to be hard is hard for all of them, and a model with task effects shared across configurations would make the posterior uncertainty about the p_c positively dependent, which we therefore take to be the plausible direction. With cost and duration fixed, b is a member exactly when p_b exceeds p_a for every $a \in D_b$, the $|D_b|$ configurations with neither higher cost nor longer duration, so no other configuration bears on b. Table 7 re-couples the 16,000 draws of each configuration with an exchangeable Gaussian copula, reordering them by the ranks of equicorrelated normal scores so that every marginal is unchanged, for ρ from -0.09, close to the most negative value $-1/11$ open to 12 exchangeable configurations, to the comonotone $\rho = 1$. Positive dependence moves each probability towards 0 or 1, and 8 of 12 configurations

Table 6: Within-harness Terminal-Bench v4 completion contrasts (Figure 4), each oriented so that its posterior mean is positive: posterior mean, 95% interval and $P(>0)$ under the central κ prior.

Block	Contrast	Mean (pp)	95% interval	$P(>0)$
Codex/max	GPT-5.6 Sol − GPT-5.6 Luna	$+18.7$	$[7.4, 29.7]$	>0.999
Codex/max	GPT-5.6 Sol − GPT-5.6 Terra	$+14.2$	$[2.9, 25.3]$	0.993
Codex/max	GPT-5.6 Terra − GPT-5.6 Luna	$+4.5$	$[-5.7, 14.5]$	0.807
Claude Code/max	Opus 5 − Sonnet 5	$+37.9$	$[26.9, 48.6]$	>0.999
Claude Code/max	Fable 5 − Sonnet 5	$+31.0$	$[20.6, 41.2]$	>0.999
Claude Code/max	Opus 5 − Opus 4.8	$+27.6$	$[15.5, 39.1]$	>0.999
Claude Code/max	Fable 5 − Opus 4.8	$+20.7$	$[9.4, 31.9]$	>0.999
Claude Code/max	Opus 4.8 − Sonnet 5	$+10.4$	$[1.8, 19.2]$	0.992
Claude Code/max	Opus 5 − Fable 5	$+6.9$	$[-6.2, 19.8]$	0.853
Grok Build/none	Grok 4.6 − Grok 4.5	$+7.4$	$[-2.0, 16.8]$	0.942
mini-SWE-agent/high	Gemini 3.8 Flash − Gemini 3.7 Flash	$+7.9$	$[-0.7, 16.4]$	0.965

have membership probability above 0.7 at every ρ. Writing H_a for the distribution function of a's posterior draws of p_a and $H_a(t^-)$ for its left limit, no coupling of these marginals can push b's membership probability outside $[L_b, U_b]$ with $L_b = \sup_t \left(\sum_{a \in D_b} H_a(t) - (|D_b| - 1) - H_b(t) \right)^+$ and $U_b = 1 - \sup_t \left(H_b(t) - \min_{a \in D_b} H_a(t^-) \right)^+$, bounds that the empirical law of the draws attains and which keep seven configurations at or above 0.80 and leave GLM-5.3 anywhere in [0.38, 1]. Fable 5 and Grok 4.5 cannot exceed 0.16 and 0.17 respectively, and Opus 4.8 and Sonnet 5 are excluded under every coupling of the draws, since none of their draws reaches the lowest draw of Fable 5.1, which is cheaper and faster than both. Monte Carlo standard errors are at most 0.004 for the probabilities, whose largest standard deviation over 20 further reorderings is 0.002, and 0.006 for the bounds, by batch means over the eight chains.

Table 7: Coupling sensitivity of Terminal-Bench v4 Pareto membership, rows as in Figure 3. $|D_b|$ counts the configurations with neither higher cost nor longer duration, "Reported" is the membership probability of Figure 3 (independent coupling, as sampled), the next five columns re-couple the draws by an exchangeable Gaussian copula with correlation ρ ($\rho = 0$ being a fresh independent coupling) and the last two give the bounds each configuration's membership probability can take under any coupling of the draws with the same marginals. Each entry is computed from 16,000 posterior draws.

