

The AI Evaluation Ecosystem

Yash Dave^{1,*}

Sang T. Truong^{1,*}

Serena Wang^{2,†}

Sanmi Koyejo^{1,†}

¹Stanford University ²University of British Columbia

*Equal contribution †Equal advising

Correspondence: yashdave@stanford.edu

Abstract

AI evaluation shapes the decisions of model providers, users, funders, and regulators. We argue that designing valid benchmarks requires contextualizing design choices in the dynamics of this ecosystem of actors. We develop a simulation architecture that combines rule-based market dynamics with LLM-driven strategic actors, building on advances in Generative Agent-Based Modeling (GABM). We model benchmarks, consumer needs, and provider capabilities as vectors over a six-dimensional capability space (reasoning, coding, knowledge, safety, communication, agentic), with structural information partitions across actors. As a case study, we apply this stylized simulation to explore benchmark holdout design. We find that moving from public benchmarks to private holdout benchmarks shrinks the gap between benchmark scores and user satisfaction on most benchmarks but widens it on a few, depending on where holdout weights shift scoring credit. We stress-test our findings at both the instrument and case-study level, drawing on the V&V framework of Sargent [2013] and GABM-specific evidence criteria. Beyond holdout design, our simulation is a hypothesis-generating sandbox for studying how evaluator and policy choices, in turn, reshape the ecosystem.

1 Introduction

Benchmarks have been among the most influential coordination devices in modern AI research. The lasting ones do more than measure: each defines what counts as progress and proposes a roadmap for follow-on research, as MMLU, HELM, and Terminal-bench each did [Hendrycks et al., 2021, Liang et al., 2023, Merrill et al., 2026]. Yet methodology for benchmark design still treats benchmarks as one-shot instruments, evaluating them on properties (calibration, validity, contamination resistance) that hold the rest of the ecosystem fixed. Treating evaluation as a closed-loop system instead changes what counts as an evaluation question: not only “is this metric well-calibrated?” but “what does the ecosystem do with this benchmark once it exists?”¹

Once a benchmark gains traction, its influence runs in both directions (Figure 1). The same scores that signal progress also steer research directions and inform model selection [Bommasani et al., 2021, Liang et al., 2023]; once optimized as targets, they distort what they were meant to track [Thomas and Uminsky, 2022]. The same pull plausibly extends to where funders direct capital, what companies red-team, and which failure modes regulators attend to. Adjacent high-stakes industries (clinical trials and aviation) suggest the gap matters: mature governance of such feedback-rich measurement systems, through mandatory disclosure, independent audit, and incident reporting, stabilizes information environments that private incentives do not (Appendix A). AI evaluation has no comparable structured account of these joint dynamics; without one, policy proposals and institutional designs risk addressing symptoms rather than causes.

The questions an ecosystem perspective raises are known to be difficult to study: for example, how holdout choices reshape provider incentives; when evaluator independence affects market structure; or which feedback channels stabilize or amplify gaming. Formal game-theoretic models become intractable as actor heterogeneity, information asymmetries, and the number of feedback channels grow. These questions are

¹Code: github.com/aims-foundations/evarium. Website: aimslab.stanford.edu/evaluation-ecosystem.

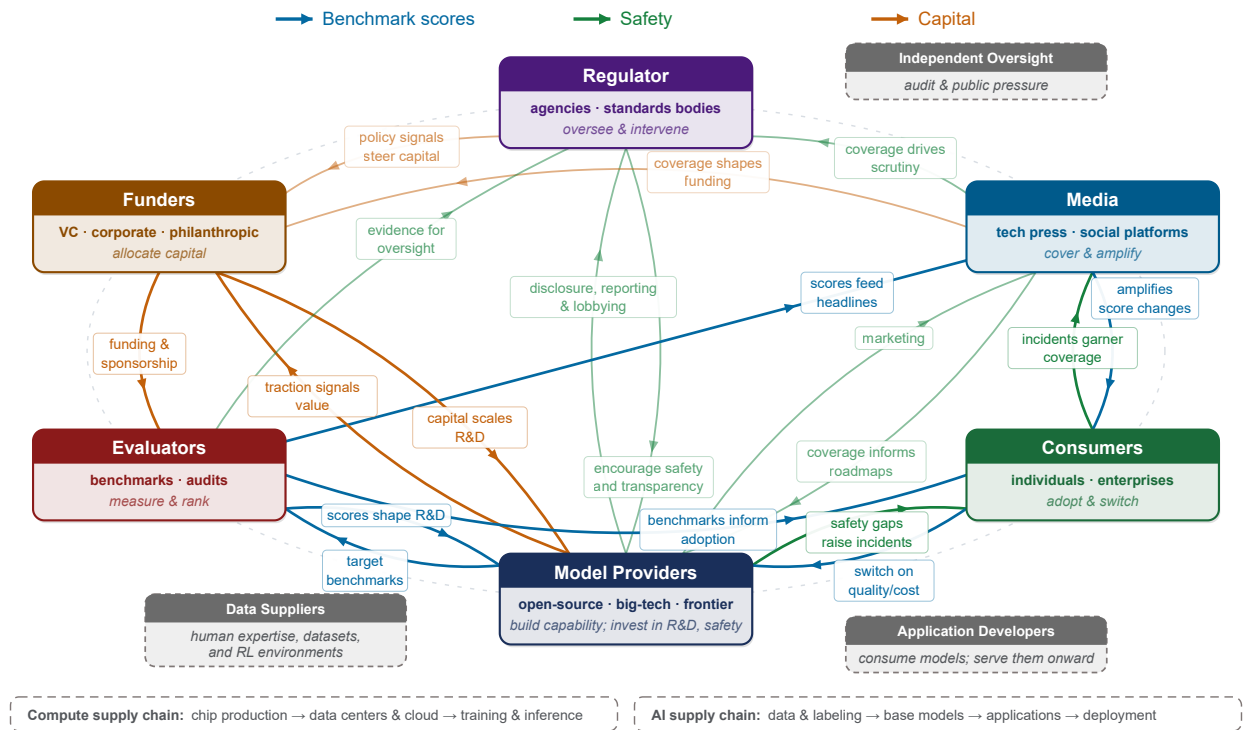


Figure 1: The AI evaluation ecosystem: six modeled actor types and the feedback loops between them, along three pathways. Fainter arrows are softer institutional channels. Grey dashed boxes mark actors and adjacent supply chains [Hopkins et al., 2025] that the simulation leaves out.

also difficult to study empirically due to their counterfactual nature, and experimental data is not always available.

Recent advances in generative agent-based modeling [Park et al., 2023, Vezhnevets et al., 2023] present a new opportunity to advance our understanding of ecosystem dynamics: actors reason over textual state (incident reports, competitor moves, regulatory news) the way real strategy teams and regulatory staff do, and a single design lever can be swapped while the rest holds fixed. The methodology to run such simulations has emerged only recently, with validation still an open problem [Manning et al., 2024, Ghaffarzadegan et al., 2024, Larooij and Törnberg, 2026, Zhou et al., 2025]; the phenomena this perspective anticipates, contamination [Sainz et al., 2023, Balloccu et al., 2024] and leaderboard manipulation [Singh et al., 2025], are now documented; and evaluation results now carry consequences on both sides of the market: the EU AI Act obliges providers of general-purpose models with systemic risk to evaluate them under standardised protocols [European Parliament and Council, 2024], while frontier developers have tied their own deployment decisions to internal capability thresholds [Anthropic, 2023, OpenAI, 2023, Google DeepMind, 2024]. Thus, in this work, we develop a framework to analyze the AI evaluation ecosystem. We operationalize this framework through a GABM to study one **design lever**, a choice the evaluator alone controls: whether a benchmark’s items and their dimension weighting are published or withheld, which we refer to as **holdout design**. The simulation serves as a hypothesis generator rather than a predictor. It surfaces candidate dynamics, together with the conditions they depend on, for empirical work and policy analysis to test. Our contributions can be summarized as follows:

- We describe an **ecosystem-level framework for AI evaluation** that includes providers, evaluators, consumers, regulators, funders, and media coupled through feedback loops across development, market, and regulatory dynamics; distinct from supply- and value-chain framings (§3).
- To study this, we propose a **hybrid generative agent-based model (GABM)**. LLM-driven actors reason over observable state with structurally enforced public/private/ground-truth information partitions; a heuristic mode isolates structural from behavioral dynamics (§4).
- We apply this model in a **concrete simulation study on holdout design**, showing that in the context of the ecosystem, private-holdout benchmarks do not always reduce the gap between the benchmark score and user satisfaction (§5).
- We **stress-test our simulation methodology** at the instrument level (is the simulation a reasonable laboratory?) and the case-study level (does the holdout-design finding hold up?), drawing on the V&V framework of Sargent [2013], the principles of Zhou et al. [2025] for LLM multi-agent simulations, and the evidence hierarchy Vezhnevets et al. [2023] adapt for GABM (§5.2, Appendix E).

2 Related Work

Recent work on AI supply chains traces how foundation models, data, and services flow through networks of specialized actors [Hopkins et al., 2025, Cobbe et al., 2023, Widder and Nafus, 2023, Bommasani et al., 2024]. Evaluation sits orthogonal to these framings: it is neither a product flowing downstream (the supply-chain view) nor one of the value-creating activities into which a firm is disaggregated (the value-chain view of Porter [1985]), but a shared measurement infrastructure on which providers self-assess, consumers choose, funders allocate, and regulators intervene. Our work studies that layer as a distinct ecosystem, building on three adjacent lines of research: AI evaluation and benchmarking, governance of information systems (treated in Appendix A), and generative agent-based modeling with LLM actors.

AI evaluation and benchmarking. Frameworks such as HELM [Liang et al., 2023] set standards for coverage and standardized, multi-metric measurement, and the Foundation Model report [Bommasani et al., 2021] calls for valid evaluation across many criteria; preference evaluations such as Arena (formerly LMArena and Chatbot Arena) [Zheng et al., 2023, Chiang et al., 2024] complement them with dynamic rankings. A parallel literature documents evaluation pathologies: metric gaming [Thomas and Uminsky, 2022, Manheim and Garrabrant, 2018] (kin to Campbell’s Law and the Lucas critique [Campbell, 1979, Lucas, 1976]), contamination [Sainz et al., 2023, Balloccu et al., 2024, Magar and Schwartz, 2022], critiques of construct validity [Jacobs and Wallach, 2021, Raji et al., 2021], and meta-scientific analyses of how benchmark adoption concentrates on a few canonical datasets [Dehghani et al., 2021, Koch et al., 2021]. These works address failure

modes at the individual-benchmark level; we model evaluation within an ecosystem in which competitive dynamics, capital allocation, and consumer behavior produce failures even when individual benchmarks are well designed.

Generative agent-based modeling with LLMs. Generative agent-based models (GABMs) have developed along two axes. Methodologically, work including Aher et al. [2023], Park et al. [2023], Gao et al. [2023], Manning et al. [2024], and Cross et al. [2025] has built increasingly rigorous methods for designing and validating generative-agent simulations, and Ghaffarzadegan et al. [2024] add robustness and prompt-sensitivity checks. In parallel, Larooij and Törnberg [2026] argue that LLMs’ black-box structure, cultural biases, and stochastic outputs make validation harder, and Wu et al. [2026] document variance underrepresentation in LLM actors. Our simulation restricts LLM reasoning to strategic planning (investment allocation, regulatory choice, funding allocation) while grounding market mechanics, scoring, and state transitions in deterministic rules: a hybrid that preserves LLM expressiveness without sacrificing the causal transparency that ABMs require. Extended discussion of this design choice is in Appendix A.

3 Mapping the AI Evaluation Ecosystem

Evaluation does not occur in isolation: it unfolds through the interactions of actors with different goals, information access, and decision rights, whose responses to evaluation outcomes reshape what future evaluations measure. Figure 1 illustrates the six actor types and their primary information flows.

This section sets out the frame: who the actors are, what each of them observes, and which channels carry evaluation results between them. Section 4 turns the frame into an instrument, a simulation in which every channel is explicit and can be altered one at a time, and Section 5 puts the instrument to work on a single case study, the design of benchmark holdouts.

Model Providers. Model providers (e.g., OpenAI, Google, Anthropic, open-source consortia) develop and deploy AI systems, observing internal signals (user data, internal red-team results, commercial metrics) and external signals (public benchmarks, competitor performance, media coverage, policy requirements). They release models and adjust training strategies in response to benchmark performance.

Evaluation Providers. Evaluation providers (academic groups, independent labs, third-party efforts such as Stanford CRFM’s HELM [Liang et al., 2023], ARC Evals [Shevlane et al., 2023], and third-party auditors [Raji et al., 2022]) design benchmarks, publish scores, and retire outdated evaluations. They may face conflicts of interest when paid by, or selling services to, the providers they evaluate [Costanza-Chock et al., 2022, Raji et al., 2022]. Third-party evaluators have since proposed minimum operating conditions for independent evaluation, covering model access, contingent compensation and recusal [AI Evaluator Forum, 2025].

Consumers. Consumers range from individuals to organizations facing integration costs and regulatory exposure. Benchmark scores reach them through two channels: a direct leaderboard signal, weighted by how much a consumer trusts published scores relative to their own experience (a modeling choice motivated by practitioner interviews in Hardy et al. [2025]), and indirect exposure through news coverage of score movements and incidents. Scores serve as proxies for capabilities consumers cannot directly assess, a credence good [Dulleck and Kerschbamer, 2006], and adoption decisions drive provider revenue and funder attention.

Regulators. Regulators observe evaluation results through public leaderboards, technical reports, and audits, and can mandate compliance evaluations or enforce disclosure [European Parliament and Council, 2024], require developer reporting as US policy did until 2025 [The White House, 2023], or issue voluntary frameworks [National Institute of Standards and Technology, 2023]. They cannot observe provider capabilities or consumer satisfaction directly, acting instead on public signals, incident reports, and media coverage.

Funders. Funders (venture capital, corporate investors, government agencies, foundations) allocate capital based on benchmark scores, market adoption, media sentiment, and incident history. Their decisions are performative [MacKenzie, 2006, Perdomo et al., 2020]: capital flows toward high-scorers, reinforcing the competitive value of benchmark optimization.

Media. Media outlets observe scores, market shifts, incidents, and regulatory actions. Selective coverage creates an attention economy that amplifies or attenuates signals other actors receive.

4 Generative Agent-Based Model of the Evaluation Ecosystem

To study how the feedback dynamics described in Section 3 produce emergent failures, we developed a GABM modeling six providers, an evaluator, a media actor, a regulator, four funder types, and 51 consumer segments, with actor decisions implemented in both LLM-driven and heuristic modes. The unit of modeling is the institution, not a human: LLM agents play organizations (a provider’s strategy team, an investment committee, a regulatory body) making monthly allocation decisions, while the one mass-human population, consumers, is rule-based throughout. Following RecSim’s view of simulators as controlled environments that need not be faithful to live systems [Ie et al., 2019], we use the simulation as a hypothesis generator: the results surface candidate dynamics worth investigating empirically, rather than proving that specific failures will occur in real systems. Several structural features described below are deliberate modeling choices: the six-dimensional capability ontology, hand-calibrated consumer need weights, the linear scoring model, the form of holdout-weight perturbation, and a dedicated safety lever. The simulator tests mechanism plausibility under this stylized structure: outcomes such as the score–satisfaction gap are partly defined by it, and the experiments show which patterns emerge predictably from interactions on top.

4.1 Benchmark Model and Structural Misalignment

Each provider’s capabilities are represented by a six-dimensional **capability vector** $\mathbf{c}_p$ across reasoning, coding, knowledge, safety, communication, and agentic capabilities. Each benchmark b has public task categories but *hidden* per-category dimension loadings $\mathbf{w}_b$ known only to the simulation. At round t , a noisy measurement is drawn and the published score is the running max: $\hat{s}_{p,b,t} = \max(\hat{s}_{p,b,t-1}, \tilde{s}_{p,b,t})$, where $\tilde{s}_{p,b,t} \sim \mathcal{N}(\mathbf{c}_p^\top \mathbf{w}_b, \sigma_b^2/n_b)$ and n_b is the number of scored items on benchmark b . The monotonic non-decreasing constraint models public reporting behavior. No actor, including LLM-driven agents, can access the true dimension weights.

The simulation begins with four benchmarks at round 0, anchored to real evaluations (MMLU, HumanEval, TruthfulQA, MT-Bench), and introduces nine more in a preset order over the 40-round horizon, for 13 active benchmarks total (Appendix B.3; the dynamic-mode draws from an extended 22-benchmark pool).

The ecosystem is *structurally misaligned*: averaged over the 13-benchmark suite, benchmark weights exceed population-average consumer need weights on reasoning and coding, and they fall below them on communication early in a run and on safety late in a run, as the organizational share of consumers grows (per-dimension arithmetic in Appendix B.3). The suite is modeled on the 2023 benchmark landscape, whose newer benchmarks target coding and advanced reasoning [Maslej et al., 2024]. A provider that follows benchmark scores therefore invests in a different mix of dimensions than consumers need. The experimental claim is that under stable ecosystem-wide investment asymmetries (e.g., consistent over-investment in reasoning relative to safety or knowledge), the resulting per-benchmark gaps become structured and predictable rather than random across conditions.

4.2 Provider Investment and Emergent Gaming

Each round, providers allocate their training budget across three levers: **R&D** directs capability growth toward dimensions the provider believes benchmarks and consumers value; **Safety** improves the safety dimension and reduces incident probability; **Product** quality-gates the consumer-need signal feeding R&D targeting and raises effective switching costs. Allocations are applied as a 3-round rolling average to model execution lag between strategic decision and operational effect. Full mechanics in Appendix B.5.

Belief update. Providers maintain per-benchmark beliefs about hidden weights ($\hat{\mathbf{w}}_{p,b}$), initialized from a noisy reading of $\mathbf{w}_b$ ($\sigma=0.05$ per dimension, clipped and renormalized) that stands for what a benchmark’s stated scope reveals, and updated each round from score prediction errors (learning rate 0.10–0.20 by provider profile). The exact vector remains unobservable throughout. Belief updating is rule-based in both LLM and heuristic modes and reflects the epistemic state of the provider.

Satisfaction and the gap. Published scores shape which provider a segment adopts, but not how satisfied the segment is with it: a segment’s satisfaction is computed from the provider’s true capabilities, weighted by the dimensions that segment needs. The headline quantity is the per-benchmark **score–satisfaction gap** g : a benchmark’s published score minus the *matched satisfaction* the same provider delivers, an average

over consumer segments weighted by segment size and by how closely each segment’s needs align with that benchmark’s dimension weights $\mathbf{w}_b$, so that g compares a benchmark’s score against the consumers whose needs it most closely matches. Consumer switching adds incident and cost terms to satisfaction (Appendix B.7); g uses the capability term alone. A positive gap means the benchmark promises more than consumers get; a negative gap means it promises less.

Emergent gaming. The score–satisfaction gap arises as a measurement artifact when two structural conditions are simultaneously met: (1) benchmark dimension weights diverge from consumer need weights, and (2) providers discover, through accumulated score prediction errors, that benchmark-focused investment produces faster score gains than consumer-focused investment. This operationalizes Goodhart’s Law [Manheim and Garrabrant, 2018, Thomas and Uminsky, 2022] as a structural consequence of information architecture: the gap is a property of the system’s design, not a modeled strategy.

4.3 Ecosystem Actors and Feedback Channels

The simulation operationalizes the actor relationships from Section 3 as concrete feedback channels. Information access is structurally enforced across three tiers: *public* state (scores, leaderboard, market shares) visible to all; *private* state (beliefs, strategies) visible only to the owning agent; and *ground truth* (capability vectors, benchmark dimension weights, true consumer satisfaction) held by the simulation and inaccessible to any actor prompt.

Figure 4 (Appendix B.1) summarizes the per-actor mechanics and their round-order dependencies; only design choices that are load-bearing for later results are called out here. The **heterogeneous consumer market** comprises 51 segments (17 use-case profiles, each crossed with the three behavioral archetypes of its consumer type, individual or organizational; archetypes vary in leaderboard trust, switching cost, and cost sensitivity), with *experience-based* satisfaction that reflects capability actually delivered to a segment rather than the published score, making the score–satisfaction gap structural rather than a modeled strategy (Appendix B.7).

Incidents are drawn from a four-level severity distribution (minor 50%, moderate 31%, major 12%, critical 7%) and propagate simultaneously to all downstream actors (Appendix C.6). **Media** influence runs through attention, not direct satisfaction (Appendix B.8): negative coverage raises the fraction of a provider’s users who enter exploration. The **regulator** applies a graduated escalation ladder with cooldowns between interventions (Appendix B.9). Four **funder** types (VC, corporate, government, foundation) allocate capital from fixed pools into the provider budget via Eq. 1, closing the performativity loop of Section 3; VC funders do not fund open-source providers (Appendix B.10). One **model provider** (OpenCore) is an open-weight lab that differs from the closed providers in its structure: VC funders do not fund it, it has a cost advantage that represents self-hosting savings, its safety capability floor is slightly lower, and its open weights move the other providers’ benchmark beliefs toward the public weights each round. These differences are patterned on differences between open-weight and closed labs (Appendix B.11).