			Gaussian copula, ρ					All couplings	
Configuration	$\lvert D_b \rvert$	Reported	-0.09	0	0.5	0.9	1	Lower	Upper
GPT-5.6 Sol	1	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000
GPT-5.6 Luna	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Gemini 3.7 Flash	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Fable 5.1	5	0.999	0.999	0.999	1.000	1.000	1.000	0.977	1.000
Opus 5	6	0.987	0.983	0.987	0.999	1.000	1.000	0.889	1.000
GPT-5.6 Terra	1	0.985	0.981	0.985	0.999	1.000	1.000	0.883	1.000
Grok 4.6	1	0.964	0.961	0.967	0.995	1.000	1.000	0.809	1.000
GLM-5.3	5	0.761	0.750	0.761	0.838	0.986	1.000	0.382	1.000
Fable 5	7	0.025	0.027	0.023	0.002	0.000	0.000	0.000	0.162
Grok 4.5	2	0.025	0.031	0.026	0.003	0.000	0.000	0.000	0.167
Opus 4.8	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Sonnet 5	11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

τ^3-bench tradeoff, further detail. Beyond the quality difference, the posterior mean differences (Claude minus GPT) are $-\$0.40$ $[-0.50, -0.29]$ in negative cost and $+446$ s [313, 577] in negative duration. The mean-vector tradeoff probability is 1.0, and for a new workload Claude and GPT dominate with probabilities 0.012 and 0.012, the tradeoff probability being 0.973 under independent random effects. Figure 7 shows the differences cell by cell: the domains separate clearly on duration and telecom stands apart on cost, while quality differences scatter around zero. Held-out LPD on the hardest complexity stratum is -20.8 with heterogeneity versus -322.0 without and on the telecom domain -66.3 versus $-2{,}333.8$ (Figure 4). The exact-marginal factor model removes the sampler failures of the augmented representation (188 versus 0 divergences at $4 \times (5{,}000 + 5{,}000)$ iterations), improves held-out LPD over independent random effects by $+15.96$ summed over the four holdouts ($+16.71$ on telecom, within ± 0.9 elsewhere) and implies a cost–duration correlation with posterior

mean 0.45, $P(\rho > 0) = 0.96$, 95% interval $[-0.02, 0.79]$. Rescaling the heterogeneity prior by 0.5 or 2 shifts the duration difference by at most 7.8 s.

D PRIOR-SOURCE LEDGER

Table 8 lists every literature source considered for the lifecycle layer, together with the role each is allowed to play. Of these, only the switching-cost anchor enters the Terminal-Bench v4 action scenarios in Section 5.5, whereas the others define separate source-scale lanes or structural constraints.

Table 8: Prior-source ledger (abridged). "Role here" is the most that each source is allowed to support.

Quantity	Source	Reported estimand	Role here
Switching cost / annual spend	Peukert (2019)	IT-outsourcing switching cost, mean about one third of annual professional-services spend (heterogeneity only)	scenario anchor (log-SD lanes 0.4/0.7/1.2)
Developer output	Cui et al. (2026)	pooled completed-task effect of AI assistance across three field experiments	source lane with transport sensitivity
Developer task time	Fan et al. (2026)	AI/control task-time ratio in a lab study of Python tasks	conditional time sensitivity without a review-hours bridge
Developer task time	Becker et al. (2025)	log-time effect for experienced open-source developers	direction only
Developer task time	Peng et al. (2023)	single-task completion-time ratio	structural contrast only
Support productivity	Brynjolfsson et al. (2025)	resolutions per hour, handle time, resolution rate, manager requests	source lane (manager request is not fallback)
Escalation mix	Wang et al. (2026)	category composition and post-escalation outcomes in matched AI-eligible chats	source lane (selected subset)
Telecom service	Wang et al. (2023)	duration, human demand, complaints	structural only
Correction burden	Hamill & Goseva-Popstojanova (2017)	concentration of fix effort	structural only
Review structure	Bacchelli & Bird (2013)	organisational review mechanisms	structural only
Airline outcomes	none	no commensurable estimate	constraint lane only

E ADDITIONAL RESULTS

Table 9: Probability that switching is the only admissible action for every horizon, switching-cost lane and cost bridge (workload 1, $\rho = 0$, 7,200-second deadline, fallback unscored). The Grok pair is 0 throughout.

Pair	Lane	H	$b = 0.25$	$b = 1$	$b = 4$
Opus 4.8→5	tight	0.5	0.045	0.000	0.000
Opus 4.8→5	tight	1	0.506	0.001	0.000
Opus 4.8→5	tight	2	0.916	0.045	0.000
Opus 4.8→5	tight	3	0.946	0.246	0.000
Opus 4.8→5	primary	0.5	0.228	0.004	0.000
Opus 4.8→5	primary	1	0.584	0.044	0.000
Opus 4.8→5	primary	2	0.852	0.228	0.004
Opus 4.8→5	primary	3	0.915	0.425	0.018
Opus 4.8→5	broad	0.5	0.474	0.117	0.010
Opus 4.8→5	broad	1	0.677	0.271	0.042
Opus 4.8→5	broad	2	0.828	0.474	0.117
Opus 4.8→5	broad	3	0.883	0.596	0.198
Fable 5→5.1	tight	0.5	0.320	0.000	0.000
Fable 5→5.1	tight	1	0.709	0.022	0.000
Fable 5→5.1	tight	2	0.753	0.320	0.000
Fable 5→5.1	tight	3	0.754	0.602	0.004
Fable 5→5.1	primary	0.5	0.417	0.024	0.000
Fable 5→5.1	primary	1	0.656	0.146	0.001
Fable 5→5.1	primary	2	0.741	0.417	0.024
Fable 5→5.1	primary	3	0.751	0.576	0.077
Fable 5→5.1	broad	0.5	0.511	0.186	0.024
Fable 5→5.1	broad	1	0.644	0.347	0.078
Fable 5→5.1	broad	2	0.713	0.511	0.186
Fable 5→5.1	broad	3	0.735	0.597	0.277

Lifecycle grid, further detail. For Fable 5→5.1 the probability that switching is the only admissible action ranges from 0.0002 (tight lane, $H = 0.5$) to 0.60 (tight lane, $H = 3$) across horizons and lanes at bridge $b = 1$ and equals 0.146 under the primary lane at one year. For Opus 4.8→5 it is between 0 and 0.60 at $b = 1$ (0.044 in the primary lane at one year), and for Grok 4.5→4.6 it is 0 in every scenario, with both actions always admissible. Under the 3,600-second deadline the feasibility probability of both actions for the Opus and Fable pairs is 0, and across the workload and fallback axes the maximum change is exactly 0. Literature lanes on developer task time (Fan et al., 2026) and customer-service productivity (Brynjolfsson et al., 2025; Wang et al., 2026) are not joined to these actions, because their bridge to benchmark actions is unresolved, and the developer task-time lane has an AI/control time ratio of 0.78 [0.74, 0.83].

Static lifecycle-pair EVSI (raw estimator). The first M5 implementation for the three lifecycle pairs reused 4,096 M4b draws and evaluated 5-, 10- and 20-task successor panels across 18 transport contexts, four normalised loss classes, three constraint classes and 64 preposterior replicates. Its largest raw EVSI was 0.0029 (Grok pair, 10-task panel) with 0 of 64 simulated panels changing the action. In 28 decision contexts the estimated EVSI exceeded EVPI, which is impossible for the estimand. At zero acquisition cost, 164 of the 312 finite raw estimates (53%) were negative before clipping, and 336 of 648 contexts had no admissible action. The largest action-change probability was 15/64 = 0.23 (MC SE 0.05), and the Pareto support set changed in 0 of 54 checks. These figures are diagnostics of Monte Carlo drift rather than measures of decision value.

Terminal-Bench v4 value of information, further detail. The grid-mean gross EVSI for 20-task panels is about 10^{-4}, against a maximum of 0.0056 normalised units (GLM-5.3, 20-task panel, weights $(0.7, 0.3, 0)$), and the grid-mean EVPI is 0.0025 (maximum 0.023) (Figure 5). The three configurations displayed in Section 5.6 were not fixed in advance. EVSI was computed for all twelve configurations and panel sizes, and the three with the largest maximum gross EVSI at 20 tasks were selected afterwards, well clear of the fourth largest, 0.0002 (GPT-5.6 Luna). At the same 20-task boundary weight, the paired estimator gives GLM-5.3 0.00584 (MC SE 0.00020), Fable 5.1 0.00536 (MC SE 0.00018) and GPT-5.6 Sol 0.00469 (MC SE 0.00016), with $P(\text{action change})$ within 0.010 of 0.48 for all three, each with Monte Carlo SE 0.011 (Table 10 shows the corresponding raw estimator for the lifecycle pairs, where the paired estimator has not been run).

Table 10: Raw static EVSI estimator for the lifecycle pairs at zero acquisition cost, by pair and panel size: decision contexts, contexts with a finite raw estimate, negative raw estimates before clipping, maximum raw EVSI, contexts with EVSI above EVPI and the largest number of 64 simulated panels that changed the action. A Monte Carlo drift diagnostic, not decision value.