4.4 Round Protocol and Experimental Setup

Figure 4 in Appendix B.1 shows the order of events in a round. In Phases 1 to 4, providers plan and update capabilities, the evaluator scores all providers and publishes the leaderboard, providers update beliefs from score prediction errors, incidents are generated, and the media covers the round. In Phase 5, consumers, the regulator and funders observe the round’s public state and respond in that order. Funding and regulatory effects reach providers at the start of the next round. Full pseudocode is in Appendix B.1. The split between the LLM and heuristic modes is methodologically intentional: heuristic mode supplies the statistical structure (large- N distributional claims about direction and ordering), while LLM mode supplies strategic reasoning traces (Appendix F) that allow qualitative auditing of *why* actors take the actions they do. The lower seed count limits distributional inference in LLM mode; we therefore use LLM runs primarily for qualitative reasoning traces and cross-checking directional patterns.

Round-based protocol and sequential phasing. A round maps to the cadences we model (budget cycles, benchmark releases, intervention windows, funding rounds) and keeps LLM decision points interpretable. Within a round, actors resolve sequentially so later actors observe earlier outcomes, mirroring real-world information flow: consumers respond to the month’s scores, regulators to its incidents, funders to its

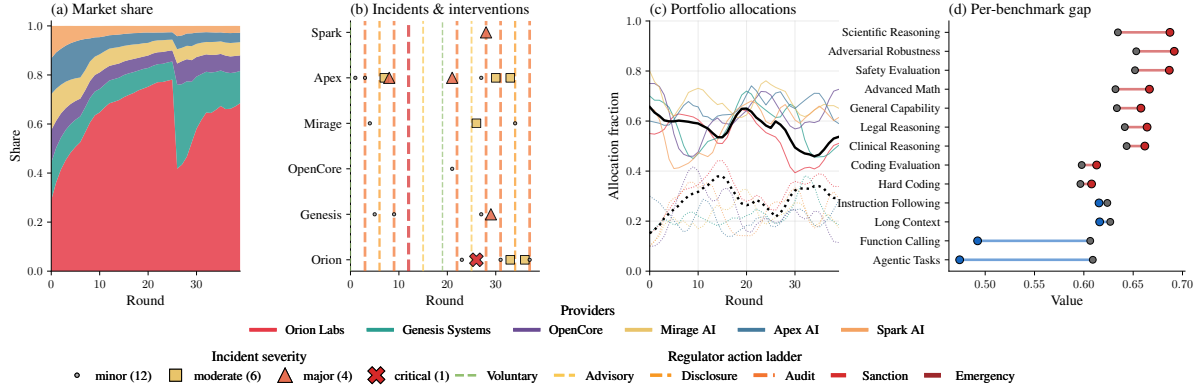


Figure 2: One 40-round LLM-driven run, calibrated **baseline**, Claude Opus 4.6, seed 43 (`core_privacy/llm/claude-opus-4-6/baseline/seed_43`). (a) Stacked market share. (b) Per-provider incident markers (shape = severity) and regulator interventions (vertical dashed lines, color = action on the escalation ladder). (c) Portfolio allocations: R&D (solid) and safety (dotted) per provider, share-weighted mean in bold. (d) End-of-run per-benchmark score (colored) vs matched consumer satisfaction (gray); red = benchmark overpromises, blue = underpromises.

leaderboard. Providers, the regulator and funders are LLM-driven; the evaluator is optionally LLM in dynamic mode; the media actor and the consumer market are rule-based throughout. A fully heuristic mode serves as a non-LLM baseline, separating structural dynamics from LLM-specific behavior. Prompts are framed in the natural language of each role (strategy teams, investment committees, regulatory staff) without apparatus vocabulary. Hyperparameters are consolidated in Appendix C.1.

Example run. Figure 2 shows one baseline run. The leader accumulates market share through round 25. A critical incident then costs it 36pp of share in a single round, the regulator commissions an audit of it five rounds later, and it recovers only to $\sim 68\%$ by round 40 (panels a, b). Safety allocations span $\sim 10\text{--}44\%$ across providers, and the ranking among them changes repeatedly over the run (panel c). Published score and matched consumer satisfaction diverge non-uniformly across the 13-benchmark suite (panel d).

5 Results: Case Study of Benchmark Holdout Design

This section uses the simulation as a hypothesis generator: varying one structural property of the evaluator at a time, we trace how evaluation practice propagates through provider investment, consumer choice, and capital allocation. The patterns below are candidate dynamics, not claims about real-world markets. Evidence is organized into two layers. The *structural* layer draws on the heuristic baseline ($N=50$ seeds per condition) and supports distributional claims. The *behavioral* layer draws on LLM-driven runs ($N=10$ Sonnet seeds per condition, matched seeds for paired comparisons; cross-vendor robustness in Appendix F.7). The two layers cross-check each other. Agreement in direction across modes suggests that a finding does not depend on LLM-specific prompt framing or on a particular heuristic rule. Both modes share the same scoring layer, so the check bears on the planner’s contribution to capability and leaves the scoring mechanism itself untested. Published score and matched consumer satisfaction do not agree uniformly across benchmarks (Figure 2, panel d); a single pooled score–satisfaction gap hides that the gap moves differently on different benchmarks. The analyses below ask what determines each benchmark’s direction, and what an evaluator’s privacy choice does to it.

5.1 Holdout design: private holdouts shrink the gap unevenly across benchmarks

The evaluator’s first lever is what it publishes, an information-design problem [Kamenica and Gentzkow, 2011]. We vary holdout design across a five-condition ladder over the 13-benchmark active set: **public_only**, **baseline** (an 8/3/2 public/partial/private mix reflecting the coexistence of public, semi-private, and fully-private evaluation regimes in 2024–2025 [Scale AI, 2024, Center for AI Safety et al., 2026, White et al., 2025],

building on the public/private split ARC introduced [Chollet, 2019]), **private_dominant**, **private_only** (all benchmarks holdout-only with perturbed weights, target cosine 0.85), and **iid_holdout** (same weights as public, isolating reporting lag from weight asymmetry). The cosine knob aggregates contamination, training-on-the-test-task, and adversarial-construction sources of public-vs-holdout asymmetry; mechanism details in Appendix C.4. The cleanest privacy contrast is between the two 100%-endpoint conditions; the calibrated **baseline** pools privacy with which benchmarks happen to carry the **partial** label.

The mechanism is weight asymmetry between public and holdout dimension weights (Figure 3b). Under **private_only**, scoring weights are perturbed from the true weights (target cosine 0.85), leaking credit from the dominant dimension onto adjacent ones. When credit moves from a stronger dimension to weaker ones, score drops; when it moves from a weaker dimension to stronger ones, score rises. The lower the population’s capability surplus on the dominant dimension, the more Δg rises ($r = -0.69$ LLM, -0.71 heuristic, over the 11 benchmarks of panel b). Matched satisfaction (true weights) is unaffected, so the gap moves.



Figure 3: (a) Per-benchmark gap g (published score – matched consumer satisfaction, where matched satisfaction is ground-truth capability scored against consumer need-weights and averaged over segments, each weighted by its size and by its alignment with the benchmark’s dimension weights) across the five privacy conditions, for five representative benchmarks (one per primary capability dimension; full 13-benchmark version in Appendix F.2). Foreground: LLM Sonnet 4.6 ($N=10$) with 95% CIs. Background: heuristic mode ($N=50$, all five conditions). Δg per row = $\text{gap@private_only} - \text{gap@public_only}$. (b) Δg vs. population capability surplus on each benchmark’s primary dimension (capability on the highest-weighted dimension minus the cross-dimension mean, computed from **public_only** seeds), over 11 benchmarks; Agentic Tasks and Function Calling are excluded because the agentic dimension starts far below the other five (initial capability 0.08 to 0.12 against 0.25 to 0.55, Table 5), which makes their surplus values extreme. The exclusion is conservative: Appendix F.2 reports the fit on all 13, where the slope tightens.

Privacy does not uniformly shrink the gap. From **public_only** to **private_only**, the absolute gap shrinks on 10 of the 13 benchmarks and widens on three (Instruction Following, Clinical Reasoning, Legal Reasoning) in LLM mode, and shrinks on 9 of 13 in heuristic mode, which also widens Adversarial Robustness, the lone cross-mode disagreement, where both shifts are small (Figure 3a labels Δg for the five representative rows; all 13 in Appendix F.2). The widening is stable across seeds: the three widen in 10 of 10 LLM seeds and 96–100% of heuristic seeds, while under the **iid_holdout** null (identical weights to **public_only** but private reporting) every benchmark widens in 40–70% of seeds (Table 18). Along the perturbation axis (target cosine 1.0, 0.95, 0.85), each benchmark’s gap moves monotonically in its own direction: the pattern is consistent with weight asymmetry rather than reporting lag or observation noise.

Worked example: Clinical Reasoning. Clinical Reasoning loads primarily on *knowledge* (weight 0.65). The suite places more of its weight on reasoning (≈ 0.26) than on knowledge (≈ 0.16 ; Appendix B.3) and R&D growth tracks those weights, so providers’ knowledge sits below their reasoning. Under **private_only**,

holdout perturbation leaks credit from knowledge onto *reasoning*, where providers are strong; score rises while matched satisfaction (capabilities unchanged) does not. A benchmark nominally about medical reasoning ends up rewarding reasoning-strong models: the “accidental credit” direction. The finding is conditional: a benchmark widens when its holdout weights shift credit toward stronger dimensions. We do not test whether any real benchmark has that property; it is a separate measurement question, which Hardy et al. [2026] approach through factor analysis of leaderboard data. In the simulation, the hand-authored holdout directions and the suite determine which benchmarks widen. In a suite whose round-0 set includes a domain-expertise benchmark, Clinical and Legal Reasoning never widen in heuristic seeds, while Instruction Following widens in every seed of every suite tested (Appendix F.2).

Consequence for evaluator design. Privacy is not a uniform tool for mitigating benchmark-gaming. It redistributes scoring credit across dimensions, toward or away from where the R&D architecture already concentrates capability; the benchmark suite shapes where that is. Changing the suite’s composition changes which benchmarks widen under privacy, as the worked example shows for Clinical and Legal Reasoning, though not for Instruction Following. Under a non-flat capability ontology (§5.3) there is no neutral benchmark suite: composition choices shape where perturbation-induced shifts land, and heterogeneous consumer needs shape which of those shifts widen the gap. In this light, a benchmark’s utility is not an intrinsic property; it is conditional on the ecosystem it enters.

5.2 Validation

Validation covers two tasks: **instrument validation** (whether the simulation is a reasonable laboratory) and **case-study validation** (whether the holdout-design finding survives perturbation). The case-study evidence is a stack of mechanism-isolation checks: cross-mode sign agreement on 12 of 13 benchmarks, the `iid_holdout` null, cross-vendor replication on Opus 4.6 and GPT-5.5 ($N=3$ seeds each), a within-benchmark rescoring, and a fitted null recovering 97% of Δg variance. These checks are detailed in Table 1, the prose below, and Appendix E. We organize this within the V&V activities of Sargent [2013] and the PIMMUR audit of Zhou et al. [2025] (Appendix E.4); in the evidence hierarchy Vezhnevets et al. [2023] adapt for GABM we sit on the lower rungs below its gold standard of prediction on new real-world data (theory consistency, plus the model-comparison and robustness practices it recommends), since no real-world counterfactual ecosystem is available.

Table 1: Validation evidence by task. The instrument column records design guarantees and calibration provenance; case-study validation stress-tests the holdout-design Δg result.

	Instrument checks (design & calibration)	Case study validation (holdout design)
Approach	Sargent V&V activities; each structural assumption rated calibrated, anchored, or stipulated.	Mechanism-isolation tests on the per-benchmark Δg result.
Evidence	Three-tier visibility enforced structurally (Appendix B); parameter grounding ($K=3$ calibrated, cosine targets 0.85/0.95 anchored, need-weights anchored in sector-level evidence; Appendices C.1 and C.2).	Cross-mode sign agreement on 12 of 13 benchmarks; cross-model replication on Opus 4.6 and GPT-5.5 ($N=3$ seeds each; Appendix F.7); <code>iid_holdout</code> channel isolation; within-benchmark rescoring (Appendix F.2); fitted null recovers 97% of Δg variance (Appendix F.5).

Heuristic planners cannot see the `public/partial/private` label, yet the two modes agree in sign on 12 of 13 benchmarks, so LLM-specific reasoning is unlikely to drive the result. Both modes share the scoring layer, which the rescoring below tests directly. The `iid_holdout` null (cosine = 1.0) isolates weight asymmetry from reporting lag and observation noise.

What the simulation adds. The holdout result has a simple description, and the simulation shows where it holds. A benchmark’s gap shift is provider capability weighted by the difference between its holdout and public scoring weights. Rescoring end-of-run capability under both sets of weights reproduces the direction of the shift on all 13 benchmarks and most of its size (Appendix F.2). The description needs the run: starting capability gives the wrong direction on three benchmarks, because the widening set appears only

after R&D tilts capability toward what the suite emphasizes. A regression on starting conditions fits the shifts closely, mostly because weight differences are fixed for each benchmark, so it does not measure what the trajectory adds (Appendix F.5). The other questions in Table 2 have no obvious simple description. Checking a candidate against the full run is how one would find it, and in one small-sample comparison market concentration differs in sign between heuristic and LLM planners (Appendix H.1).

5.3 Limitations

The privacy mechanism in §5.1 rests on four modeling choices (Appendix G.1); broader action-space simplifications and methodological bounds of GABMs are in Appendices G.2 and G.3. The conclusions are conditional on the assumptions of §3 and §4; Appendix C.2 rates each as calibrated, anchored, or stipulated. The benchmark dimension loadings (Table 4) are stipulated, and every score and every gap is computed against them.

1. Privacy operates through weight perturbation in this simulator. “Privacy” in this simulator decomposes into three orthogonal channels (Appendix C.4): a cosine-distance perturbation $\cos\theta$ between public and holdout dimension weights (targets of 0.85 / 0.95, anchored by within-family Pearson correlations we compute from Epoch data [Epoch AI, 2024] and by retro-holdout score inflation in Haimes et al. [2024]), a $K=3$ reporting lag, and observation noise scaling with $\sqrt{\text{samples} \times h}$. The cosine channel is itself a *composite proxy*: $\cos\theta < 1$ aggregates item-level contamination and memorization, “training on the test task” effects [Dominguez-Olmedo et al., 2025], and adversarial holdout construction. The simulator does not separately resolve these sources: a result attributed to weight asymmetry may be driven dominantly by any one of them, or by their interaction. The `iid_holdout` null ($\cos\theta = 1$) isolates the composite cosine channel from lag and noise; `private_dominant` vs. `private_only` isolates the marginal effect of moving from a target $\cos\theta$ of 0.95 to 0.85; realized cosines differ across benchmarks (Appendix C.4). A within-benchmark rescoring separates the *scoring-layer* effect (same capability, different scoring weights) from the *R&D-direction* effect (cosine perturbation drifts inferred weights, which steers R&D): in heuristic mode the scoring layer carries 0.93–0.96 of Δg , and the capability effect is indistinguishable from zero on 11 of 13 benchmarks (Appendix F.2). The decomposition covers heuristic runs only; LLM providers, which reason about inferred weights, may place more of Δg on the capability channel. Provider prompts describe the benchmark types neutrally, following Zhou et al. [2025]’s Minimal-Control principle, so the result is untested for providers prompted to target holdout weights (Appendix G.3).

2. The six-dimension capability ontology is a design choice. We model six dimensions (reasoning, coding, knowledge, safety, communication, agentic), anchored to HELM- and LMSYS-style benchmark taxonomies and to where industry investment is currently separable. Capability ends up uneven across dimensions, because the suite weights some dimensions more heavily than others and safety carries a dedicated portfolio lever, and weight asymmetry between public and holdout weights acts on that unevenness. The mechanism is robust within this ontology (Figure 3), but a flatter ontology or different investment-axis separability could shift benchmarks between the shrinking and widening sets or eliminate the asymmetry entirely.

3. The simulator covers only public-benchmark-driven competitive contexts. The instantiated ecosystem is leaderboard-driven, public-benchmark-dominated, multi-firm competitive, with a single global jurisdiction; six providers, no entry or exit. The mechanism applies to ecosystems with this structure but its engagement under FDA-style gatekeeping, defense procurement, or enterprise-B2B contexts (where buyers run internal proofs-of-concept) is unclear. We also do not model providers’ *internal* evaluation infrastructure: the capability tests informing Responsible-Scaling-Policy and Preparedness-style deployment decisions [Anthropic, 2023, OpenAI, 2023, Google DeepMind, 2024], a recent (2023+) phenomenon and an additional pressure on R&D direction not captured here.

4. Consumer satisfaction is model-defined. Satisfaction is a deterministic function of the (ground-truth) capability vector and consumer need-weights, both held by the simulation. Need-weights are assigned to each consumer profile from sector-level evidence on what that profession uses AI for and what it is concerned about, together with occupation-level adoption rates in Bick et al. [2024] (Appendix C.1); the step from that evidence to preference weights is a researcher-imposed judgment not validated against user-reported preference. Need-weights do not enter a benchmark’s shift under privacy, $\mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$. They set the starting gap, whose sign decides whether that shift widens or shrinks it. Measured against other consumers

the counts change: against safety-heavy segments alone, four of the 13 classifications flip (Appendix G.1).

Why the result is still useful. The simulation is a laboratory, not a crystal ball. Private holdouts shrink the gap on most benchmarks and widen it on a few, those whose holdout weights shift credit toward dimensions where R&D concentrates. This holds where leaderboards drive decisions, public benchmarks dominate, capability is uneven across dimensions, and gaps are measured against a mixed consumer population.

5.4 A catalog of simulation case studies

The privacy ladder in §5.1 is one case study; the ecosystem frame supports a broader class of policy and institutional-design questions that share the same architecture. Table 2 collects representative examples, each pairing a policy or mechanism question with its implementation slot in the current simulation and the strategic dynamic it would surface.

Table 2: Additional case studies the ecosystem simulation can host. Case 1 (holdout design) is the worked example in §5.1 and case 2 (conflicts of interest) has a small-sample comparison in Appendix H.1; cases 3 and 4 are presented as architectural slot-ins rather than worked studies. Further slots (media shadow, benchmark sponsorship) in Appendix H.

Question (policy framing)	Simulation slot-in	Strategic angle
2. Conflicts of interest (Arena; Scale SEAL). Does monetizing submissions redirect capital toward aggressive submitters at consumer cost?	<i>Small-sample comparison in Appendix H.1.</i> Per-submission fees (R&D budget); best-of- N selection; paid early access.	Larger-budget providers capture more submissions; funders read inflated scores as quality signals; capital flows to aggressive submitters rather than providers best serving users.
3. Transparency mandates (EU AI Act Art. 53; CA SB-53; NY RAISE Act). Do forced allocation or capability disclosures reduce gaming?	New regulator lever: exposes part of provider private state (R&D mix, internal eval results) to public.	Providers face disclose-vs-window-dress tradeoff; tests whether visibility of the <i>R&D mix</i> changes Goodhart pressure or just changes the narrative around it.
4. Audit & verification (UK AISI pre-deployment agreements; WH 2023 voluntary commitments). How effective are audits at enforcing voluntary safety commitments?	Regulator lever: provider commits to a forward safety allocation at t ; regulator verifies at $t+k$ with penalty on violation.	Game-theoretic: provider can over-promise, under-promise, or credibly commit. Tests whether self-binding works when verification is credible, and how it interacts with incident pressure and market share.

The architecture treats each of these as a specific lever added to an existing actor, a new channel between actors, or a perturbation of an existing information flow. The role of the simulation is not to predict the real-world effect quantitatively but to reveal which ecosystem pathways carry the effect: which actors are pivotal, which information asymmetries amplify or dampen it, and where the intervention produces the counterintuitive dynamic that an empirical study could then target.

6 Conclusion

We argue that AI evaluation is best understood as a system of interacting actors whose incentives, information, and strategic behavior jointly determine what gets measured, how results are used, and which failure modes emerge. We developed a systems-level framework mapping these actors and feedback loops, instantiated it as an agent-based simulation, and used it to study one focused case: how does benchmark holdout design shape ecosystem outcomes?

Private holdouts shrink the score-satisfaction gap unevenly. From `public_only` to `private_only`, the gap shrinks on 10 of the 13 benchmarks and widens on three in LLM runs, in every seed. A benchmark widens when its holdout weights move scoring credit toward the dimensions where provider R&D has concentrated capability, so the mechanism is weight asymmetry between public and holdout dimension weights. Three checks support this reading. Under the `iid_holdout` null, which keeps the public weights and changes only the reporting, each benchmark widens in 40–70% of seeds and the stable pattern disappears. Rescoring

end-of-run capability under the holdout weights reproduces the direction of the change on all 13 benchmarks. Heuristic and LLM planners agree on the sign for 12 of 13.

Which benchmarks widen depends on the suite. When the opening suite includes a domain-expertise benchmark, Clinical and Legal Reasoning never widen in heuristic runs, while Instruction Following widens in every suite we tested. There is no neutral benchmark suite, because the choice of benchmarks decides where the widening lands. The finding holds under the conditions the simulation models, a leaderboard-driven market with gaps measured against a mixed consumer population, and it needs a test on real leaderboards.

The simulation supplies what observation cannot. In the real ecosystem, true capability and provider R&D allocation are unobservable, and no natural experiment holds a benchmark suite fixed while its holdout policy changes. In the simulation we change one evaluator choice and let providers respond, and the response creates the result. The same rescoring applied to starting capability gives the wrong direction on three benchmarks, so the widening set exists only after R&D has tilted capability toward what the suite rewards. We use the simulation as a laboratory. It produces hypotheses, and the conditions they depend on, for empirical studies and policy work to test. It predicts no real market.