Pair	Tasks	Contexts	Finite	Negative	Max raw EVSI	EVSI>EVPI	Max action change
Opus 4.8→Opus 5	5	72	40	14	8.7e-05	0	0/64
Opus 4.8→Opus 5	10	72	40	38	7.9e-06	0	0/64
Opus 4.8→Opus 5	20	72	40	1	2.1e-03	2	0/64
Fable 5→Fable 5.1	5	72	40	28	5.6e-05	0	10/64
Fable 5→Fable 5.1	10	72	40	36	1.8e-04	0	15/64
Fable 5→Fable 5.1	20	72	40	22	5.4e-04	0	14/64
Grok 4.5→Grok 4.6	5	72	24	24	-1.7e-05	0	0/64
Grok 4.5→Grok 4.6	10	72	24	0	2.9e-03	13	0/64
Grok 4.5→Grok 4.6	20	72	24	1	3.9e-05	13	0/64

Table 11: Native Terminal-Bench v4 held-out task-fold diagnostic (balanced stylised profile, five folds over the same four jobs). "Set-valued rate" is the fraction of folds in which the Pareto rule returns a set rather than one action. Returning a set is not claimed to reduce loss.

Policy	Mean normalised held-out loss	Set-valued rate
Quality only	1.000	0
Cost only	0.800	0
Duration only	0.600	0
Equal weights	0.400	0
Pareto restriction	0.417	1

Scalarisation, further detail. On the 66-cell Terminal-Bench v4 weight grid the preference regions are split among Gemini 3.7 Flash (0.45), GPT-5.6 Sol (0.21), GPT-5.6 Luna (0.17), Fable 5.1

(0.14) and GLM-5.3 (0.03) (Figure 3), computed from the posterior means with cost and duration at their reported values. At these means Opus 5, GPT-5.6 Terra and Grok 4.6 still win none of the 501,501 weights of a grid with step 0.001. Per posterior draw their membership probabilities of 0.987, 0.985 and 0.964 compare with probabilities of 0.227, 0.210 and 0.004 of winning at least one of the 5,151 weights of the grid with step 0.01, so that each is a supported point in only a minority of the draws in which it is a frontier member, and Grok 4.6 in almost none of them. In the native four-configuration panel, three configurations are near-certain members (Terra 0.998, Sol > 0.999, Opus 5 0.990; Sonnet 5 0), yet the weight grid gives them 24, 36 and 6 of the 66 cells (Sonnet 5: 0), and excluding the error-flagged trials gives membership probabilities of 0.9995, 0.9999, 0.993 and 0. With one job per configuration, job, service and execution-window effects are confounded in this panel, and its probabilities are conditional on the observed jobs.

Reported cost and duration. The public summaries give each configuration only a total cost and a mean trial duration, so Table 12 multiplies both by mean-preserving log-normal noise, drawn afresh for every posterior draw, with independent log-standard deviations s of 0.02 (of the order of the smallest gap that protects a membership), 0.05 (roughly the relative sampling error of a mean over 66 tasks with five trials each when the per-trial coefficient of variation is near one) and 0.10, and with a total of 0.15 anchored to the marginal posterior log-standard deviations of 0.11 to 0.20 of per-trial cost and duration in the native panel, with half of that noise variance shared by all configurations. This anchor is marginal and conservative, because configurations that run on the same tasks share much of that spread, and a factor common to all of them changes no comparison. Grok 4.6 falls from 0.964 to 0.683 at $s = 0.02$ and to 0.466 under the native anchor because only a mean trial 36.9 s (1.6%) shorter keeps it clear of GPT-5.6 Sol, which costs 29.2% less and completes more in 99.6% of draws, and once the noise far exceeds that gap the two durations are ordered almost at random, so that its probability levels off near $0.964 - 0.996/2 \approx 0.466$, with GPT-5.6 Terra co-dominating in a minority of the lost draws. Opus 5 falls from 0.987 to 0.942 and 0.699 at the same two levels because Fable 5.1, whose mean trial is 18.8% shorter and which completes more in 81.6% of draws, costs only 4.6% more than Opus 5. The memberships of GPT-5.6 Sol, Gemini 3.7 Flash, GPT-5.6 Luna, Fable 5.1 and GPT-5.6 Terra stay at or above 0.98 at every level, whereas GLM-5.3 and Fable 5 rise, to 0.837 and 0.357 under the native anchor, in the mirror image of these losses: the configurations that dominate them at the reported values hold only small edges, GPT-5.6 Sol being 6.8% cheaper than GLM-5.3 and Fable 5.1 14.1% cheaper and 7.4% faster than Fable 5, and noise erodes such edges as readily as it erodes that of Grok 4.6. The native panel, in which cost and duration are jointly uncertain and paired by task, keeps GPT-5.6 Terra, GPT-5.6 Sol and Opus 5 as members (Terra 0.998, Sol > 0.999, Opus 5 0.990; Sonnet 5 0), but it contains neither Grok 4.6 nor Fable 5.1 and so cannot settle either fragile membership.