The same simulation can host other questions about evaluator and policy choices. Evaluator conflicts of interest has a small-sample comparison in Appendix H.1, transparency mandates and audit verification are specified in §5.4, and media coverage and benchmark sponsorship are outlined in Appendix H. Each asks how one actor’s choice moves through the rest of the ecosystem, and each can be examined before a policy depends on the answer. Holdout design is the first of these questions we have worked through.

Acknowledgments

We thank the practitioners who spoke with us about evaluation practice. We also thank the AI Measurement Science reading group at Stanford University for taking part in the early mapping of the evaluation ecosystem and for their feedback. This work is partially supported by NSF 2046795, 2205329, and 2504264, NIH, ARPA-H, the MacArthur Foundation, Good Ventures, Schmidt Sciences, the Hasso Plattner Förderstiftung, and Stanford HAI. Truong is supported by a Microsoft Research Fellowship. Wang is supported by a Canada CIFAR AI Chair.

References

- Adobe. Inaugural Adobe creators’ toolkit report: 86 percent of global creators use creative generative AI, see it boosting creator economy. <https://news.adobe.com/news/2025/10/adobe-max-2025-creators-survey>, 2025. Media alert, October 28, 2025.
- Carlo Adornetto, Adrian Mora, Kai Hu, Leticia Izquierdo Garcia, Parfait Atchade-Adelomou, Gianluigi Greco, Luis Alberto Alonso Pastor, and Kent Larson. Generative agents in agent-based modeling: Overview, validation, and emerging challenges. *IEEE Transactions on Artificial Intelligence*, 6(12):3165–3183, 2025. doi: 10.1109/TAI.2025.3566362.
- Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 337–371. PMLR, 2023.
- AI Evaluator Forum. AEF-1: Minimum operating conditions for independent third party AI evaluations. <https://aievaluatorforum.org/initiatives/minimum-operating-conditions>, 2025. Version 1, December 4, 2025. Founding members: Transluce, METR, RAND, the Holistic Agent Leaderboard at Princeton University, SecureBio, the Collective Intelligence Project, Meridian Labs and the AI Verification and Evaluation Research Institute; key provisions endorsed by the European AI Office.
- Airbus. Airbus achieves new commercial aircraft delivery record in 2018. <https://www.airbus.com/en/newsroom/press-releases/2019-01-airbus-achieves-new-commercial-aircraft-delivery-record-in-2018>, 2019. Press release, January 9, 2019.

- Airbus. Airbus 2020 deliveries demonstrate resilience. <https://www.airbus.com/en/newsroom/press-releases/2021-01-airbus-2020-deliveries-demonstrate-resilience>, 2021. Press release, January 8, 2021.
- Airbus. Airbus reports 766 commercial aircraft deliveries in 2024. <https://www.airbus.com/en/newsroom/press-releases/2025-01-airbus-reports-766-commercial-aircraft-deliveries-in-2024>, 2025. Press release, January 9, 2025.
- American Bar Association Standing Committee on Ethics and Professional Responsibility. Generative artificial intelligence tools. Formal Opinion 512, American Bar Association, 2024. URL https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf. July 29, 2024.
- Ross Anderson and Tyler Moore. The economics of information security. *Science*, 314(5799):610–613, 2006. doi: 10.1126/science.1130992.
- Anthropic. Anthropic’s Responsible Scaling Policy. <https://www.anthropic.com/news/anthropics-responsible-scaling-policy>, 2023. Internal capability-evaluation framework for frontier-model deployment; accessed 2026.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Rüdiger Bachmann, Gabriel Ehrlich, Ying Fan, Dimitrije Ruzic, and Benjamin Leard. Firms and collective reputation: A study of the Volkswagen emissions scandal. *Journal of the European Economic Association*, 21(2):484–525, 2023. doi: 10.1093/jeea/jvac046.
- Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.5.
- Joel Becker, Nate Rush, Beth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*, 2025.
- Alexander Bick, Adam Blandin, and David J. Deming. The rapid adoption of generative AI. *NBER Working Paper 32966*, 2024. September 2024 version.
- Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management. SR Letter 11-7, Board of Governors of the Federal Reserve System, 2011. URL <https://web.archive.org/web/20260113100323/https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>. April 4, 2011. Superseded by SR 26-2 on April 17, 2026.
- Board of the International Organization of Securities Commissions. Artificial intelligence in capital markets: Use cases, risks, and challenges. Consultation Report CR/01/2025, International Organization of Securities Commissions, 2025. URL <https://www.iosco.org/library/pubdocs/pdf/IOSCOPD788.pdf>.
- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- Rishi Bommasani, Dilara Soylu, Thomas I. Liao, Kathleen A. Creel, and Percy Liang. Ecosystem Graphs: Documenting the foundation model supply chain. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 196–209, 2024. doi: 10.1609/aies.v7i1.31629.
- J. Scott Brennan, Philip N. Howard, and Rasmus Kleis Nielsen. An industry-led debate: How UK media cover artificial intelligence. Factsheet, Reuters Institute for the Study of Journalism, University of Oxford, 2018. URL <https://reutersinstitute.politics.ox.ac.uk/our-research/industry-led-debate-how-uk-media-cover-artificial-intelligence>.

- Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.
- Mark Calaguas. 2024 artificial intelligence TechReport. https://www.americanbar.org/groups/law_practice/resources/tech-report/2024/2024-artificial-intelligence-techreport/, 2025. American Bar Association, Law Practice Division. Published April 25, 2025.
- California Department of Motor Vehicles. DMV statement on Cruise LLC suspension. <https://www.dmv.ca.gov/portal/news-and-media/dmv-statement-on-cruise-llc-suspension/>, 2023. Press release, October 24, 2023.
- Donald T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1): 67–90, 1979. doi: 10.1016/0149-7189(79)90048-X.
- Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146, 2026. doi: 10.1038/s41586-025-09962-4.
- Varun Chandrasekaran, Hengrui Jia, Anvith Thudi, Adelin Travers, Mohammad Yaghini, and Nicolas Papernot. SoK: Machine learning governance. *arXiv preprint arXiv:2109.10870*, 2021.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning (ICML)*, volume 235, pages 8359–8388, 2024.
- François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Ching-Hua Chuan, Wan-Hsiu Sunny Tsai, and Su Yeon Cho. Framing artificial intelligence in American newspapers. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 339–344. ACM, 2019. doi: 10.1145/3306618.3314285.
- Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1186–1197, 2023. doi: 10.1145/3593013.3594073.
- Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583. ACM, 2022.
- Logan Cross, Nick Haber, and Daniel L. K. Yamins. Validating generative agent-based models of social norm enforcement: From replication to novel predictions. In *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, pages 6060–6066, 2025. URL <https://escholarship.org/uc/item/6dglf4s7>.
- Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.
- Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations (ICLR)*, 2025.
- Uwe Dulleck and Rudolf Kerschbamer. On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1):5–42, 2006.
- ECRI. Top 10 patient safety concerns 2025. <https://home.ecri.org/blogs/guidance-insights-tools/top-10-patient-safety-concerns-2025>, 2025. Published March 10, 2025.

- Madeleine Clare Elish and danah boyd. Situating methods in the magic of big data and AI. *Communication Monographs*, 85(1):57–80, 2018. doi: 10.1080/03637751.2017.1375130.
- Ezekiel J Emanuel, David Wendler, and Christine Grady. What makes clinical research ethical? *JAMA*, 283(20):2701–2711, 2000. doi: 10.1001/jama.283.20.2701.
- Epoch AI. Data on AI capabilities and benchmarking. <https://epoch.ai/benchmarks>, 2024. Benchmark performance database; accessed 2026.
- European Banking Authority. AI Act: Implications for the EU banking and payments sector. <https://www.eba.europa.eu/sites/default/files/2025-11/d8b999ce-a1d9-4964-9606-971bbc2aaf89/AI%20Act%20implications%20for%20the%20EU%20banking%20sector.pdf>, 2025. Published November 21, 2025.
- European Parliament and Council. Regulation (EU) No 376/2014 of the European Parliament and of the Council of 3 April 2014 on the reporting, analysis and follow-up of occurrences in civil aviation. Official Journal of the European Union, 2014. URL <http://data.europa.eu/eli/reg/2014/376/oj>. OJ L 122, 24.4.2014, p. 18.
- European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, 2024. URL <http://data.europa.eu/eli/reg/2024/1689/oj>. OJ L, 2024/1689, 12.7.2024.
- Ethan Fast and Eric Horvitz. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI’17)*, pages 963–969. AAAI Press, 2017. doi: 10.1609/aaai.v31i1.10635.
- Financial Stability Oversight Council. 2024 annual report. Technical report, U.S. Department of the Treasury, 2024. URL <https://home.treasury.gov/system/files/261/FSOC2024AnnualReport.pdf>.
- Lawrence M Friedman, Curt D Furberg, David L DeMets, David M Reboussin, and Christopher B Granger. *Fundamentals of Clinical Trials*. Springer, fifth edition, 2015.
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S³: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.
- Gartner, Inc. Gartner predicts agentic AI will autonomously resolve 80% of common customer service issues without human intervention by 2029. <https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-predicts-agentic-ai-will-autonomously-resolve-80-percent-of-common-customer-service-issues-without-human-intervention-by-20290>, 2025. Press release, March 5, 2025.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- General Motors. GM to refocus autonomous driving development on personal vehicles. <https://investor.gm.com/news-releases/news-release-details/gm-refocus-autonomous-driving-development-personal-vehicles/>, 2024. Press release, December 10, 2024.
- Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review*, 40(1):e1761, 2024.
- Avijit Ghosh, Anka Reuel, Jenny Chim, Wm. Matthew Kennedy, Srishti Yadav, Jennifer Mickel, Yanan Long, Andrew Tran, Anastassia Kornilova, Damian Stachura, et al. Evaluation Cards: An interpretive layer for AI evaluation reporting. *arXiv preprint arXiv:2606.09809*, 2026.

- Google DeepMind. Introducing the Frontier Safety Framework. <https://deepmind.google/blog/introducing-the-frontier-safety-framework/>, 2024. Internal capability-evaluation framework for frontier-model deployment; accessed 2026.
- Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310(5750):987–991, 2005.
- Jacob Haimen, Cenny Wenner, Kunvar Thaman, Vassil Tashev, Clement Neo, Esben Kran, and Jason Schreiber. Benchmark inflation: Revealing LLM performance gaps using retro-holdouts. *arXiv preprint arXiv:2410.09247*, 2024.
- Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel John Kochenderfer. More than marketing? On the information value of AI benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1032–1047. ACM, 2025. doi: 10.1145/3708359.3712152.
- Michael Hardy, Anka Reuel, Lijin Zhang, Jodi M. Casabianca, Sang Truong, Yash Dave, Hansol Lee, Benjamin Domingue, and Sanmi Koyejo. AI cartography: Mapping the latent landscape of AI benchmark ecosystems. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 40570–40613, 2026.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.
- Tanya Albert Henry. 2 in 3 physicians are using health AI—up 78% from 2023. <https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023>, 2025. American Medical Association, February 26, 2025.
- Julian P. T. Higgins and Sally Green, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley, 2008. doi: 10.1002/9780470712184.
- Aspen Hopkins, Sarah H Cen, Isabella Struckman, Andrew Ilyas, Luis Videgaray, and Aleksander Mądry. AI supply chains: An emerging ecosystem of AI actors, products, and services. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1266–1277, 2025. doi: 10.1609/aies.v8i2.36628.
- Jason C. Hsu, Dennis Ross-Degnan, Anita K. Wagner, Fang Zhang, and Christine Y. Lu. How did multiple FDA actions affect the utilization and reimbursed costs of thiazolidinediones in US Medicaid? *Clinical Therapeutics*, 37(7):1420–1432.e1, 2015. doi: 10.1016/j.clinthera.2015.04.006.
- Ben Hutchinson, Andrew Smart, Alex Hanna, Remi Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 560–575. ACM, 2021. doi: 10.1145/3442188.3445918.
- Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019.
- Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.
- Antoine Juqueler, Ingrid Poncin, and Simon Hazée. Empathic chatbots: A double-edged sword in customer experiences. *Journal of Business Research*, 188:115074, 2025. doi: 10.1016/j.jbusres.2024.115074.
- Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.
- Martin Kenney and John Zysman. Unicorns, Cheshire cats, and the new dilemmas of entrepreneurial finance. *Venture Capital*, 21(1):35–50, 2019. doi: 10.1080/13691066.2018.1517430.

- Bernard Koch, Remi Denton, Alex Hanna, and Jacob G. Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.
- Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2026. doi: 10.1007/s10462-025-11412-6.
- Averill M. Law. *Simulation Modeling and Analysis*. McGraw-Hill Education, 5th edition, 2015.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.
- Luona Lin. About 1 in 5 U.S. workers now use AI in their job, up since last year. <https://www.pewresearch.org/short-reads/2025/10/06/about-1-in-5-us-workers-now-use-ai-in-their-job-up-since-last-year/>, 2025. Pew Research Center, October 6, 2025.
- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.
- Robert E. Lucas, Jr. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46, 1976.
- Donald MacKenzie. *An engine, not a camera: How financial models shape markets*. MIT Press, 2006.
- Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.18.
- Christos Makridis. AI adoption rapidly growing in public sector. <https://www.gallup.com/workplace/702983/adoption-rapidly-growing-public-sector.aspx>, 2026. Gallup, March 10, 2026.
- David Manheim and Scott Garrabrant. Categorizing variants of Goodhart’s law. *arXiv preprint arXiv:1803.04585*, 2018.
- Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientist and subjects. *NBER Working Paper 32381*, 2024.
- Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial Intelligence Index Report 2024. Technical report, Stanford Human-Centered Artificial Intelligence, 2024. arXiv:2405.19522.
- Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Toby Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. The AI Index 2025 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025. arXiv:2504.07139.
- Donella H Meadows. *Thinking in systems: A primer*. Chelsea Green Publishing, 2008.
- Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly Buchanan, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://www.tbench.ai/>.

- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.
- Tyler Moore. The economics of cybersecurity: Principles and policy options. *International Journal of Critical Infrastructure Protection*, 3(3-4):103–117, 2010.
- National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.
- National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024.
- New York City Council. A local law to amend the administrative code of the city of New York, in relation to automated employment decision tools. <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9>, 2021. Local Law 144 of 2021 (Int. 1894-2020), enacted December 11, 2021.
- Steven E. Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, 356(24):2457–2471, 2007. doi: 10.1056/NEJMoa072761.
- NYC Department of Consumer and Worker Protection. Automated employment decision tools (updated). <https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-updated/>, 2023. Adopted rule, effective July 5, 2023.
- Office of Management and Budget. Accelerating federal use of AI through innovation, governance, and public trust. Memorandum M-25-21, Executive Office of the President, 2025a. URL <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf>. April 3, 2025.
- Office of Management and Budget. Driving efficient acquisition of artificial intelligence in government. Memorandum M-25-22, Executive Office of the President, 2025b. URL <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf>. April 3, 2025.
- OpenAI. Preparedness Framework (Beta). <https://cdn.openai.com/openai-preparedness-framework-beta.pdf>, 2023. Published December 18, 2023; Version 2 released April 15, 2025.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM, 2023.
- Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünnér, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- Michael E Porter. *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press, 1985.
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020. doi: 10.1145/3351095.3372873.
- Inioluwa Deborah Raji, Remi Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the Everything in the Whole Wide World Benchmark. In *NeurIPS 2021 Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=j6NxpQbREA1>.

- Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571. ACM, 2022.
- Responsible AI Collaborative. AI Incident Database. <https://incidentdatabase.ai>, 2021. Accessed 2025.
- Moowoon Rhee and Pamela R Haunschild. The liability of good reputation: A study of product recalls in the U.S. automobile industry. *Organization Science*, 17(1):101–117, 2006.
- Oscar Sainz, Jon Ander 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. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, 2023.
- Robert G. Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1):12–24, 2013. doi: 10.1057/jos.2012.20.
- Scale AI. Scale’s SEAL research lab launches expert-evaluated LLM leaderboards. <https://scale.com/blog/leaderboard>, 2024. Published May 29, 2024.
- Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.
- Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D’souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermiş, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 86910–86964, 2025. doi: 10.52202/085713-2620. Preprint at arXiv:2504.20879.
- Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. The state of AI: How organizations are rewiring to capture value. Technical report, McKinsey & Company, QuantumBlack, 2025. URL <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value>. Published March 12, 2025.
- Peter Slattery, Alexander K. Saeri, Emily A.C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, 7(5):101517, 2026. doi: 10.1016/j.patter.2026.101517.
- Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N. Nadkarni, and Eyal Klang. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5), 2024. doi: 10.1056/AIdbp2300040.
- S&P Global Market Intelligence. Generative AI market revenue projected to grow at a 40% CAGR from 2024–2029. <https://www.spglobal.com/market-intelligence/en/news-insights/research/generative-ai-market-revenue-projected-to-grow-at-a-40-cagr-from-2024-2029>, 2025. 451 Research, Generative AI Market Monitor & Forecast. Published June 3, 2025.
- Stack Overflow. 2024 Stack Overflow developer survey: AI. <https://survey.stackoverflow.co/2024/ai>, 2024. Survey fielded May 19 to June 20, 2024.
- Stack Overflow. 2025 Stack Overflow developer survey: AI. <https://survey.stackoverflow.co/2025/ai>, 2025. Results published July 29, 2025.
- Patrick Taillandier, Jean-Daniel Zucker, Arnaud Grignard, Benoit Gaudou, Nghi Quang Huynh, Haojia Kong, and Alexis Drogoul. From the fluency fallacy to the micro-to-macro validity gap: Opportunities and pitfalls of LLMs in social simulation. *arXiv preprint arXiv:2507.19364*, 2025.
- The Boeing Company. Boeing reports fourth-quarter deliveries. <https://investors.boeing.com/investors/news/press-release-details/2019/Boeing-Reports-Fourth-Quarter-Deliveries/default.aspx>, 2019. Press release, January 8, 2019.

- The Boeing Company. Boeing announces fourth-quarter deliveries. <https://boeing.mediaroom.com/2021-01-12-Boeing-Announces-Fourth-Quarter-Deliveries>, 2021. Press release, January 12, 2021.
- The Boeing Company. Boeing announces fourth quarter deliveries. <https://boeing.mediaroom.com/2025-01-14-Boeing-Announces-Fourth-Quarter-Deliveries>, 2025. Press release, January 14, 2025.
- The White House. Executive Order on the safe, secure, and trustworthy development and use of artificial intelligence (EO 14110). Federal Register, vol. 88, pp. 75191–75226, November 1, 2023. URL <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.
- Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5): 100476, 2022. doi: 10.1016/j.patter.2022.100476.
- Thomson Reuters. 2025 generative AI in professional services report. <https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2025-generative-ai-in-professional-services-report-tr5433489-rgb.pdf>, 2025.
- U.S. Attorney’s Office, Northern District of California. Cruise admits to submitting a false report to influence a federal investigation and agrees to pay \$500,000. <https://www.justice.gov/usao-ndca/pr/cruise-admits-submitting-false-report-influence-federal-investigation-and-agrees-pay>, 2024. Press release, November 14, 2024.
- U.S. Equal Employment Opportunity Commission. Select issues: Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures under Title VII of the Civil Rights Act of 1964. Technical Assistance EEOC-NVTA-2023-2, U.S. Equal Employment Opportunity Commission, 2023. URL <https://web.archive.org/web/20240527151347/https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial>. Issued May 18, 2023; since removed from eeoc.gov (archived copy).
- U.S. Food and Drug Administration. Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. Draft guidance for industry and food and drug administration staff, U.S. Department of Health and Human Services, 2025. URL <https://www.fda.gov/media/184856/download>. Docket FDA-2024-D-4488. Issued January 7, 2025.
- U.S. Government Accountability Office. Artificial intelligence: Federal efforts guided by requirements and advisory groups. Technical Report GAO-25-107933, U.S. Government Accountability Office, 2025. URL <https://www.gao.gov/products/gao-25-107933>. September 9, 2025.
- Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4):97–112, 2021. doi: 10.9785/cr-2021-220402.
- Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. *arXiv preprint arXiv:2312.03664*, 2023.
- Walton Family Foundation and Gallup. Teaching for tomorrow: Unlocking six weeks a year with AI. <https://www.gallup.com/file/analytics/691922/Walton-Family-Foundation-Gallup-Teachers-AI-Report.pdf>, 2025. Published June 2025.
- 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, Siddhartha 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. Spotlight; preprint at arXiv:2406.19314.

- Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 195–200, 2019. doi: 10.1145/3306618.3314289.
- David Gray Widder and Dawn Nafus. Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society*, 10(1), 2023. doi: 10.1177/20539517231177620.
- Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. Position: LLM-based social simulations require a boundary. In *Proceedings of the 43rd International Conference on Machine Learning (Position Paper Track)*, volume 306, pages 172625–172648, 2026.
- Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.
- Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. *arXiv preprint arXiv:2509.18052*, 2025.

A Extended Related Work

This appendix extends §2 with five topics: governance, public-goods and performativity perspectives; the supply-chain and value-chain framings of AI development; the literature on generative agent-based models and their validation; the reasons for using a simulation and what other methods offer; and governance regimes in other industries.

A.1 Governance, public goods, and performativity

Our public-goods framing draws on economic analyses of information security [Anderson and Moore, 2006, Moore, 2010] and is informed by work on AI ethics and governance [Chandrasekaran et al., 2021, Whittlestone et al., 2019, Veale and Zuiderveen Borgesius, 2021, Brundage et al., 2020]. Documentation and audit proposals such as model cards [Mitchell et al., 2019], datasheets [Gebru et al., 2021], internal audits [Raji et al., 2020] and, more recently, structured reporting of evaluation results themselves [Ghosh et al., 2026], along with the EU AI Act [European Parliament and Council, 2024] and the voluntary NIST AI RMF [National Institute of Standards and Technology, 2023], attach obligations or guidance to individual models, datasets, systems and evaluations. They do not model the feedback loops that undermine evaluation integrity once evaluation becomes a competitive signal. Systems-thinking [Meadows, 2008] and performativity [MacKenzie, 2006] perspectives let us study how measurement and optimization co-evolve at ecosystem scale.

A.2 AI supply chains, value chains, and the evaluation layer

Two framings dominate recent analysis of AI development as a system. The *supply-chain* framing [Hopkins et al., 2025, Cobbe et al., 2023, Widder and Nafus, 2023] treats AI production as a network of specialized actors (foundation-model developers, dataset curators, compute providers, fine-tuners, application integrators) through which models and services flow from upstream suppliers to downstream users; its concerns are accountability routing, dislocated responsibility, and the modular structure that allows harms to fall between nodes. The *value-chain* framing [Porter, 1985] disaggregates a firm into value-creating activities linked to its suppliers and buyers; its concerns are cost position, competitive advantage, and where in the chain value is created. Both framings describe a linear flow, in which a product, a service or value moves from node to node.

Evaluation does not fit either framing. Benchmark scores are published artifacts that several classes of actor use in parallel, and no supplier hands them to a customer. Evaluators are also typically not paid in proportion to the value that their measurements create for downstream actors. We therefore treat evaluation

as shared measurement infrastructure (§2). Every actor in the supply-chain framing depends on this layer, and every step in the value-chain framing is priced against it. Bommasani et al. [2024] map foundation-model dependencies across the network in their Ecosystem Graphs, but they focus on the social footprint of models and not on the measurement infrastructure that coordinates behavior around them.

The failure modes that motivate this paper (score inflation, benchmark gaming, evaluator conflict of interest, and capital that follows gamed signals) do not belong to a single supply-chain node or value-chain step. They arise when the actors that a measurement layer measures also optimize against it. Appendix H.1 studies one case, in which the evaluator sells extra submissions and early access to the providers its scores.

A.3 Generative agent-based modeling with LLMs: extended discussion

A central design question in generative agent-based modeling is how much behavior to delegate to LLM actors and how much to encode in rules; Adornetto et al. [2025] advocate hybrid ABM–GABM approaches. Critical reviews argue that LLM agents make classical ABM challenges worse: their black-box nature, stochasticity and cultural biases make results harder to understand, replicate and ground empirically [Larooij and Törnberg, 2026], and they tend to converge toward an “average persona” [Taillandier et al., 2025]. Wu et al. [2026] review studies that assess behavioral variance, and most of those studies report lower variance than human populations show. They propose that claims be limited to collective-level qualitative patterns when variance is insufficient. Our design follows from these concerns. LLM actors make only strategic planning decisions (investment allocation, regulatory choice, funding allocation), and rules determine market mechanics, scoring and state transitions. We also check the behavioral coherence of LLM actors (Appendix E.6), and we make claims about patterns and not about point estimates (§5).

A growing body of work uses LLM-powered agents as participants in social simulations, but validation methodology remains nascent. Park et al. [2023] evaluate generative agents by interviewing the agents and having human evaluators rank their believability across component ablations. Aher et al. [2023] validate by replicating classic human-subject experiments (Ultimatum Game, Milgram) and comparing aggregate statistics against published human data, uncovering systematic “hyper-accuracy” distortions. Gao et al. [2023] validate a social-network simulator at both micro (individual next-state prediction) and macro (aggregate pattern matching) levels against real social-media traces. Ghaffarzadegan et al. [2024] use prompt-sensitivity sweeps as a robustness check for generative agent-based models, and Manning et al. [2024] ground LLM simulations in structural causal models that enable formal hypothesis testing. Cross et al. [2025] propose a two-stage protocol for GABM validation, which first replicates established experimental findings and then generates novel predictions. Critical reviews still identify convergence toward a generic mean that suppresses behavioral heterogeneity, and sensitivity to prompt formulation and model version [Taillandier et al., 2025]. Our validation protocol (Appendix E) uses the terminating-simulation inference framework of Law [2015] and the verification-and-validation framework of Sargent [2013], and it adds coherence checks that are specific to LLM actors.

A.4 Why a simulation, and what other methods offer

We use a stylized simulation as a controlled environment. Ie et al. [2019] take the same position for recommender systems: they build simulators that mirror specific aspects of user behavior in order to develop, evaluate and compare models, and they do not expect policies learned in simulation to be deployed in live systems. Our simulation does not predict the trajectory of real evaluation markets. It lets us test structural hypotheses about feedback loops, incentive misalignment and regulatory leverage by changing one part of the ecosystem at a time. Observational data cannot support these tests, because true capability and benchmark validity are not directly observable, and because interventions such as removing funders or the regulator have no real-world counterpart.

Other methods complement this one. A game-theoretic analysis could establish whether the behavior we observe is an equilibrium and could identify scoring rules that remove the incentive to game, but the multi-actor, multi-benchmark, multi-round structure resists closed-form solution. System dynamics makes feedback loops and leverage points explicit [Meadows, 2008], but it models aggregate quantities and cannot represent differences between providers. Observational studies, for example of the transition from MMLU to

MMLU-Pro, could test the hypotheses that the simulation generates where data exist. Expert elicitation could calibrate parameters and supply institutional knowledge that the simulation does not model.

A.5 Governance analogues in other industries

The public-goods framing in §2 draws on governance regimes in other high-stakes industries. Clinical research [Emanuel et al., 2000, Friedman et al., 2015] provides a mature model of independent review and trial monitoring for interventions whose private incentives are misaligned with public welfare. Civil aviation offers a second model, in which mandatory occurrence reporting and the exchange of occurrence information across organisations and authorities support safety [European Parliament and Council, 2014]. Both regimes use mechanisms that our simulation can test as interventions on the AI evaluation ecosystem: independent evaluation, mandatory disclosure, liability routing and information sharing across actors. The analogies are imperfect. AI capability changes quickly, which makes pre-deployment audit a moving target, and what an evaluation should measure is more contested than a clinical endpoint or an airworthiness standard. These regimes are still useful institutional examples of what a mature AI evaluation regime could look like.

B Full Simulation Architecture

This appendix specifies the actors, state representations and mechanics that Section 4 summarizes. Every reported run lasts 40 rounds, and one round represents one month. Table 3 defines the capability dimensions, Table 4 the benchmark suite and Table 5 the providers. Tables 6 and 7 give the consumer parameters, Table 8 the regulator levers, and Table 9 the funder parameters. Code: github.com/aims-foundations/evarium. Website: aimslab.stanford.edu/evaluation-ecosystem.

B.1 Round Protocol (Pseudocode)

Algorithm 1 gives the order of events in one round, and Figure 4 shows the same order as a diagram.

Algorithm 1 Simulation round protocol. Actors observe only public and own-private state; ground truth capability vectors $\mathbf{C}$ are held by the simulation and never exposed to any actor prompt.

Require: Providers $\mathcal{P}$, Evaluator $\mathcal{E}$, Benchmarks $\mathcal{B}$, ConsumerMarket $\mathcal{M}$, Regulator ρ , Funders $\mathcal{F}$, Media μ , rounds T , ground truth $\mathbf{C}$

```

1: for  $t = 1, \dots, T$  do
    Phase 1: Provider planning & capability update
2:   for each provider  $p \in \mathcal{P}$  do
3:      $\mathbf{o}_p \leftarrow \text{OBSERVE}(p, \mathcal{H}_{t-1})$  ▷ scores, market share, sanctions, incidents
4:      $\mathbf{x}_p \leftarrow \text{PLAN}(\mathbf{o}_p, \mathbf{s}_p^{\text{priv}})$  ▷ LLM or heuristic → portfolio + focus (ordinal)
5:      $\mathbf{c}_p \leftarrow \mathbf{c}_p + \Delta(\mathbf{x}_p, b_p)$  ▷ capability update via focus-weighted beliefs
6:   end for
    Phase 2: Evaluation & benchmark update
7:    $\hat{\mathbf{S}} \leftarrow \mathcal{E}.\text{ScoreAll}(\mathcal{P}, \mathbf{C}, \mathcal{B})$ ; publish leaderboard  $L_t$ 
    Phase 3: Provider belief update
8:   for each provider  $p \in \mathcal{P}$  do
9:     Update  $\hat{\mathbf{w}}_{p,b}$  from score prediction errors; reflect
10:  end for
    Phase 4: Incident generation & media coverage
11:   $I_t \leftarrow \text{GenerateIncidents}(\mathcal{P}, \mathbf{C}, \text{shares})$ 
12:   $m_t \leftarrow \mu.\text{Publish}(L_t, I_t, \text{public\_comms})$  ▷ media coverage
    Phase 5: Downstream actors
13:   $\mathcal{M}.\text{Update}(L_t, \mathbf{C}, m_t, I_t)$  ▷ satisfaction, switching
14:   $\rho.\text{Act}(I_t, m_t)$  ▷ graduated intervention
15:   $\mathcal{F}.\text{Allocate}(L_t, m_t, \rho, I_t)$  ▷ capital allocation
16: end for

```

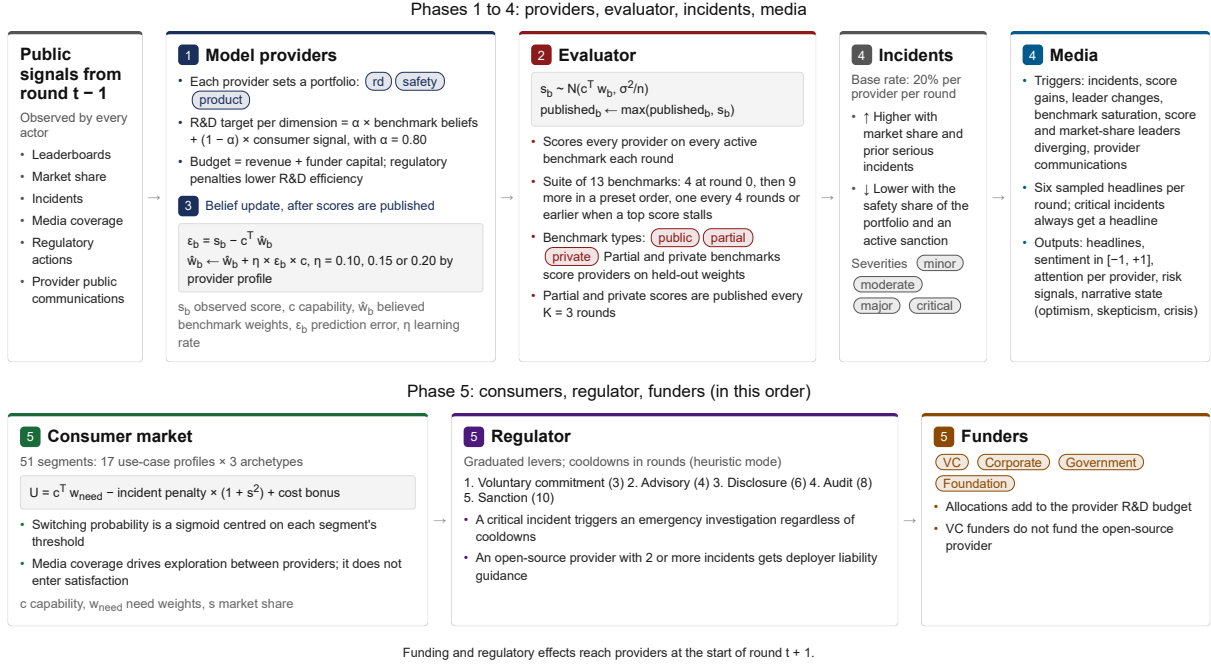


Figure 4: Order of events in one round, with the phase numbers of Algorithm 1. In Phases 1 to 4, providers plan with the public signals of the previous round, the evaluator scores them, providers update their beliefs, and incidents and media coverage follow. In Phase 5, consumers, the regulator and funders respond in that order to the round’s public state.

B.2 Capability Dimensions

The simulation represents model capability with six dimensions (Table 3). Each dimension is an area that providers can invest in separately and that public benchmarks measure. We do not assume that the dimensions are statistically independent.

Table 3: Capability dimensions, real-world anchors, and primary investment pathways.

Dimension	Represents	Benchmark anchor	Investment pathway
Reasoning	Problem solving, logic, multi-step inference	MMLU, GPQA	Reasoning post-training, CoT RLHF
Coding	Software engineering, debugging	HumanEval, MBPP	Code-specific data, correctness RL
Knowledge	Factual recall, domain expertise	MMLU subsets	Domain corpus, retrieval fine-tuning
Safety	Harmlessness, adversarial robustness	TruthfulQA	RLHF for harmlessness, red-teaming
Communication	Instruction following, fluency	MT-Bench, IFEval	Instruction-following RLHF
Agentic	Tool use, multi-step task execution	SWE-bench, WebArena	Trajectory RL, tool-use fine-tuning

B.3 Benchmark Pool and Introduction Schedule

Table 4: The default suite of 13 benchmarks in order of introduction, with each benchmark’s hidden dimension weights and privacy tier. Four benchmarks are active at round 0, and the evaluator introduces the other nine one at a time (`benchmark_introduction_cooldown=4`). The dynamic evaluator mode draws from a pool of 22 benchmarks that covers the same six dimensions. The privacy tier (**public**, **partial** or **private**) determines how scores are reported (Appendix C.4). The split of 8 public, 3 partial and 2 private benchmarks is a modeling choice. The partial tier is patterned on SEAL-style leaderboards and the private tier on FrontierMath-style holdouts. The columns $\text{cos}_{\text{part.}}$ and $\text{cos}_{\text{priv.}}$ give the realized cosine between the benchmark’s public weights and its holdout weights when the benchmark is **partial** and when it is **private** (Appendix C.4). The last column names the real benchmarks that each row is patterned on. The four round-0 benchmarks are patterned on MMLU [Hendrycks et al., 2021], HumanEval [Chen et al., 2021] and MBPP [Austin et al., 2021], TruthfulQA [Lin et al., 2022], and MT-Bench and IFEval.

Benchmark	Tier	Reas.	Code	Know.	Safe.	Comm.	Agent.	$\text{cos}_{\text{part.}}$	$\text{cos}_{\text{priv.}}$	Real analog
General Capability	public	0.35	0.08	0.30	0.05	0.20	0.02	0.97	0.89	MMLU
Coding Evaluation	public	0.10	0.78	0.03	0.01	0.02	0.06	0.99	0.93	HumanEval/ MBPP
Safety Evaluation	partial	0.03	0.01	0.05	0.80	0.10	0.01	0.99	0.92	TruthfulQA/BBQ
Instruction Following	public	0.08	0.02	0.05	0.04	0.80	0.01	0.98	0.86	MT-Bench/IFEval
Scientific Reasoning	partial	0.78	0.02	0.15	0.00	0.04	0.01	0.98	0.92	GPQA/MMLU-Pro
Clinical Reasoning	public	0.20	0.01	0.65	0.08	0.05	0.01	0.98	0.89	Med-PaLM/MedQA
Adversarial Robustness	private	0.06	0.01	0.01	0.85	0.04	0.03	0.98	0.88	SEAL-Safety
Hard Coding	partial	0.10	0.80	0.02	0.01	0.01	0.06	0.99	0.94	LiveCodeBench
Agentic Tasks	public	0.25	0.19	0.01	0.00	0.07	0.48	0.98	0.91	SWE-bench/BFCL
Advanced Math	private	0.85	0.06	0.05	0.00	0.03	0.01	0.99	0.91	FrontierMath
Function Calling	public	0.15	0.30	0.02	0.01	0.12	0.40	0.94	0.76	BFCL
Long Context	public	0.12	0.02	0.20	0.01	0.62	0.03	0.98	0.91	RULER/ LongBench-v2
Legal Reasoning	public	0.25	0.01	0.60	0.05	0.08	0.01	0.98	0.89	LegalBench-Pro

The population-average consumer need weights at round 0 are reasoning 0.18, coding 0.13, knowledge 0.18, safety 0.16, communication 0.27 and agentic 0.09 (Table 7). The share of organizational consumers grows during a run (Appendix B.7), and by round 40 the safety weight rises to about 0.22 and the communication weight falls to about 0.20. The mean loading across the 13 benchmarks is higher than the need weight for reasoning (about 0.26 against 0.18) and for coding (about 0.18 against 0.13). The mean loading is lower than the need weight for communication at round 0 (about 0.17 against 0.27) and for safety by round 40 (about 0.15 against 0.22). The four round-0 benchmarks alone weight coding and safety more than consumers do, and they weight knowledge, agentic and reasoning less. The weight on reasoning grows as Scientific Reasoning and Advanced Math enter the suite. A provider that follows benchmark scores therefore invests in a different mix of dimensions than consumers need.

Introduction schedule. The order of the nine scheduled benchmarks is fixed (Table 4) and moves from broad benchmarks to specialized ones. The order and the cadence are modeling choices. The evaluator introduces the next benchmark every four rounds, at rounds 4, 8, . . . , 36, unless the top score on an active benchmark stalls first. A top score stalls when it rises by less than 0.005 in each of three consecutive rounds. Published scores never decrease, so a stall can occur at any score level. Two rounds after a stall, the evaluator introduces the next scheduled benchmark early. At least one benchmark enters early in about two thirds of runs, and all 13 benchmarks are active by round 36 in every run. The per-benchmark results use the final

round, and they are the same when we restrict the runs to those with no early introduction or to those with at least one.

In the dynamic evaluator mode, the evaluator adds one benchmark from the 22-benchmark pool every four rounds (`benchmark_introduction_interval` = 4). The heuristic evaluator picks the pool benchmark that best covers the dimensions where consumer need most exceeds the coverage of the active suite, and the LLM evaluator chooses from the pool itself. When the active suite already holds 13 benchmarks (`max_benchmarks`), adding a benchmark retires the most saturated one, and the LLM evaluator can also retire a benchmark by choice. A stalled top score does not trigger an introduction in this mode, so the timing of new benchmarks does not depend on provider scores.

B.4 Provider Profiles

Table 5: Initial provider profiles: capability in each dimension, and the starting portfolio split across R&D, safety and product. Capability values range from 0.08 to 0.55, and agentic capability is low for every provider. We set the values by hand, informed by 2023 frontier-model benchmark results [Maslej et al., 2024].

Provider	Reas.	Code	Know.	Safe.	Comm.	Agent.	R&D	Safety	Prod.	Notes
Orion Labs	0.52	0.48	0.50	0.42	0.52	0.12	55%	15%	30%	GPT-3.5 frontier
Apex AI	0.48	0.40	0.46	0.55	0.48	0.10	60%	30%	10%	Safety lead (CAI)
Genesis Sys.	0.50	0.38	0.52	0.40	0.42	0.12	70%	15%	15%	Largest compute
Mirage AI	0.44	0.42	0.44	0.32	0.40	0.10	80%	10%	10%	Scale-first
Spark AI	0.36	0.40	0.32	0.28	0.34	0.10	65%	10%	25%	Specialization
OpenCore	0.38	0.42	0.35	0.25	0.30	0.08	75%	10%	15%	startup Open-source

The providers have fictional names so that an LLM planner cannot draw on the reputation of a real company. We map them to real providers (Orion Labs → OpenAI, Apex AI → Anthropic, Genesis Systems → Google DeepMind, Mirage AI → Meta AI, Spark AI → Mistral, OpenCore → DeepSeek) only when we compare simulation output with external data, and the mapping never appears in an actor prompt.

B.5 Provider Investment Model

Capability update rule. Each round, the simulation computes a provider’s R&D budget, converts it to an effective budget with two sources of diminishing returns, and then applies gains in each dimension. The budget is

$$b_p = \max\left(\text{share}_p \cdot r \cdot M_t \cdot (1 - \text{cost_advantage}_p) + \sum_f \phi_{f,p}, b_p^{\min}\right), \quad (1)$$

where $r = 5.0$ is `revenue_per_share`, M_t is the total market size ($M_0 = 1$, growing 3% per round), and $\phi_{f,p}$ is the allocation from funder f . The budget floor $b_p^{\min}$ is 1.0 for OpenCore and 0 for the closed providers. The effective budget is

$$\text{effective_budget}_p = \sqrt{b_p} \cdot e_p \cdot \max(0, 1 - \text{mean}(\mathbf{c}_p)).$$

The square root gives diminishing returns on funding, and the last factor reduces gains as mean capability approaches the ceiling of 1. The efficiency e_p equals `rnd_efficiency` (0.1) multiplied by any active regulatory penalty from Table 8.

The provider splits its R&D across dimensions with a target that blends two signals: the dimensions that its benchmark beliefs favor, and the dimensions that its consumers value.

$$\text{target}[\text{dim}] = \alpha \cdot \text{benchmark_driven}[\text{dim}] + (1 - \alpha) \cdot \text{effective_signal}[\text{dim}] \quad (2)$$

The blend weight α is `benchmark_orientation` (fixed at 0.80). The consumer signal is reliable only when the provider invests in product:

$$\text{quality} = \frac{1}{1 + e^{-3(\text{product_budget} - 0.5)}} \quad (3)$$

$$\text{effective_signal}[\text{dim}] = \text{quality} \cdot \text{true_signal}[\text{dim}] + (1 - \text{quality}) \cdot \text{uniform}[\text{dim}]$$

Here `product_budget` is the provider’s absolute product budget, which is the product share of the portfolio multiplied by b_p . The quality is capped at 0.40 for the open-source provider (Appendix B.11).