Table 12: Sensitivity of Terminal-Bench v4 Pareto membership to the reported cost and mean duration (Section E), rows as in Figure 3. "Fixed" is the membership probability of Figure 3, the next three columns multiply cost and duration by independent mean-preserving log-normal noise with log-standard deviation s, and the last uses a total log-standard deviation of 0.15 anchored to the native panel, with a fraction $f = 0.5$ of its variance shared by all configurations. Each entry averages 20 noise replicates over 16,000 posterior draws, with Monte Carlo standard errors of at most 0.001 over noise replicates and 0.005 over the eight chains.

	Reported			Independent, s			Native
Configuration	Cost ($)	Duration (s)	Fixed	0.02	0.05	0.10	0.15, f=0.5
GPT-5.6 Sol	2,542	2,388	1.000	1.000	0.998	0.997	0.997
GPT-5.6 Luna	347	4,088	1.000	1.000	1.000	1.000	1.000
Gemini 3.7 Flash	1,262	1,689	1.000	1.000	1.000	1.000	1.000
Fable 5.1	6,244	3,894	0.999	0.999	0.999	0.990	0.989
Opus 5	5,969	4,793	0.987	0.942	0.775	0.700	0.699
GPT-5.6 Terra	1,734	2,515	0.985	0.985	0.985	0.983	0.982
Grok 4.6	3,592	2,352	0.964	0.683	0.539	0.471	0.466
GLM-5.3	2,728	5,831	0.761	0.762	0.799	0.834	0.837
Fable 5	7,265	4,203	0.025	0.028	0.158	0.343	0.357
Grok 4.5	2,094	3,162	0.025	0.025	0.027	0.097	0.108
Opus 4.8	6,481	5,163	0.000	0.000	0.001	0.002	0.002
Sonnet 5	9,604	6,510	0.000	0.000	0.000	0.000	0.000

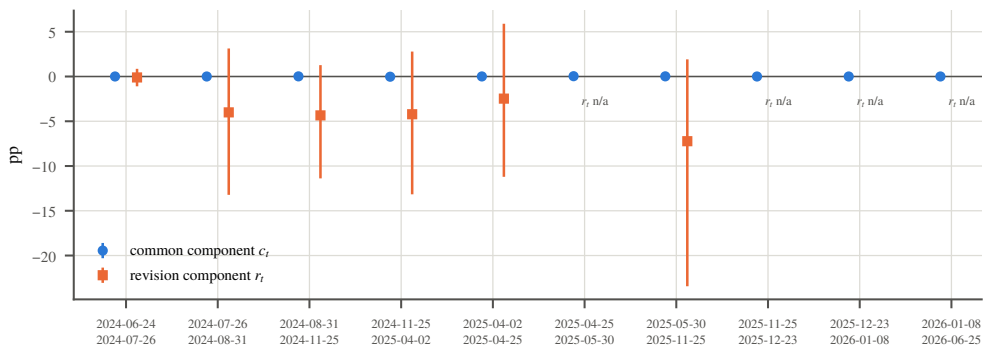


Figure 8: LiveBench regime bridge: common component c_t and revision component r_t per adjacent release transition, posterior means and 95% intervals. The revision component r_t is not identified where no task changed.

Table 13: Temporal identification audit by source (summarised by term in Table 2). Each row is one source and estimand, and "Outcome" is the predeclared disposition.