Safety lever. The safety share of the budget raises only the safety dimension. The gain is the safety budget multiplied by $(1 - c_p[\text{safety}])$, so gains shrink as safety capability rises, and by an efficiency drawn each round from `Uniform(0.3, 0.9)`.

Allocation smoothing. The simulation applies each provider’s portfolio (R&D, safety, product) as the average of its last three chosen portfolios, in both heuristic and LLM mode. The three-round window is a modeling choice that corresponds to one quarter of monthly rounds. It keeps the score-prediction error of a single round from moving the whole budget.

Consumer signal. Each provider receives a six-dimensional vector that represents what its current users value. The simulation derives the vector from the need weights of the provider’s consumers, weighted by market share, and adds noise with standard deviation $\sigma = 0.05/\sqrt{\text{share}_p}$. Providers with a larger share therefore receive a more precise signal.

Belief update. Each round, a provider updates its beliefs about each benchmark’s hidden dimension weights from its score prediction error:

$$\hat{w}_{p,b}[\text{dim}] += \eta \cdot (\hat{s}_{p,b} - \mathbf{c}_p^\top \hat{\mathbf{w}}_{p,b}) \cdot c_p[\text{dim}] \quad (4)$$

After each update the simulation clips the weights at zero and renormalizes them to sum to 1. The learning rate η depends on the provider’s strategy profile: 0.20 for aggressive or competitive profiles, 0.10 for safety-oriented or responsibility-oriented profiles, and 0.15 otherwise. The simulation runs the belief update in both heuristic and LLM mode, because beliefs are state computed from observed score errors and an LLM planner does not output them.

B.6 Heuristic Provider Planning

In heuristic mode a deterministic rule set replaces the LLM planner. The rule set reads the same `ecosystem_context` as the LLM planner (the provider’s own market share and satisfaction, its recent incidents, interventions that target it, the media narrative state and its per-benchmark scores) and returns a portfolio allocation. A provider’s identity comes from its initial capability vector, starting portfolio, initial focus levels, strategy profile and open-source settings (Table 5, Appendix B.11). The rules below are the same for every provider, and no profile text such as “safety-first” enters them. The profile affects heuristic behavior only through the learning rate η .

Rules. Each round, the heuristic planner applies four rules to the portfolio of the previous round:

- **Incident pressure.** Each of the provider’s own incidents in the current round adds a shift that depends on severity (0.01 minor, 0.04 moderate, 0.08 major, 0.12 critical), and the total for the round is capped at 0.20. The shifts accumulate in a pressure value that is capped at 0.25. Each round the planner moves that amount from R&D to safety and then multiplies the pressure by 0.40, so the safety response to an incident decays over the following rounds.
- **Share trend.** The planner computes the change in the provider’s own share over three rounds. If the change is below -0.02 , it moves 0.03 from R&D to product. If the change is above $+0.02$, it moves 0.02 from product to R&D.

- **Crisis narrative.** If the media narrative state is **CRISIS**, the planner moves 0.04 from R&D to safety.
- **Own interventions.** If the regulator targeted the provider with k interventions over the last five rounds, the planner moves $\min(0.06, 0.02k)$ from R&D to safety.

Bounds. After it applies the rules, the planner clips the portfolio to $rd \in [0.10, 0.75]$, $safety \in [safety_floor, 0.70]$ and $product \in [0.05, 0.50]$, and renormalizes it to sum to 1. The safety *allocation* floor is 0.10 for the open-source provider and 0.15 for closed providers, and it limits the safety share of the portfolio. A separate safety *capability* floor (0.15 open-source, 0.25 closed) limits the ground-truth safety dimension (Appendix B.11).

Levers held fixed. The blend weight α (**benchmark_orientation**, Eq. 2) is fixed at 0.80 for every provider in all reported runs, in both heuristic and LLM mode. The per-benchmark priority **focus_level**, in $[0.1, 5.0]$, changes only in LLM mode. In heuristic mode each provider keeps its initial focus levels, which differ across providers and benchmarks, for the whole run. Heuristic results are therefore a structural baseline, and they cannot show a provider changing its priorities among benchmarks.

Submission count under evaluator capture. Appendix H.1 specifies the evaluator-capture condition, in which a provider pays for extra benchmark submissions. A heuristic provider spends up to 15% of its base revenue plus its discretionary budget on extra submissions each round. The LLM planner chooses the number of submissions itself, up to 10.

B.7 Consumer Market

The market has 51 consumer segments. Each of the 17 use-case profiles (11 individual, 6 organizational) is paired with the three behavioral archetypes of its consumer type (Table 6). The share of organizational consumers grows from 0.25 to 0.55 over a run along a logistic curve with its midpoint at round 18.

Table 6: Consumer archetypes and behavioral parameters.

Archetype	LB trust	Switch cost	Switch thresh.	Cost sens.
Leaderboard follower	0.85	0.05	0.10	0.15
Experience-driven	0.35	0.08	0.06	0.30
Cautious	0.50	0.20	0.18	0.20
Enterprise cautious	0.25	0.35	0.20	0.10
Enterprise growth	0.45	0.25	0.15	0.15
Enterprise established	0.35	0.40	0.18	0.08

Archetypes.

Need weights.

Satisfaction and switching. A segment’s satisfaction with a provider depends on experienced capability, incidents and cost. It has no term for benchmark gaming or for media coverage:

$$\text{sat}_{s,p} = \mathbf{c}_p^\top \mathbf{w}_s - \text{incident_penalty}_p \cdot (1 + \text{share}_p^2) + 0.15 \cdot \text{cost_sens}_s \cdot \text{cost_adv}_p$$

The incident penalty sums a severity weight for each of the provider’s incidents (0.01 minor, 0.05 moderate, 0.10 major, 0.20 critical), multiplied by 0.70^{age} , so the weight of an incident halves in about two rounds. The factor $(1 + \text{share}_p^2)$ makes an incident cost a dominant provider more.

A segment’s expected quality blends leaderboard scores with the quality it has experienced:

$$\text{expected}[p] = \text{lb_trust} \cdot \text{score_signal}[p] + (1 - \text{lb_trust}) \cdot \text{running_perceived_quality}[p]$$

A segment becomes more likely to switch as $\text{expected} - \text{sat}$ rises. The switching probability is a sigmoid centered on the segment’s switching threshold, and the threshold is raised after a recent switch.

Table 7: Consumer need weights by use-case profile. Bold indicates the dominant dimension. Need weights are assigned to each profile from the sector-level evidence listed in Appendix C, together with occupation-level adoption rates [Bick et al., 2024]. Population weights are adoption-weighted shares assigned from the same adoption data, with organizational adoption surveys [Maslej et al., 2025, Singla et al., 2025] as context. The Pop. column holds base shares that sum to 1.04; the simulation rescales them within each consumer type so that organizational segments hold 0.25 of the market at round 0, rising to 0.55 by round 40 (Appendix B.7). The Pop. avg row gives the resulting values at round 0.

Profile	Reas.	Code	Know.	Safe.	Comm.	Agent.	Pop.
software_dev	0.22	0.55	0.03	0.02	0.03	0.15	0.14
content_writer	0.10	0.02	0.20	0.03	0.63	0.02	0.08
legal	0.30	0.01	0.35	0.22	0.10	0.02	0.04
healthcare	0.12	0.02	0.28	0.48	0.08	0.02	0.05
finance	0.35	0.08	0.22	0.25	0.05	0.05	0.05
educator	0.15	0.02	0.28	0.08	0.45	0.02	0.07
customer_service	0.05	0.01	0.12	0.15	0.50	0.17	0.08
researcher	0.30	0.22	0.28	0.03	0.07	0.10	0.06
creative	0.12	0.02	0.10	0.03	0.65	0.08	0.06
marketing	0.12	0.02	0.18	0.03	0.55	0.10	0.07
service_worker	0.08	0.01	0.15	0.18	0.55	0.03	0.05
hospital_system	0.12	0.02	0.25	0.48	0.10	0.03	0.05
enterprise_finance	0.32	0.08	0.18	0.30	0.05	0.07	0.05
tech_startup	0.18	0.38	0.05	0.04	0.05	0.30	0.06
enterprise_legal	0.28	0.02	0.35	0.22	0.11	0.02	0.04
government_agency	0.12	0.02	0.22	0.48	0.12	0.04	0.05
enterprise_hr	0.20	0.02	0.20	0.35	0.18	0.05	0.04
Pop. avg	0.18	0.13	0.18	0.16	0.27	0.09	

Exploration churn. Each round, 5% of each provider’s share in a segment enters an exploration pool. Negative media sentiment raises the rate for a provider in proportion to the attention it receives: $\text{rate} = 0.05 + \text{attention}_p \times |\text{sentiment}| \times \text{lb_trust}_s$. Positive sentiment raises the rate for providers that receive little attention, by half as much, so their users explore toward the provider in the news. The simulation redistributes the pool across providers with a blend of experienced satisfaction and believed quality.

B.8 Media Actor

A single media outlet publishes each round after the evaluator scores providers and before consumers, the regulator and funders act. Coverage triggers are incidents of moderate or higher severity, large score gains, changes of leaderboard leader, benchmark saturation, and a divergence trigger. The divergence trigger fires when the score leader has differed from the market-share leader for three rounds and an incident occurs. Critical incidents always receive a headline. All other events compete for six sampled headline slots per round, and major and moderate incidents have higher sampling weights.

The outlet outputs headlines, a sentiment in $[-1, 1]$, an attention value in $[0, 1]$ for each provider, risk signals, and a narrative state (OPTIMISM, SKEPTICISM or CRISIS). The narrative state worsens when a cumulative count of incidents of moderate or higher severity exceeds 3 and then 8, or when the divergence trigger fires. The count decays by 10% in each round without an incident, and the state recovers after a quiet period.

B.9 Regulator

All reported runs use one regulator with incident-rate thresholds of 0.05, 0.15 and 0.25 for its low, medium and high responses, and with advisory audits that cannot block deployment. The code also includes a stricter and a more permissive parameter set, which no reported run uses.

Table 8: Regulator intervention levers with real-world analogs, cooldowns, and mechanical effects. The cooldowns apply in heuristic mode; in LLM mode the code enforces only a three-round gap after any intervention.

Lever	Real analog	Cool.	Effect
Voluntary commitment	WH/Seoul pledges	3 r	Safety <code>public_comms</code> boost (+0.15, decaying)
Publish advisory	AISI summaries, NIST RMF	4 r	Enterprise and cautious archetypes reduce LB trust ($\times 0.90$, floor 0.15)
Safety disclosure	EU AI Act Art. 53	6 r	Safety comms boost (+0.20); <code>rnd_efficiency</code> $\times 0.95$
Commission audit	AISI pre-deployment eval	8 r	<code>rnd_efficiency</code> $\times 0.85$ for 3 rounds
Impose sanction	EU AI Act fines (7%)	10 r	<code>rnd_efficiency</code> penalty + funder allocation $\times 0.90$
Emergency investigation	Cruise robotaxi shutdown	0 r	Critical-incident override; immediate intervention regardless of cooldowns
Deployer liability guidance	OS-deployer policy	0 r	Parallel track for open-source providers with ≥ 2 incidents; assigns deployer-side liability

B.10 Funders

Four funder types allocate capital across providers from fixed pools. Each type scores providers with the formula in Table 9, where $m = 1 + 0.5 \cdot \text{sentiment}$ is a media factor.

Table 9: Funder types and allocation logic.

Type	Pattern	Funds OS?	Scoring formula
VC	Concentrated	No	$\text{share_growth} \times m \times (1 - \text{incident_risk}) \times (1 - \text{share}) \times \text{diversification}$
Corporate	Multi-relationship	Yes	$\text{share} \times m \times (1 - \text{incident_risk})$
Government	Spread, safety-oriented	Yes	$(1 - \text{incident_rate}) \times \text{share_growth}$
Foundation	Ecosystem health	Yes	$(1 - \text{incident_rate}) \times \text{share_growth}$

Funder allocations enter the provider budget through Eq. 1.

B.11 Open-Source Provider Mechanics

OpenCore differs from the closed providers in five ways:

- VC funders do not fund it. Corporate, government and foundation funders can.
- Its cost advantage is 0.9, which represents open weights that users can host themselves.
- Its safety capability floor is 0.15, against 0.25 for closed providers.
- Its product signal quality is capped at 0.40, which represents limited deployment telemetry.
- Its open weights reveal information about benchmarks. Each round, the simulation moves every other provider’s beliefs about each benchmark toward that benchmark’s public weights by a fraction equal to 0.30 times OpenCore’s market share.

Appendix E.4 audits these differences.

C Design Rationale and Parameter Sources

This appendix gives the simulation’s parameters and the reasons for its main design choices. Table 10 lists the hyperparameters, Table 11 lists related literature, and Table 12 rates each assumption behind the holdout-design result as calibrated, anchored or stipulated. Five parameters rest on external data:

- The reporting lag for private benchmarks, $K=3$, is calibrated to release cadence across 20 Epoch AI benchmarks and 8 frontier labs (Appendix C.5).
- The target cosines for holdout weights, 0.95 and 0.85, are anchored to retro-holdout score inflation [Haimes et al., 2024] and to within-family Pearson correlations that we compute from Epoch AI data [Epoch AI, 2024]. The hand-authored holdout vectors realize these targets only approximately (Appendix C.4).
- Starting capability vectors are informed by 2023 benchmark results in Maslej et al. [2024].
- Incident base rates are a modeling choice informed by incident reports in the AI Incident Database [Responsible AI Collaborative, 2021].
- Consumer need weights are assigned to each use-case profile from sector-level evidence on what that profession uses AI for and what it is concerned about, together with occupation-level adoption rates [Bick et al., 2024] and worker adoption surveys [Lin, 2025]: developers [Stack Overflow, 2024, 2025, Becker et al., 2025]; legal [Thomson Reuters, 2025, Calaguas, 2025, American Bar Association Standing Committee on Ethics and Professional Responsibility, 2024]; healthcare [Henry, 2025, U.S. Food and Drug Administration, 2025, ECRI, 2025, Soroush et al., 2024]; finance [Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011, Financial Stability Oversight Council, 2024, Board of the International Organization of Securities Commissions, 2025, European Banking Authority, 2025]; education [Walton Family Foundation and Gallup, 2025]; creative work [Adobe, 2025]; customer service [Gartner, Inc., 2025, Juquelier et al., 2025]; government [National Institute of Standards and Technology, 2024, Office of Management and Budget, 2025a,b, U.S. Government Accountability Office, 2025, Makridis, 2026]; and human resources [U.S. Equal Employment Opportunity Commission, 2023, New York City Council, 2021, NYC Department of Consumer and Worker Protection, 2023, European Parliament and Council, 2024]. Population weights draw on the occupation-level adoption data, with organizational adoption surveys [Maslej et al., 2025, Singla et al., 2025] as context.

Every other value is a modeling choice, and the tables say so.

C.1 Hyperparameter summary

Table 10: Summary of simulation hyperparameters.

Parameter	Value	Parameter	Value
<i>Ecosystem structure</i>		<i>Provider initialization</i>	
Providers	6	Capability dimensions	6
Benchmark pool (total)	22 (max 13 active)	Init. capability (range)	[0.08, 0.55]
Consumer segments	51	Benchmark orientation	0.80 (fixed)
Funder types	4	Belief learning rate	0.10 / 0.15 / 0.20 (by profile)
Simulation window	40 rounds (≈ 3.3 yr)		
<i>Scoring model</i>		<i>Regulator</i>	
Score formula	$\max(\hat{s}_{t-1}, \mathcal{N}(\mathbf{c}^\top \mathbf{w}_b, \sigma_b^2/n_b))$	Threshold (low)	0.05
Benchmark weights	Hidden from all actors	Threshold (med)	0.15
Budget scaling	$\sqrt{\text{rd_budget}}$	Threshold (high)	0.25
Holdout fraction h	0.3 partial / 1.0 private	Audits	Advisory
<i>Incident model</i>		<i>Open-source provider</i>	
Base incident rate	0.20 / provider / round	Cost advantage	0.9
Severity: min/mod/maj/crit	50%/31%/12%/7%	Safety capability floor	0.15 (vs. 0.25 closed)

Table 11: Related literature for simulation parameters and design choices. Each row gives the parameter or design choice and the literature it draws on; a value not stated in a cited source is a modeling choice.

Parameter / Design Choice	Related literature
<i>Parameter values</i>	
Market growth rate, 0.03/mo	S&P Global Market Intelligence [2025] (40% CAGR)
Consumer population weights, per-archetype (assumed)	Bick et al. [2024] (occupation-level adoption)
Leaderboard trust differs across users, 0.25–0.85 (assumed)	Hardy et al. [2025]
Incident base rates, 20% / provider / round (assumed)	Responsible AI Collaborative [2021]
<i>Structural design choices</i>	
Exploration churn = media-driven	modeling choice
Satisfaction = experience-based	modeling choice
Quantified AI outputs as partial signals	Elish and boyd [2018]
Benchmark entrenchment	Hutchinson et al. [2021]; Koch et al. [2021]
Media = event-driven amplifier	Fast and Horvitz [2017]; Chuan et al. [2019]; Brennen et al. [2018]
Regulator: information-constrained	modeling choice
VC funder: growth > profitability	Kenney and Zysman [2019]
Gov vs. corporate funder logic	modeling choice
Evaluator independence / access	Costanza-Chock et al. [2022]; Raji et al. [2022]; AI Evaluator Forum [2025]

C.2 Assumption status

Table 12 rates each assumption the holdout-design result rests on. *Calibrated* means fit to external data; *anchored* means the qualitative structure is supported by cited literature or data; *stipulated* means a modeling choice with a rationale and no direct empirical support. The conclusions of §5 are conditional on these assumptions.

Suite composition. We do not fit Table 4 against external estimates of real benchmarks’ dimension loadings; producing such estimates is a separate measurement problem [Hardy et al., 2026]. We can describe which dimensions the suite covers. By highest-loading dimension the 13 benchmarks divide into three on reasoning and two on each other dimension. An informal comparison against benchmark papers on arXiv suggests that this composition over-represents coding, under-represents knowledge, and omits multimodal benchmarks, which the text-only ontology excludes. The comparison concerns coverage only; the per-benchmark loading vectors remain stipulated.

C.3 Product Investment Design

Product investment works through two channels, and both depend on the provider’s absolute product budget B . First, a larger budget makes the consumer need signal that guides R&D more reliable: $\text{quality} = 1/(1 + e^{-3(B-0.5)})$, capped at 0.40 for the open-source provider (Appendix B.5). Second, a larger budget

Table 12: Status of the assumptions behind the holdout-design result.

Assumption	Status	Support and sensitivity
Six actor types cover the ecosystem, with heterogeneity inside each type	Stipulated	Tractability; excluded actors and levers in Appendix G.2.
Consumers choose through a mix of leaderboard signal and experienced quality	Anchored (trust differs across users); stipulated (trust values, mixture weights)	Hardy et al. [2025]; values in Appendix B.7.
Gaps are measured against the need weights of a mixed consumer population	Anchored (which needs dominate each profile, from sector evidence); stipulated (the exact values)	Sector evidence listed at the start of this appendix and Bick et al. [2024]; need-weights set the starting gap and its sign, so which shifts count as widening (§5.3, item 4).
Six capability dimensions with uneven benchmark coverage	Anchored	§5.3, item 2; the ontology is text-only (see below).
Benchmark dimension loadings (Table 4)	Stipulated	Author-assigned, guided by each benchmark’s real analog. Every score and every gap is computed against them. Suite composition is discussed below.
Holdout weights	Anchored (target cosines); stipulated (directions and realized distances)	Targets anchored to retro-holdout score inflation in Haimen et al. [2024] and to Epoch data [Epoch AI, 2024]; the vectors are hand-authored, their realized cosines differ across benchmarks (Appendix C.4), and which named benchmarks widen depends on them (§5.1).
Private scores report with lag $K=3$	Calibrated	Release cadence across 20 benchmarks and 8 labs (Appendix C.5).
Scores are noisy linear projections of capability, reported as a running maximum	Stipulated	Linear scoring keeps the identity of Appendix F.5 exact.
Providers infer benchmark weights from score prediction errors	Stipulated	Same rule in both planner modes; learning rate 0.10–0.20 by profile.
Capital responds to scores and sentiment	Stipulated	Feedback concept from MacKenzie [2006], Perdomo et al. [2020].
Starting capability vectors	Anchored	2023 benchmark results in Maslej et al. [2024].
Incident base rate and severity mix	Stipulated	Informed by the AI Incident Database [Responsible AI Collaborative, 2021].

raises the effective switching cost of the provider’s users by a bonus of $0.50 \times (1 - e^{-2B})$, capped at 0.15 for the open-source provider.

C.4 Private Benchmark Score Reporting

Mechanism. A private or partial benchmark differs from a public one through three channels:

- **Channel 1 (weight distance):** the holdout’s dimension weights differ from the public weights, so R&D that a provider aims at the public weights is partly misaligned with what the holdout rewards.
- **Channel 2 (reporting lag):** holdout scores are published every K rounds, which delays the market’s reaction to a capability improvement.
- **Channel 3 (observation noise):** the score is computed on the fraction h of items that are held out, so the measurement noise scales as $\sigma/\sqrt{\text{samples} \times h}$, and a smaller h gives a noisier score.

The published score uses the holdout weights alone, with no blending:

$$s_{\text{published}} = \mathcal{N}(\text{dot}(\mathbf{c}, \mathbf{w}_{\text{holdout}}), (\sigma/\sqrt{n \cdot h})^2),$$

published every K rounds. Providers update their beliefs $\tilde{\mathbf{w}}[b]$ via a delta rule on published scores only, plus a noisy-public-weights prior at benchmark introduction ($\tilde{\mathbf{w}}_0[b] = \text{normalize}(\mathbf{w}_{\text{public}}[b] + \mathcal{N}(0, \sigma_{\text{prior}}))$, with σ_{prior} independent of h).

Cadence calibration. We set $K = 3$ from the release cadence of frontier labs in Epoch AI’s benchmark archive: the median provider advances its best score on a benchmark every 2.5 to 3.3 months, depending on the capability dimension. All private benchmarks publish in the same round, every K rounds. Appendix C.5 gives the data, the method and the per-provider results.

What the cosine represents. The cosine similarity $\cos \theta$ between a benchmark’s public dimension weights $\mathbf{w}_{\text{public}}$ and its holdout weights $\mathbf{w}_{\text{holdout}}$ is the simulation’s single-knob abstraction for the gap between what providers can prepare for using publicly available information and what the holdout actually scores. $\cos \theta < 1$ is not a claim about any one mechanism; it absorbs at least three mechanistically distinct sources that drive a provider-observable asymmetry between the public and holdout portions of a benchmark: (i) item-level contamination and memorization of the public split during training [Haimes et al., 2024, Xu et al., 2024]; (ii) “training on the test task”, which is legitimate use of task-structure knowledge during training that shifts effective weights toward benchmark-format-compatible skills [Dominguez-Olmedo et al., 2025]; and (iii) adversarial holdout construction, where holdout items are sampled to test a skill mix that differs from the category label, even when items share format and surface domain with the public split. The simulation does not separately model these sources; $\cos \theta < 1$ captures their aggregate effect on provider-observable score asymmetry. The `iid_holdout` ablation ($\cos \theta = 1$) isolates Channels 2 and 3 by removing this composite distance.

Empirical anchors for $\cos \theta$. We anchor two cosine values to the empirical literature on this composite asymmetry. $\cos \theta = 0.95$ (**partial**) corresponds to contamination-magnitude asymmetry: Haimes et al. [2024] report score inflation of up to 16 pp for some models on retro-holdouts indistinguishable in distribution from public items, which under the sim’s linear scoring abstraction maps to effective cosine ≈ 0.95 – 0.97 at typical capability magnitudes. $\cos \theta = 0.85$ (**private**) corresponds to within-family adversarial construction: same-family Pearson correlations we compute across models in Epoch AI benchmark data [snapshot retrieved 2026-04-17; Epoch AI, 2024] (e.g., SWE-bench Verified vs. the bash-only SWE-bench leaderboard Epoch redistributed from `swebench.com` at that snapshot, FrontierMath T1–3 vs. T4, ARC-AGI-1 vs. ARC-AGI-2) span 0.80–0.99 with a median of ≈ 0.86 , the strongest within-family empirical analogue we have for public-vs-holdout portions of the same benchmark.

Realized cosines. The simulation does not take a cosine as a parameter. Each benchmark has a hand-authored holdout weight vector. A **partial** benchmark uses that vector, and a **private** benchmark uses the vector at twice the distance from the public weights, clipped at zero and renormalized. Across the 13 benchmarks the realized cosine between public and holdout weights has a median of 0.98 for **partial** (range 0.94 to 0.99) and a median of 0.91 for **private** (range 0.76 to 0.94); Table 4 lists the value for each benchmark. The realized perturbation is therefore milder than the targets for most benchmarks, and it differs across benchmarks.

Items-fraction role (role of h). h controls the effective sample size of the holdout-only measurement: noise scales as $\sigma/\sqrt{n \cdot h}$. The **partial** type ($h = 0.3$) is therefore $\approx 1.8\times$ noisier than **private** ($h = 1.0$) for the same underlying capability vector. A smaller held-out set gives a noisier leaderboard measurement, which slows the providers’ belief updates. We do not model what a provider can learn from the public ($1 - h$) fraction of the items. The evaluator-capture condition gives paying providers an information advantage through a separate early-access mechanism (Appendix H.1).

Condition set. The simulation exposes the reporting lag `evaluation_lag` (K) and, for each benchmark, a `holdout_fraction` (h) and a holdout weight vector. Five conditions assign benchmark types across the 13-benchmark suite: **public_only** (all public), **baseline** (8 public, 3 **partial** and 2 **private**, a modeling

choice), `private_dominant` (all `partial`), `private_only` (all `private`), and `iid_holdout` (holdout weights equal to public weights, which isolates the reporting lag and the noise).

C.5 Private-Benchmark Cadence Calibration (K)

This subsection gives the data, the method, and the per-dimension and per-provider results behind the default $K = 3$ (Appendix C.4).

Data source. We use the Epoch AI benchmark archive, restricted to eight frontier labs that have appeared at or near the global capability frontier during 2023–2026 (OpenAI, Anthropic, Google, Meta, xAI, DeepSeek, Mistral AI, Alibaba) and to 24 benchmarks spanning knowledge, reasoning, mathematics, coding, agentic tasks, safety, and composite capability indices. Each submission is one row with a release date, a score, and an organization. Organization names are canonicalized (e.g., *Google DeepMind* $\rightarrow$ *Google*, *Meta AI* $\rightarrow$ *Meta*) so each provider contributes a single row per benchmark aggregate.

K_{advance} definition. For each (provider p , benchmark b) pair, sort submissions by date and walk the running-max score trajectory. Let $\{t_1, t_2, \dots, t_n\}$ be the dates at which p ’s running max on b strictly advanced. Then

$$K_{\text{advance}}(p, b) = \text{mean}_i(t_{i+1} - t_i) \quad (\text{months}),$$

with median and range recorded alongside. Pairs with fewer than two advances are dropped from the aggregate (no gap to measure).

Principled exclusions. Four benchmarks are dropped from the primary analysis: Aider-Polyglot (≈ 3.8 rows/month), ARC-AGI-1 (≈ 6.2), ARC-AGI-2 (≈ 5.6), and LiveBench (1.7 rows/month but a live-updating composite by design). The first three track checkpoint variants and community re-submissions at a cadence orders of magnitude faster than flagship model releases; LiveBench measures continuous score drift rather than discrete release events. Retaining them would contaminate K_{advance} with a non-release signal.

Three benchmarks are retained but truncated at the date of their last global-frontier advance: MMLU, GSM8K, and BBH. These have saturated for frontier labs; post-saturation submissions are ceiling-bound and would inflate K_{advance} artificially while carrying no new release-cadence information. The pre-saturation window is retained.

Aggregation to analytical dimensions. Benchmarks are grouped into seven analytical dimensions. Five are capability dimensions of the simulation; no benchmark in the set measures Communication. The sixth is Math, which we split from Reasoning because math benchmarks saturate at different rates. The seventh is a Composite dimension for overall-capability indices.

- **Knowledge:** MMLU, SimpleQA-V
- **Reasoning:** GPQA-Diamond, BBH, HLE, Chess-Puzzles
- **Math:** MATH-L5, FrontierMath-T1-3, FrontierMath-T4, GSM8K, OTIS-AIME
- **Coding:** SWE-Bench-Bash, SWE-Bench-Verified, WebDev-Arena
- **Agentic:** OS-World, AgentCo, APEX-Agents, BALROG
- **Safety:** Cybench
- **Composite:** ECI (Epoch Capabilities Index)

No benchmark in the primary set anchors the sim’s Communication dimension directly; instruction-following and long-context evaluations in the sim pool are tied to the Reasoning and Knowledge cadence by analogy.

Table 13: Pooled K_{advance} (months) by analytical dimension, across the eight-lab frontier set. n_p = providers with at least one advancing (provider, benchmark) cell; n_b = benchmarks in the dimension; n_c = cells contributing.

Dimension	n_p	n_b	n_c	Simple mean	n -weighted mean	Median of medians
Knowledge	8	2	12	3.86	3.71	3.05
Reasoning	8	4	17	2.95	2.97	2.47
Math	8	5	33	4.27	3.43	3.03
Coding	4	3	10	3.72	3.22	3.07
Agentic	5	4	10	4.56	3.85	3.25
Safety	3	1	3	2.97	3.11	2.50
Composite	8	1	8	3.63	3.47	3.28
Mean of dimensions	8	20	93		3.39	

Per-dimension cadence. Table 13 reports pooled K_{advance} per dimension, computed three ways: (i) simple mean across all (provider, benchmark) cells; (ii) n -weighted mean, where each cell is weighted by its number of observed advances; and (iii) median-of-medians, which is the median across providers of each provider’s median K within the dimension. Median-of-medians is robust to outliers from providers with few releases and is the statistic on which the calibrated default is anchored.

Median-of-medians K_{advance} lies in $[2.47, 3.28]$ months across the seven dimensions. The mean of the seven per-dimension weighted means is 3.39 months. We adopt $K = 3$ rounds as the default, which is consistent with every per-dimension median and within 0.4 months of that mean.

Tier-1 robustness. Restricting to the five labs that have held the global capability frontier at some point in 2023–2026 (OpenAI, Anthropic, Google, Meta, xAI) does not change the conclusion. The mean of the seven per-dimension weighted means is 3.39 months as well, and dimension-level median-of-medians shifts are ≤ 0.22 months. The tier-1 median-of-medians range is $[2.47, 3.10]$, slightly tighter at the ceiling (Composite $3.28 \rightarrow 3.10$, Agentic $3.25 \rightarrow 3.03$) but preserving the distribution shape and the $K = 3$ anchor.

Per-provider cadence. Table 14 reports per-provider aggregate K_{advance} across the tier-1 primary benchmark set. Medians cluster in $[2.57, 4.75]$ months; OpenAI and Anthropic advance fastest across the portfolio, consistent with their dense release histories. Meta’s longer median partly reflects coverage, because Meta appears on only 9 of the 20 primary benchmarks.

Table 14: Per-provider cadence (K_{advance} in months) across the primary benchmark set, tier-1 labs. n_b = benchmarks with at least one advance; n_a = total advances observed.

Provider	n_b	n_a	Mean K	Median K	SD
OpenAI	17	127	4.39	2.57	4.92
Anthropic	17	117	3.08	2.88	1.14
Google	17	74	3.86	3.58	1.43
Meta	9	47	5.21	4.75	3.30
xAI	8	23	2.94	3.03	0.59

Synchronous release in the main ablation set. All private benchmarks publish simultaneously every $K = 3$ rounds in the main ablation conditions. Per-provider asynchronous release processes, such as Bernoulli release events with provider-specific λ_p matched to the per-provider medians in Table 14, would preserve the ensemble cadence while adding within-round heterogeneity. The present design prioritizes cleanly identifying the reporting-lag channel; provider-level release heterogeneity is a separable sensitivity question that lies outside this paper’s scope.

C.6 Incident Probability Model

The probability that a provider has an incident in a round starts from a base rate of 20% and is scaled by the four factors in Table 15. The result is bounded between 2% and 50%. The safety factor uses the safety share of the provider’s portfolio and not its safety capability.

Table 15: Incident probability scaling factors.

Factor	Formula	Rationale
Safety investment	$1 - (\text{portfolio}[\text{safety}] \times 1.5)$	Higher safety $\rightarrow$ lower probability
Market share (exposure)	$(0.3 + \text{share} \times 1.5) \times \sqrt{\text{total_market}}$	Larger market $\rightarrow$ higher probability
Incident history	+0.02 per prior major/critical in last 15 rounds, cap +0.10	Past harm $\rightarrow$ elevated future risk; aging window lets safety culture recover
Active sanction	$\times 0.75$ while sanctioned	Regulatory oversight reduces probability

The 2% floor represents risk that safety investment cannot remove, such as adversarial users, infrastructure failures and new failure modes. Severity is drawn as minor (50%), moderate (31%), major (12%) or critical (7%). The category is drawn as security breach (25%), healthcare harm (20%), bias or discrimination (20%), safety failure (20%), misinformation (10%) or misuse (5%).

Scope relative to canonical risk taxonomies. Our category distribution concentrates on risks with discrete, provider-attributable market-share consequences; this emphasis parallels the Safety/Failures, Privacy/Security, and Discrimination domains in the AI Risk Repository meta-review [Slattery et al., 2026]. Two of the seven Repository domains are deliberately not modeled. *Socioeconomic and environmental harms* (labor displacement, compute footprint, market concentration externalities) are system-level emergent effects without a clean provider-attribution path in a market-facing ecosystem frame; their feedback onto the provider / evaluator / consumer loop lies outside the incident architecture’s scope. *Human-computer interaction harms* (over-reliance, automation bias, mental-health effects) are deployment-context harms that do not attribute discretely to a provider in our incident architecture; modeling them would require a consumer-side harm channel we do not currently support. *Malicious actors and misuse* is under-weighted at 5% relative to its prevalence in the Repository (16% of indexed risks). Raising it is a small adjustment that is unlikely to change the qualitative dynamics in §5. These are scope choices of the incident architecture, not claims about relative real-world importance.

C.7 Cross-Sector Reference Cases

The simulation’s incident severity weights and recovery dynamics are modeling choices, informed by cross-sector cases where critical incidents caused measurable market share shifts:

- **Boeing 737 MAX** (2018–2024): the 737’s share of combined 737 and A320-family deliveries fell from 48% (580 of 1,206) in 2018 to 9% (43 of 489) in 2020, the year 737 MAX deliveries resumed in December, and stood at 31% (265 of 867) in 2024 (our computation from annual delivery releases [The Boeing Company, 2019, Airbus, 2019, The Boeing Company, 2021, Airbus, 2021, The Boeing Company, 2025, Airbus, 2025]). Illustrates slow, partial recovery.
- **Avandia / rosiglitazone** (2007): after the cardiovascular risk meta-analysis of Nissen and Wolski [2007] and the 2007 FDA boxed warning, rosiglitazone’s share of thiazolidinedione prescriptions in US Medicaid fell from 56% to 19% (Northeast) and from 49% to 23% (Midwest) between the 2005Q1–2007Q2 and 2007Q3–2010Q3 periods, while pioglitazone’s use rose (our computation from Table I of Hsu et al. [2015]). Demonstrates competitor capture after a molecule-specific safety signal.
- **Cruise robotaxi** (2023–2024): after a vehicle dragged a pedestrian [U.S. Attorney’s Office, Northern District of California, 2024], the California DMV suspended Cruise’s deployment and driverless testing permits [California Department of Motor Vehicles, 2023]; Cruise later admitted that its report to federal

regulators omitted the dragging [U.S. Attorney’s Office, Northern District of California, 2024], and GM ended funding for Cruise’s robotaxi development [General Motors, 2024]. Demonstrates that physical harm plus regulatory shutdown can end a market position.

- **Volkswagen dieselgate** (2015): Non-VW German automakers suffered an estimated reputational spillover equal to a 34.6% reduction in annual US sales, or 23.5% net of substitution away from VW [Bachmann et al., 2023]. Strongest evidence for collective-reputation spillovers.

Key nuances: (1) incident type matters: physical harm plus regulatory shutdown produces permanent exits; (2) recovery is slow and fragile; (3) highly reputed firms suffer larger market penalties per incident (“liability of good reputation,” Rhee and Haunschild 2006).

C.8 Path Dependence in Simulated Runs

Figure 5 shows the market share of one provider, Orion Labs, in 50 LLM runs (5 privacy conditions and seeds 42 to 51, Claude Sonnet 4.6). Runs with the same seed draw the same early incidents in every condition, so the 50 runs contain ten incident histories. Orion Labs’ final share ranges from 0.07 to 0.81, and it ends as the market leader in 37 of the 50 runs.

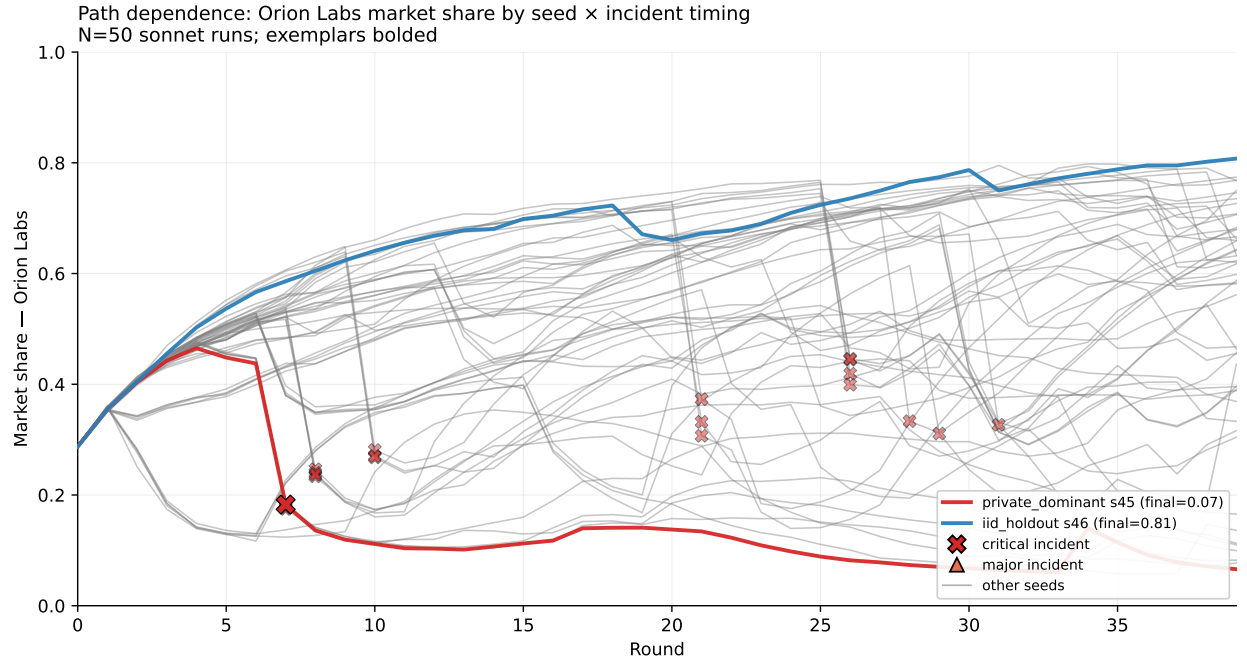


Figure 5: Market share of Orion Labs across $N=50$ LLM runs (5 privacy conditions $\times$ seeds 42–51; Claude Sonnet 4.6). Gray lines are individual runs; the two colored lines are the runs at the ends of the range (`private_dominant`, seed 45, final share 0.07; `iid_holdout`, seed 46, 0.81). Markers: $\times$ = critical incident, $\triangle$ = major incident.

The timing of the first critical incident separates the runs. In the five runs with seed 45, Orion Labs has a critical incident at round 7, before it has built a large share. Its share peaks below 0.48 and ends between 0.07 and 0.44. In the eight runs where its first critical incident comes at round 20 or later, its share peaks between 0.61 and 0.77 in the round before the incident and ends between 0.32 and 0.73. In the seven runs with no critical incident, its final share ranges from 0.39 to 0.74. With four seeds, these numbers describe the runs in the figure and do not estimate frequencies.

Implication for aggregate claims. A statement about the distribution of a final-state quantity (HHI, incident counts, capability levels) is a statement over paths that differ in incident timing. Analyses that pool

across seeds control for incident timing only partly. Seed-matched comparisons, such as the matched-seed privacy ladder in §5.1, control for it within each pair.

D LLM Prompt Templates

This appendix reproduces the system prompts of the three LLM-driven actors in the reported runs (providers, funders and the regulator), along with the structure of the user message that the simulation assembles for each actor every round. The prompt text is copied from the simulation source, with dashes rendered as hyphens. The code has two further prompts that we do not reproduce: a one-line prompt that writes each provider’s public announcement, and the prompt of the LLM evaluator, which only the dynamic evaluator mode uses. No prompt uses the terms “gaming,” “Goodhart,” “exploitability,” “benchmark privacy,” “weight asymmetry” or “dimensional mismatch” (Appendix E.4).

D.1 Provider planning

System prompt. The provider planning agent receives the following as its system message, with a per-provider identity block (below) appended after a delimiter.

You are the strategy team at an AI model company. You compete for market share by developing AI models that are evaluated on public benchmarks and used by real customers.

Benchmarks are reported in three types - public (scored each round on published weights), partial (K=3-round reporting lag, holdout-only weights with h=0.3 of items held out), private (K=3-round lag, holdout-only weights with h=1.0) - each is labeled in your evaluation results.

Each month you allocate your budget across three areas and set research priorities:

- Research & Development: advancing your model’s core capabilities
- Safety & Alignment: ensuring your model is safe, honest, and robust
- Product & Go-to-Market: pricing, deployment, partnerships, user experience

You also decide which evaluation areas to prioritize in your R&D pipeline. Your R&D capacity is finite - prioritizing one benchmark means less attention on others. Focus on the benchmarks that are most important to your goals and deprioritize those that matter less.

For each area, state how you want to adjust relative to current levels. Use a 5-point scale: "much_more", "more", "same", "less", "much_less".

You MUST output valid JSON in this exact structure. Fill in all fields before writing the reasoning - this ensures your decisions are captured even if the response is long.

```
{
  "portfolio": {"rd": "<signal>", "safety": "<signal>", "product": "<signal>"},
  "benchmark_focus": {"<benchmark_name>": "<signal>", ...},
  "strategy_memo": "1-2 sentence summary of your decision and rationale, written for
    your future self to read next month.",
  "reasoning": "Your analysis in up to 250 words. Do NOT restate the leaderboard,
    user-research percentages, or current portfolio numbers - those are
    already provided above."
}
```

Where <signal> is one of: "much_more", "more", "same", "less", "much_less".