Source	Temporal estimand	Key evidence	Outcome
LiveBench, 10 adjacent releases	revision r_t vs drift c_t (Section 4)	revision hypermean -3.50 pp $[-10.64, 1.65]$, $P(<0) = 0.90$; every revision interval covers 0; holdout LPD mixed $+0.74$ on the chronological stress split but -0.94 summed over development pairs	inconclusive
LiveBench, measurement state	stochastic regime process	ΔLPD dynamic$-$static -0.021	not earned
LiveCodeBench releases (item level)	release-arrival trend	all four temporal models fail PPC (0/10-cell fraction 0.645 observed, $\approx$0.577 replicated); trend slope $+0.68$ logit/yr $[-1.27, 2.61]$	inconclusive
τ^3-bench, six-event holdout	dynamic entrant marks	held-out LPD static vs. dynamic -30.07 vs. -33.94	static preferred
SWE-bench Verified (OpenAI, 2024), 5 entries	arrival trend	dynamic gain $+6.78$ LPD (time coefficient 90% interval $[-0.50, 1.58]$) but system, scaffold and task mix change together	confounded
τ^2 banking, entrant-mark law (T0–T2)	$g(E)$ on evaluation clock	fails predictive checks (rank q=2.1×10^{-7}; duration median 5.327 outside [5.331, 5.383]; scalar gates pass); T1 vs T0 forward score not fitted	rejected
τ^2 banking, design recovery	T2 innovation scale κ; T1 trend κ_1	6,000 simulated datasets per model on the real folds: T2 power 0 at κ=0.25, posterior median 0.167 for every true κ; T1 power 0; support $\leq$5 occupied weeks vs. floors of 7 (T2) and 9 (T3)	not identifiable
τ^3/LCB-Pro/AstaBench dates	execution-date imputation	eligible imputed rows 0 of 34,800	not supported
τ^2 event panel	configuration identity	events with complete identity 0 of 102; occupied weeks per stratum 2–9	blocks inference

F Claim boundaries

G Extended limitations

Limits of public archives. Because the Terminal-Bench v4 histograms are anonymous, task-level covariance between configurations is not estimated from them, and because each configuration is a single submission, job-to-job variation is not identified. The native four-configuration panel, moreover, has only one job per configuration, so job, service and execution window are confounded within it. The τ^3-bench multidimensional evidence, for its part, rests on a single reconciled pair. Exploratory data analysis informed some of our modelling choices, and the complexity and domain holdouts were used during development, so the predictive comparisons should not be read as prospective confirmation. Finally, the analyst's own selection of sources and contrasts is a selection problem of a different kind from leaderboard top-k selection, and it is not corrected by selection-aware likelihoods.

Table 14: Claims and their evidential status.

Claim	Status	Evidence
Static, source-specific multidimensional frontiers are broad (no single winner)	supported (exploratory)	Sections 5.1 and 5.2
Linear scalarisation can discard high-probability frontier members	supported	Section 5.3
Admissibility of switching depends on horizon and cost bridge	supported under declared Λ	Section 5.5
EVSI is concentrated at decision boundaries, not where variance shrinks	exploratory	Section 5.6
Temporal model describes how the frontier evolves	not supported by current archives	Section 5.4
Organisation-specific adopt/retain/replace/retire decisions	not supported (no organisational outcomes)	Section 6
VOI identifies the most consequential evaluations in general	method only (paired estimator run for M5a only; no ranking of lifecycle pairs)	Section 5.6
Effort or harness causal effects	not identified	Section 5.1
Provider latency, economic cost, monetary EVSI	not measured	Section 6

Assurance. Every audit reported here was carried out by the authors' own pipeline and reviewers, which is to say by self-review, and none has been independently replicated. Some of the numbers in Section 5 and its appendices come from the 2- and 4-chain diagnostic fits of Table 4 rather than from fits meeting the production standard of Section C, and the slot registry in `values.tex` maps every reported number to its source bundle.

Scope relative to earlier work. An earlier version of this project deliberately placed procurement and deployment decisions out of scope, treating public archives only as objects of audit. The present paper brings decisions back in, but only in the restricted form of admissible action sets under declared scenario classes, and the earlier caution about unobserved organisational quantities remains in force.

What would change the conclusions. The limitations above also indicate what evidence would change our conclusions. Independently repeated, configuration-matched jobs across separated time blocks would identify service variance and, eventually, mutable-service change, and authenticated execution clocks and stable configuration identities would make the temporal decomposition testable. Organisational outcomes (review burden, correction, fallback success, exit costs), together with elicited losses and a prospective validation, would allow the decision layer to move from admissible sets to recommendations.