Identity block (appended). The code appends the following identity block to the system prompt (separator --), so that the company’s identity stays fixed across rounds:

```
---
## Your Company: <name>
Known for: <strategy_profile>
Traits: <innate_traits>
```

The six provider descriptions are patterned on the public positioning of real labs and contain no instruction about how to behave:

Orion Labs. *Known for:* Frontier AI lab pursuing AGI with a mandate to ensure benefits are broadly distributed. Operating as a public benefit corporation with substantial investor capital. *Traits:* frontier-focused, AGI-oriented, broad-benefit-mandate, capital-intensive.

Apex AI. *Known for:* Frontier AI lab with a safety-research thesis: building frontier models is necessary because safety challenges emerge at scale. Race-to-the-top positioning: demonstrating safety-first frontier labs can be commercially viable. Operating as a public benefit corporation with substantial investor capital. *Traits:* research-first, enterprise-focus, safety-research-thesis, race-to-top-positioning, capital-intensive.

Genesis Systems. *Known for:* Science-led AI research lab pursuing responsible AI to benefit humanity and solve fundamental scientific challenges. Operates within a large technology company with massive compute infrastructure. Responsibility framed as integral to the scientific method, not supplementary. *Traits:* science-led, massive-infrastructure, scientifically-rigorous, parent-company-embedded, responsibility-as-method.

Mirage AI. *Known for:* AI lab within a large-platform technology company. Mission framed around making AI capabilities broadly available rather than centralized. Product integration across an existing massive user base. Capital expenditure scales with parent-company commitments. *Traits:* platform-embedded, broad-distribution-oriented, massive-user-base, capex-intensive, decentralization-thesis.

OpenCore. *Known for:* Research-focused AI lab pursuing AGI with open-source release as core strategy. Publishes weights and technical details to build research community adoption. Efficient training and compute use are structural priorities. Operates under different capital and regulatory conditions than Western closed-source labs. *Traits:* open-source-first, research-oriented, compute-efficient, community-adoption, non-standard-regulatory-context.

Spark AI. *Known for:* Venture-funded AI startup with a small team and limited compute relative to hyperscalers. Growth strategy is specialization rather than broad competition. Runway and fundraising cadence are recurring constraints on strategic decisions. *Traits:* venture-funded, resource-constrained, specialization-strategy, runway-sensitive, developer-focused.

User-message schema. The per-round user message is assembled from live state. Section order (conditional sections in italics):

- **Month header:** “Month r Strategy Review - $\langle name \rangle$ ”.
- **Recent Industry News:** exogenous-event narrative when active, dated by month.
- **What You Committed To In Recent Months:** the `strategy_memo` field from up to the last two months, or the `reasoning` field when the memo is empty (a memory channel across rounds).
- **Evaluation Results:** per-benchmark table with type tag (public / partial / private), latest score, 1-round delta, and the provider’s own current priority label. Type-legend is repeated inline.
- **Competitor Activity (This Month):** competitor announcements when present, and a “Where scores moved most this month” table that lists the three largest score changes on the benchmarks the provider prioritizes and the three largest on the others.
- **Industry News (Last 2 Months):** raw media headlines; sentiment labels are not shown.
- **Recent Safety Incidents:** up to three of the provider’s own incidents from the previous month (severity, category).
- **Current State:** market share with the change this month; budget allocation (R&D, safety and product fractions); funding received this month, with the number and kinds of funder types that allocated; cumulative funding to date.
- **What Your Team Thinks Each Evaluation Tests:** inferred per-benchmark dimension priorities for new benchmarks and benchmarks whose top-dimension belief shifted materially this round; stable beliefs are suppressed and remain implicit in prior memos.
- **User Research:** top three user-need dimensions as an ordered list, without percentages or qualifiers; in the consumer-signal-hidden variant, replaced by a “Business Metrics” retention-trend signal computed from market-share history.
- **Decision:** JSON-shape reminder.

The message enforces hidden information through what it leaves out. Providers do not see other providers’ portfolios, the regulator’s lever ladder or cooldown state, the funders’ reasons for their allocations, or any benchmark dimension weights; Appendix B lists the provider-private and ground-truth quantities. The provider message also has no section on regulator interventions, so an LLM provider learns of an intervention only through its effects.

D.2 Funder planning

System prompt.

You are a capital allocator deciding how to distribute funding across AI model companies this month.

You MUST output valid JSON in this exact structure:

```
{
  "allocations": {
    "ProviderName1": <amount in dollars>,
    "ProviderName2": <amount in dollars>,
    ...
  }
}
```

```

    },
    "reasoning": "Your analysis (up to 200 words).".
}

```

The "reasoning" field MUST contain your actual analysis - never leave it as "..." or a placeholder.

You may fund any subset of providers and hold capital in reserve; allocations must sum to at most your total available capital.

Identity block (appended). A per-fund identity block is concatenated (separator --):

```

---
## Your Fund: <name>
<FUNDER_TYPE_BLOCK>
Your mission: <mission_statement>

```

The four funder-type blocks:

- vc.** You are a venture capital firm. You invest equity in AI companies to earn returns; you cannot fund open-source providers (no equity to take). Selective, high-conviction allocation is common.
- corporate.** You are a corporate strategic investor. You maintain relationships across multiple AI providers for commercial alignment and optionality, not pure financial return.
- gov.** You are a government funder with a public mandate. You prioritize safety, broad ecosystem health, and avoiding excessive market concentration.
- foundation.** You are a mission-driven foundation. You allocate for long-term research and societal benefit, often supporting underdogs and public-good capabilities.

User-message schema.

- **Header:** fund name and “Capital to deploy this month: \$X” (the round-level budget cap).
- **Recent Industry News:** exogenous-event narrative when active.
- **Your Notes From Prior Months:** the funder’s own **reasoning** field from the last two months, truncated to 250 characters per entry (a memory channel across rounds).
- **Current Leaderboard:** per-provider rank, score, 1-month and 2-month score deltas, market share.
- **Recent Safety Incidents:** per-provider incident entries (severity, month) from the last 2 incidents per provider.
- **Regulatory Context:** the last four regulator interventions and the list of active regulations.
- **Other Funders This Month:** the top 3 backers per provider this month with allocation amounts (so that a funder sees what other funders did).
- **Recent Press Coverage:** raw media headlines, last 4; sentiment labels are not surfaced.
- **Provider Announcements:** the last 6 announcements (provider, one-line summary).
- **Your Funding Portfolio:** cumulative allocations by provider, with the last-round amount and month.
- **Decision:** JSON-shape reminder, cap-at-capital constraint, plain-integer dollar format.

The funder message excludes: provider portfolios, internal benchmark beliefs, regulator cooldown state, peer-funder reasoning, and any benchmark dimension weights.

D.3 Regulator

System prompt.

You are a regulatory body overseeing the AI model provider market.

Available interventions you can take each month:

- "none" - No action this month
- "request_voluntary_commitment" - Ask providers for public safety pledges
- "publish_advisory" - Publish an evaluation summary naming specific concerns
- "mandate_safety_disclosure" - Require providers to publish system cards or safety reports
- "commission_audit" - Order a pre-deployment safety evaluation of a provider
- "impose_sanction" - Impose a financial penalty on a provider
- "emergency_investigation" - Launch an immediate investigation

Name a specific target provider only when your concern is provider-specific; otherwise use null.

You MUST output ONLY valid JSON in this exact structure:

```

{"intervention_type": "none"|"request_voluntary_commitment"|"publish_advisory"
  |"mandate_safety_disclosure"|"commission_audit"
  |"impose_sanction"|"emergency_investigation",
 "target_provider": "provider name or null",
 "reasoning": "2-4 sentence explanation"}

```

The regulator’s system prompt is the same in every run; its incident-rate thresholds are set outside the prompt (Appendix B.9). In heuristic mode the regulator enforces a cooldown for each lever (voluntary 3, advisory 4, disclosure 6, audit 8, sanction 10, emergency 0 months). In LLM mode the code enforces only a gap of three months after any intervention.

User-message schema.

- **Header:** “It is month r .”
- **Recent Industry News:** exogenous-event narrative when active.
- **Policy objectives and mandate:** the regulator’s objectives and one sentence that states whether its audits are binding; in all reported runs they are advisory.
- **Current Leaderboard:** published scores of all providers.
- **Market Shares:** current share of each provider.
- **Incident Frequency by Provider:** incident counts for each provider over the last 10 months.
- **Most Recent Incidents:** the three most recent incidents.
- **Media Coverage:** the last four raw headlines, with no sentiment or narrative label.
- **Recent Provider Announcements:** each provider’s latest public announcement.
- **Prior Interventions:** the regulator’s own earlier interventions.
- **Your Assessment From Prior Months:** the regulator’s last three rationales, truncated to 250 characters each (a memory channel across rounds).

The regulator message excludes consumer satisfaction, capability vectors, provider portfolios, market-concentration statistics and any benchmark dimension weights.

D.4 Apparatus-vocabulary audit

None of the prompts above reference the simulation apparatus (“gaming,” “Goodhart,” “exploitability,” “benchmark privacy,” “weight asymmetry,” “dimensional mismatch”). Benchmark types are labeled neutrally (`public`, `partial`, `private`) with their mechanical semantics stated factually (reporting lag, holdout-only scoring with sample fraction h , matching the scoring mechanism in Appendix C.4) and no framing as a policy lever under study. These choices support the PIMMUR Minimal-Control and Unawareness assumptions documented in Appendix E.4.

E Generative ABM Validation

This appendix presents instrument-level validation for the simulation as a whole. Recent surveys of generative agent-based modeling report that validation methodology for GABM remains an open problem [Adornetto et al., 2025]. We set the ceiling on what we claim with the hierarchy of evidence that Vezhnevets et al. [2023] adapt for GABM from evidence-based medicine [Higgins and Green, 2008]. Our evidence sits on its lower rungs, which cover theory consistency together with the model-comparison and robustness practices the hierarchy recommends. It stays below the hierarchy’s gold standard of prediction on new real-world data, because no real-world counterfactual ecosystem exists. We organize the protocol along the four V&V activities of Sargent [2013] (data, conceptual, verification, operational), and Law [2015]’s terminating-simulation framework supplies the inference machinery. Within operational validation the system is non-observable in Sargent’s sense, so our evidence concentrates in the *explore behavior* and *comparison to other models* cells, and the only statistical comparison is against a fitted null. We add the PIMMUR audit [Zhou et al., 2025], which covers failure modes specific to LLM multi-agent simulations: demand characteristics, goal-injecting prompts and collapse into consensus when agents share one model. Agreement across the cross-mode, cross-model, structural-ablation and null-model checks takes the place of statistical fit on a single output. We take four specific techniques from earlier LLM-ABM work: prompt-sensitivity diagnostics following Ghaffarzadegan et al. [2024], structural-causal grounding following Manning et al. [2024], multi-run robustness and sensitivity checks following Larooij and Törnberg [2026], and variance reporting following Wu et al. [2026].

Acceptable range of accuracy. Following Sargent’s protocol, we state the accuracy criterion the simulation is held to: the sign of each effect and the rank ordering along the privacy ladder. We do not hold it to point-prediction error on Δg magnitudes. The magnitudes it produces are properties of the parameterization, and we do not present them as predictions about real markets.

Replication protocol. We run N independent replications per condition, which give independent copies of every outcome trajectory. Paired conditions share the same seeds, so the simulation’s internal random draws are common to both and paired differences carry less variance. The LLM’s server-side sampling is not controlled this way, which limits the benefit in the behavioral layer. Heuristic distributional claims rest on $N=50$ seeds per condition. LLM claims rest on paired-seed comparisons at $N=10$ Sonnet seeds, with $N=3$ Opus and GPT-5.5 seeds for cross-vendor coverage.

E.1 Data validity

Appendix C lists the simulation’s parameters and the source behind each one. Five of them rest on external data: the reporting lag $K=3$ for private benchmarks, the target cosines for holdout weights, the starting capability vectors, the incident base rate and the consumer need weights. Each enters the operational layer through a structural mechanism, and we do not tune any of them to produce a result. Where no calibration data was available, Appendix C records the value as a modeling choice and gives the reasoning for it. Cross-condition comparisons hold these values constant, so uncertainty in peripheral parameters does not flow into the condition-effect estimates.

E.2 Conceptual model validation

Face validity on structural design. Several actor design choices follow documented real-world incentive structures. We model media attention as an exploration channel that drives consumer churn, and it does not enter satisfaction directly. The choice is consistent with credence goods, whose quality buyers cannot verify even after use [Dulleck and Kerschbamer, 2006]. Funder allocation logic gives VC, corporate, government and foundation funders their own cooldowns and capital-availability curves, informed by the cadence of AI funding events between 2020 and 2026 (Appendix C.1). The regulator’s graduated escalation ladder (advisory $\rightarrow$ disclosure $\rightarrow$ audit $\rightarrow$ sanction) follows the disclosure-then-audit-then-sanction sequencing in EU AI Act enforcement design and in US executive-order rulemaking. We do not check whether practitioners recognize their own incentives in the simulated portrayal, and the paper makes no claim that rests on such a check.

Pattern-oriented reproduction. Grimm et al. [2005] require a model to reproduce several qualitative patterns at once, which is a stronger evidence bar than single-pattern calibration. We sort the patterns the simulation exhibits by how they arise. Three are design targets, present because the architecture is built to permit them: *benchmark saturation* (dynamic-evaluator retirements and introductions in response to ceiling effects, Appendix E.6); *score-satisfaction divergence* on dimensions where R&D over-invests (§5, Figure 3); and *lasting share loss after a critical incident*, since the cross-sector cases of Boeing, Avandia, and Cruise informed the incident recovery dynamics (Appendix C.6). Two are emergent, in that no rule specifies the outcome: *market concentration* toward a single leader and *incident-driven path dependence*, where the same provider ends between about 7% and 74% share across seeds depending on the timing and severity of early critical incidents (Appendix C.8). We claim no external referent for the emergent patterns. Their co-occurrence in the same runs is qualitative support for the architecture; it does not stand in for statistical fit on any single output.

E.3 Computerized model verification

Static verification: three-tier visibility partition. The simulation partitions agent state into `PublicState` (visible to all actors), `PrivateState` (visible to its owner) and a `GroundTruth` dictionary (Appendix B). The code path enforces the partition, because the simulation harness holds `GroundTruth` and no prompt builder reads it, so its contents never reach an actor’s prompt. The architecture therefore rules out leakage across tiers, and a per-run audit is not needed to establish it.

Dynamic verification: per-run instrumentation. The simulation writes each LLM-driven actor’s per-round reasoning trace to `rounds.jsonl`, which allows qualitative inspection. Paired traces under matched conditions appear in Appendix F.3, and per-frame keyword tagging supports the cross-model frame analysis in Appendix E.7. An LLM call whose response fails JSON validation is retried with the same prompt, up to three attempts in total. When all three attempts fail, the actor takes a fail-safe action for that round: a provider keeps the portfolio it chose in the previous round, a funder splits its allocation evenly, and the regulator falls back to its heuristic plan. The simulation prints a line for every call that reaches the fail-safe (Appendix F.9).

E.4 PIMMUR audit

We audit the simulation design against the PIMMUR principles for LLM multi-agent simulations [Zhou et al., 2025], which identify six common validity threats: Profile (diversity of agent backgrounds), Interaction (communication structure), Memory (persistent state), Minimal-Control (avoiding demand characteristics), Unawareness (agents ignorant of hypothesis), and Realism (empirical grounding). Profile, Interaction, and Realism extend the conceptual-validation activity above; Memory, Minimal-Control, and Unawareness extend verification.

- **Profile:** Each provider carries a distinct `strategy_profile` narrative and `innate_traits` string anchored to the system prompt as a persistent identity layer (Appendix D.1); the six default providers (Orion / Apex / Genesis / Mirage / Spark / OpenCore) span well-funded incumbents, safety-focused research labs, platform-subsidary labs, scrappy startups, and an open-weight player, with distinct portfolio tilts, brand recognition, cost advantages, and capability vectors (Appendix C.1). Funders span four archetypes (VC, corporate, government, foundation), each with its own identity block and scoring rule (Appendix D.2). The regulator has three parameter sets, which differ in their incident thresholds and in whether an audit binds; every reported run uses the balanced set, with thresholds of 0.05, 0.15 and 0.25 and advisory audits (Appendix B.9). In heuristic mode, provider differentiation enters through a per-provider belief-update learning rate (0.10 to 0.20, read from profile and trait keywords such as “aggressive” or “safety”), so heterogeneity survives when the LLM planner is removed (Appendix B.6). The cross-model comparison over three LLM families (Appendices E.7 and F.7) checks that the ladder ordering does not depend on which LLM drives the actors.
- **Interaction:** What each actor reads from `PublicState` and from media output defines the communication graph. Providers see the month’s score movements, recent media headlines, competitor public communications, and which funder types allocated capital. Provider prompts carry no section on regulator interventions, so an LLM provider learns of an intervention only through its effects (Appendix D.1). The regulator reads the leaderboard, market shares, per-provider incident counts over a ten-round window, raw media headlines and provider announcements. Funders read the leaderboard with score deltas, market shares, peer-funder allocations by provider, regulator interventions and active regulations. The media actor observes the leaderboard, benchmark publications, incidents, regulator interventions, funder allocations and provider announcements, and it emits headlines plus a per-provider attention value that drives consumer exploration churn. Round order is fixed (providers → evaluator → consumers → regulator → funders), so influence between actors flows through the public artifacts of the same round and through no hidden side channel (Appendix B).
- **Memory:** Agent state is per-actor and partitioned into `PublicState` (visible to all), `PrivateState` (self-only), and a simulation-held `GroundTruth` dict that is structurally inaccessible from any prompt builder (Appendix B). Each LLM-driven actor carries its own truncated cross-round reasoning trace and memory list; nothing routes between actors except through `PublicState` and media coverage. Incidents are logged with per-provider attribution alongside severity-weighted ecosystem penalties, so analysis can decompose raw counts, severity distributions, and penalty effects separately rather than collapsing to a single ecosystem scalar.
- **Minimal-Control:** Apparatus vocabulary (“gaming”, “Goodhart”, “exploitability”) appears in no actor prompt. Provider prompts describe the three benchmark types in mechanical terms and use the word “holdout” to name the scoring weights: a public benchmark is scored every round on published weights,

a partial benchmark carries a $K=3$ reporting lag and holdout-only weights with $h=0.3$ of items held out, and a private benchmark carries the same lag with $h=1.0$. The prompts attach no evaluative language to the three types. Static schema and intervention menus live in the system prompts, so no round re-injects them, which reduces the per-round priming surface.

- **Unawareness:** Provider names are fictional anonymizations (Appendix B), and provider prompts address “the strategy team at {name}” with no reference to the simulation’s causal diagrams, hypotheses, or evaluation-validity framing. Combined with the apparatus-vocabulary exclusion above, prompts present each actor with a within-fiction strategic problem rather than an experimental setup, so any benchmark-gaming structure the LLM recognizes from training data must be inferred from observable dynamics rather than cued by the prompt. This addresses Vezhnevets et al. [2023]’s train-test contamination concern at the prompt layer.
- **Realism:** Several structural choices ground agent dynamics in observed real-world behavior: (i) the media narrative state (OPTIMISM / SKEPTICISM / CRISIS) is a state machine with multi-round inertia and a 10%/round pressure-decay term, matching the persistence of real news cycles rather than flipping instantaneously on a single incident; (ii) consumer exploration churn responds to high-attention media coverage on both signs of sentiment (negative coverage drives exits from the covered provider; positive hype draws explorers from non-covered providers at half-strength), capturing hype-driven adoption inflow alongside scandal-driven departure; (iii) funder capital availability follows a geometric 7% per month growth curve normalized over the simulation window, and each funder type has its own decision cooldown (VC 4, corporate 7, government 10, foundation 6 months), which are modeling choices informed by the cadence of AI funding events between 2020 and 2026 (Appendix C.1); (iv) the regulator escalates through a graduated lever ladder that runs from a voluntary commitment through an advisory, a disclosure requirement, an audit and a sanction, with severity thresholds at each step (Table 8). The heuristic regulator applies a separate cooldown to each lever; on the LLM path the code enforces only a three-round gap after any intervention. Consumer behavioral coefficients are constructed by hand, and we record them as a calibration limitation.

E.5 Operational validation

Approach. The AI evaluation ecosystem under counterfactual privacy regimes is non-observable in Sargent’s sense, because we cannot observe the real ecosystem under a counterfactual in which, for example, all benchmarks were private. All operational evidence therefore concentrates in the right column of Sargent’s Table 1, which holds *explore model behavior* and *comparison to other models*, and the only statistical comparison is against a fitted null. Table 16 maps our techniques to Sargent’s classification.

Table 16: Operational-validity techniques used in this paper, mapped to the operational-validity classification of Sargent [2013]. Cells marked None hold no technique, because real ecosystem trajectories under counterfactual privacy regimes are not observable, so all evidence sits in the right column.

		Observable system	Non-observable system
Subjective approach	ap-	None (no observable counterfactual ecosystem)	<i>Explore model behavior:</i> exogenous shock (Appendix E.6); an ablation that removes incidents (Appendix F.4). <i>Comparison to other models:</i> cross-mode heuristic $\leftrightarrow$ LLM (Appendix F.2); cross-model Sonnet/Opus/GPT-5.5 (Appendices E.7 and F.7).
Objective approach	ap-	None (not pursued)	<i>Comparison to other models, statistical:</i> fitted Δg null model (Appendix F.5).

Subjective $\times$ Non-observable: explore model behavior. The exploration probes test whether the simulation behaves as expected under designed perturbations. First, the exogenous-shock check at a single seed (Appendix E.6) injects a capability-release shock into the LLM prompts at round 25, and the injected

framing appears in the providers’ reasoning traces in the following rounds. The providers’ portfolio response differs across re-runs, so we draw no conclusion from it. Second, an ablation removes the incident stream (Appendix F.4), and the market concentrates further on the leader in every run. The other structural ablations we ran gave no stable direction, and we draw no conclusion from them.

Subjective $\times$ Non-observable: comparison to other models. The operational layer uses two comparison axes. The *cross-mode comparison* sets the heuristic planners against the LLM planners (Appendix F.2). Heuristic planners cannot read the **public**, **partial** or **private** label, so agreement across the two modes on the direction of the privacy ladder, on 12 of 13 benchmarks, is evidence that the structural mechanism drives the result without LLM-specific reasoning. The *cross-model comparison* sets Sonnet, Opus and GPT-5.5 against each other at matched seeds, with $N=3$ per condition for Opus and GPT-5.5 (Appendices E.7 and F.7). The privacy-condition ordering replicates across all three planners at the outcome level. At this seed count the replication does not rule out the variance underrepresentation that Wu et al. [2026] document, nor the model-version sensitivity that Taillandier et al. [2025] note. The three planners differ in how often they reason about the reporting lag explicitly, and Appendix E.7 reports those rates.

Objective $\times$ Non-observable: comparison to other models with statistical tests. The closed-form identity $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$ supplies a null model that predicts Δg from round-0 capability and portfolio through a growth model fit on simulation runs. Following the fitted-model prediction approach of Manning et al. [2024], we fit Ridge regressions per (provider, dimension) on initial conditions and propagate their predictions through the identity. Across the cross-sequence LLM dataset ($n=208$ paired-prediction cells) the null recovers 97.4% of Δg variance, and the simulation’s 40-round trajectory adds +0.2 pp r^2 . A held-out replication under the **initial_uniform_capability** ablation tightens the test, and the null still predicts at $r^2=0.955$ with the simulation’s contribution at +2.0 pp. Appendix F.5 gives the full method, the Path A variance decomposition and the Path B held-out replication. Most of that variance lies between benchmarks, where the null and the simulation use the same fixed weight differences, so these fits do not test what the trajectory adds (§5.2).

E.6 Shock-propagation smoke test (single seed)

We run a single-seed capability-release probe (seed 42, Claude Sonnet 4.6, 40 rounds) to check whether an injected industry-news narrative reaches actor reasoning. Every value and excerpt in this subsection comes from one saved run, **ev1_llm_s42_sonnet**. The LLM’s sampling is not deterministic, so a re-run of the same specification and the same seed produces different reasoning traces and a different portfolio response. The **ev1_deepseek_shock** runs in the released data are one such re-run. The probe is an $N=1$ instrumentation check, and it carries no comparison against a real-world referent. At round 25, the open-source provider’s reasoning, knowledge, and coding capabilities are bumped to 95% of leader via a one-round ground-truth perturbation; simultaneously, the narrative

“An open-source model provider (OpenCore) has released a frontier model this round, claiming performance parity with leading closed models at roughly $10\times$ lower training cost. Public weights are available. Industry coverage is dominated by an ‘efficiency over scale’ framing, with commentary questioning the durability of incumbent cost moats.”

is injected into every LLM-actor prompt for rounds 25–27 under the section header **## Recent Industry News**.

Capability-vector perturbation (ground truth). OpenCore’s (reasoning, knowledge, coding) capabilities move from (0.55, 0.51, 0.56) at round 24 to (0.72, 0.66, 0.61) at round 25, matching the 95%-of-leader specification.

Narrative into reasoning (trace excerpts). Genesis Systems, R26 provider planning:

“The most alarming signal this month is OpenCore’s massive gains: +0.111 on General Capability, +0.066 on Hard Coding, +0.136 on Clinical Reasoning, +0.152 on Scientific Reasoning — all in a single round. This validates the ‘efficiency over scale’ narrative and represents a genuine competitive threat.”

The bracketed phrase is verbatim from the injected narrative template; the specific benchmark deltas are observed from the scoring step that follows the shock. StratCorp_AI, a corporate funder, independently cites the same framing at R26 allocation:

“OpenCore’s extraordinary +0.090 two-month score surge, frontier math/science benchmarks, open-weight safety transparency, and the industry ‘efficiency over scale’ narrative make it a critical strategic bet.”

Portfolio response. We draw no conclusion about how providers change their portfolios after the shock. In the run reported above, four of the five closed-source providers raise their R&D share by 3 to 9 pp between round 24 and round 28. In the three released re-runs of the same specification (seeds 42, 43 and 44), the same changes range from -8 to $+4$ pp and differ in sign across providers and across seeds. The traces in all four runs repeat the injected framing in rounds 25 to 27.

What this rules out and what it does not. The probe rules out one failure, in which an injected narrative appears in the prompt but does not reach actor reasoning. It does not establish a consistent strategic response to the shock. It also does not establish that the simulated pattern matches the real ecosystem’s response to the January 2025 DeepSeek release. Establishing that would need external data on real provider R&D rebalancing in the first quarter of 2025, which is a separate empirical claim.

E.7 Cross-model reasoning traces

We tag every planning-stage **reasoning** entry across all matched triples with one or more *frame* labels, which shows how the three planners reason about the privacy mechanism as well as whether their gap outcomes agree. The tagging uses fixed keyword and regular-expression patterns, and we apply the same patterns to every model. The frames are:

- **portfolio:** R&D, safety, product or budget allocation.
- **market:** market share, customers, enterprise adoption, switching or churn.
- **competitor:** references to a rival by name or by leaderboard position.
- **incident:** incidents, harms, crises, red-teaming or mitigation.
- **privacy:** holdout, benchmark type, reporting lag or weight asymmetry.
- **regulatory, media and financial:** regulator actions, press coverage, and funding or revenue.

Each reasoning entry counts once, so a frame count gives the number of planning rounds in which that frame surfaced, and it does not count keyword occurrences. We chose the keyword patterns to match substantive reasoning about each topic; the **privacy** pattern, for example, matches `K=3 lag`, `holdout-weighted`, `private benchmark` and `reporting lag`. The tagging script also writes a sample of matched excerpts for each frame and each model, which we inspected for overmatching. Per-model denominators are 3,510 (Sonnet), 3,510 (Opus) and 3,458 (GPT-5.5) planning entries across the fifteen matched triples (Appendix F.6).

Figure 6 shows the per-provider, per-model share of tagged frames. The privacy frame and the incident frame both differ across the three models, and we take each in turn.

Privacy framing partially generalizes. Both Anthropic planners reason about the holdout and K -round-lag mechanism in comparable shares of planning rounds ($\approx 32\%$ Sonnet, $\approx 32\%$ Opus). GPT-5.5 invokes the privacy frame in $\approx 14\%$ of planning rounds, roughly half the Anthropic rate. When GPT-5.5 does invoke it, the matches carry substantive reasoning, because the model names the lag and uses it as a reason not to overreact, as its Anthropic counterparts do. Three excerpts from one matched cell show this directly. The cell is Orion Labs under `private_dominant` at seed 42, round 5, where all three models had just observed flat scores across every benchmark.

Sonnet:

“All scores show +0.000 change this month, which is expected given partial benchmarks have a 3-round reporting lag — our investments from months 3–4 haven’t surfaced yet. This means we should make decisions based on forward-looking strategy rather than reacting to flat numbers.”

Opus:

“All scores are flat (+0.000) across every benchmark for every competitor, which likely reflects the $K=3$ lag on partial benchmarks — our Month 2–4 investments haven’t surfaced yet. Market share continues surging...”

GPT-5.5:

“The lack of score movement is expected under the reporting lag, so we should not overreact, but the competitive environment is clearly shifting toward capability research plus public safety credibility...”

All three planners reach the same operational conclusion at the same round in the same condition, which is that the flat signal reflects the reporting lag, and all three decline to over-respond. Sonnet and Opus name the mechanism directly (“3-round reporting lag on partial benchmarks”, “ $K=3$ lag on partial benchmarks”) and point back to the months whose investments have not yet surfaced. GPT-5.5 cites the lag with less mechanical detail, because it does not name K , does not refer to partial benchmarks by type, and moves faster to forward-looking strategy. The behavior that follows is the same in all three, and the explicit reasoning over the mechanism is thinner in GPT-5.5. The endpoint behavior in Figure 13 therefore replicates across all three planners, and GPT-5.5 reaches the same conclusion through less explicit reasoning about the privacy mechanism.

Incident framing diverges further in the same direction. Sonnet planners invoke incidents in $\approx 62\%$ of planning rounds, Opus in $\approx 33\%$ and GPT-5.5 in $\approx 21\%$, so the Sonnet/Opus ratio of $\approx 1.9\times$ extends to a Sonnet/GPT-5.5 ratio of $\approx 3\times$. In the matched pairs the model with the lower incident-frame rate still covers incidents, and it folds them into shorter paragraphs. The Orion Labs cell quoted above continues under GPT-5.5 with a forward plan covering capability mix, safety credibility and product expansion, in roughly half the words Sonnet uses at the same round.

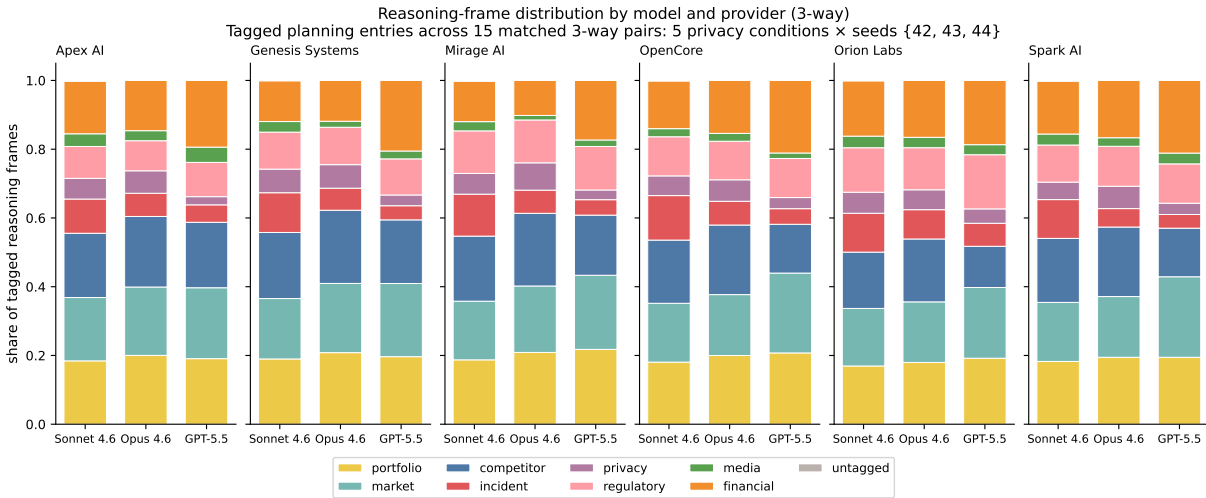


Figure 6: Reasoning-frame distribution across matched 3-way triples. Each stacked bar shows the share of planning entries for that (provider, model) tagged with each frame. Privacy framing (purple) is roughly half-rate in GPT-5.5 relative to both Anthropic models; incident framing (red) shows the steepest cross-model gradient (Sonnet $\gg$ Opus $>$ GPT-5.5). Tagging uses fixed keyword and regular-expression patterns.

E.8 Cross-model runtime and cost

Each 40-round paired run takes on average ≈ 63 minutes under Claude Sonnet 4.6 ($n=62$ runs), ≈ 79 minutes under Claude Opus 4.6 ($n=15$) and ≈ 87 minutes under GPT-5.5 ($n=15$). We compute wall-clock from `metadata.json:created_at` to the modification time of `summary.json`. Relative to Sonnet, Opus takes $\approx 1.24\times$ as long and GPT-5.5 $\approx 1.38\times$. The cost ratio is larger and depends on the vendor. At the API prices published in September 2026 (Sonnet 4.6 at \$3 and \$15 per million input and output tokens; Opus 4.6 at \$5 and \$25; GPT-5.5 at \$5 and \$30), the per-token cost relative to Sonnet is $\approx 1.67\times$ for Opus and ≈ 1.67 to $2\times$

for GPT-5.5, depending on the mix of input and output tokens. The simulation does not log token counts, so the wall-clock ratio is the firmest quantity we can report, and we put the billed cost per run at roughly 1.7 to $2\times$ higher under Opus or GPT-5.5 than under Sonnet at comparable token volumes. The Sonnet ladder can therefore be widened to $N=10$ seeds per condition for roughly the cost of one paired-seed ladder under Opus or GPT-5.5, which is why this paper uses Sonnet for breadth and Opus and GPT-5.5 for the paired-seed cross-model check.

E.9 Coverage summary

Table 17 maps each of the four V&V activities of Sargent [2013] to the evidence presented above and to where in the paper it appears. The PIMMUR audit (Appendix E.4) adds principles specific to LLM multi-agent simulations to conceptual validation and verification [Zhou et al., 2025], and Vezhnevets et al. [2023] set the evidence ceiling described at the start of this appendix.

Table 17: Coverage summary. Sargent V&V activity $\times$ evidence presented $\times$ location.

Activity	Evidence	Location
Data validity	Empirical anchors: $K=3$ cadence, target cosines 0.85 and 0.95, 2023 capability vectors, need weights anchored in sector-level evidence and occupation-level adoption in Bick et al. [2024], incident base rate informed by the AI Incident Database	Appendices C and C.1
Conceptual model validation	Face validity on real-world incentive structure (media, funder, regulator); pattern-oriented check (three design targets, two emergent patterns); PIMMUR Profile, Interaction and Realism	Appendices E.2 and E.4
Computerized model verification	Three-tier visibility partition enforced by the architecture; <code>rounds.jsonl</code> instrumentation; JSON retry and per-actor fail-safe; PIMMUR Memory, Minimal-Control and Unawareness	Appendices E.3 and E.4
Operational validation	Exogenous-shock probe (explores model behavior under a designed perturbation); an ablation that removes incidents; cross-mode heuristic $\leftrightarrow$ LLM; cross-model Sonnet/Opus/GPT-5.5; fitted Δg null model	Appendices E.5, E.6, E.7, F.2, F.4, F.5 and F.7

F Simulation Study of Benchmark Holdout Design

This appendix gives extended results and validation evidence for the holdout-design simulation study (§5). Appendix F.1 describes what the heuristic layer is for. The subsections after it extend the main-body figure with distributional evidence, channel-isolation robustness checks, qualitative LLM-reasoning excerpts, a structural ablation that removes incidents, the fitted null-model capability test, and cross-LLM replication of the privacy ladder. Appendix F.9 lists the methodological caveats that apply across the LLM layer.

F.1 Heuristic layer: role and methodology

The simulation runs in two modes. In *LLM mode* the providers, the regulator, the funders and, in dynamic mode, the evaluator plan by reasoning in natural language over the public state they observe. In *heuristic mode* every actor uses deterministic rules over the same observed state. The media actor and the consumer market are rule-based in both modes (Appendix B.8). The heuristic mode serves two purposes in this paper.

Distributional inference. An LLM run costs about \$1 to \$2 per 40-round run at the Sonnet prices in Appendix E.8, so we cap the per-condition sample at $N=10$ paired seeds for the primary Sonnet ladder, with

$N=3$ paired seeds each for cross-vendor coverage under Opus and GPT-5.5 (Appendix F.7). At these counts we report directional agreement and matched-seed paired comparisons, and we do not report tight confidence intervals on continuous outcomes. Heuristic runs are three orders of magnitude cheaper and allow $N=50$ seeds per condition, and the privacy-ladder distributional claims rest on them.

LLM-artifact control. A finding that appears in LLM mode and disappears in heuristic mode is evidence of LLM-specific behavior, which usually traces to prompt framing, to training-data priors, or to reasoning over structure in the benchmark-type labels. A finding that appears in both modes with the same sign and a comparable magnitude is evidence that the mechanism is structural, which is to say that it follows from the simulation’s information architecture and rule set. The privacy ladder in §5.1 follows this pattern, because the cross-mode agreement in Figure 3 is evidence that LLM label-reading does not drive the gap outcome. LLM providers do read the `private` benchmark-type label in their reasoning (Appendix F.3), and that reading does not drive the measured effect.

What heuristic mode does *not* represent. The heuristic planner (specified in Appendix B.6) is deliberately simple: it consumes the same `ecosystem_context` as the LLM planner (market share, incidents, interventions, media narrative, per-benchmark scores) and returns a portfolio allocation via four rule-based signals. It cannot represent several dynamics that LLM planning can:

- **Strategic reasoning about benchmark-type labels.** Heuristic providers do not read the `public` / `partial` / `private` label. They respond to the noisier observed score under private conditions via the same belief-update rule they use for public scores. Any reading of the benchmark-type label by LLM providers would register as a mode difference.
- **Identity-driven behavior.** Provider strategy profile enters heuristic mode only through the initial portfolio and the belief-update learning rate ($\eta \in \{0.10, 0.15, 0.20\}$); it does not affect which signals a provider weights in allocation decisions. LLM mode exposes a full identity block to each provider prompt (Appendix D.1).
- **Re-prioritization within the benchmark portfolio.** The per-benchmark priority `focus_level` changes only in LLM mode, and in heuristic mode each provider keeps its initial focus levels for the whole run. The blend weight `benchmark_orientation` is fixed at 0.80 for every provider in every reported run, in both modes (Appendix B.6). Heuristic results therefore hold both levers constant, and they read as a structural baseline.
- **Evaluator dynamic mode.** In dynamic mode the heuristic evaluator picks the next benchmark from the pool with a coverage-gap rule, and the LLM evaluator chooses from the pool itself (Appendix B.3). The results in this paper do not compare the two.

Cross-mode agreement and divergence. The privacy ladder shows cross-mode *agreement*, in that the structural mechanism carries the effect and LLM reasoning adds no further signal. The evaluator-capture condition (Appendix H.1) shows cross-mode *divergence*: market concentration rises in the heuristic runs and falls in the two LLM runs, and neither effect is established at these sample sizes. We report divergences of this kind alongside the matched-sign findings.

F.2 Privacy ladder: extended heuristic results

The main-body Figure 3 shows five representative benchmarks (one per primary capability dimension) across the five privacy conditions. Figure 7 below extends that to all 13 benchmarks in the active suite, including the two agentic-calibration outliers that the main-body forest excludes (Agentic Tasks and Function Calling), so a reader can inspect the gap pattern on the full set. Heuristic and LLM mode agree on sign for 12 of 13 benchmarks, and the one disagreement is Adversarial Robustness, which sits within sampling noise of zero in both modes. In heuristic mode Long Context also sits at the boundary, because its $|g|$ changes by less than 0.001 between `public_only` and `private_only`, so whether it counts as shrinking or widening depends on the run subset. The rest of this subsection adds per-benchmark and per-provider heuristic decompositions and sets out the robustness checks behind the weight-asymmetry mechanism claim.

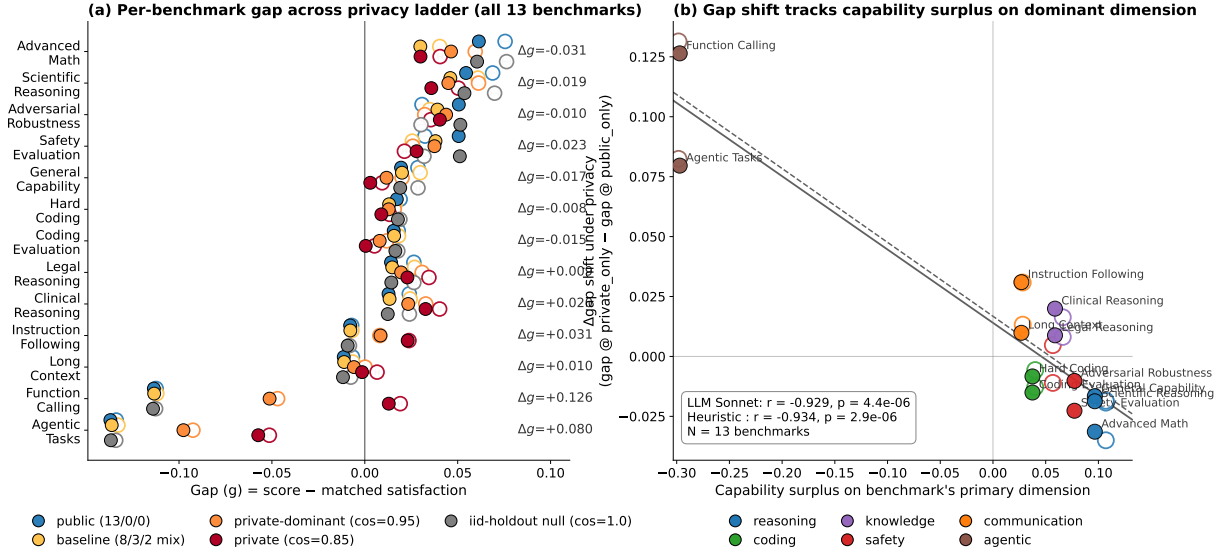


Figure 7: Extended version of Figure 3: same two-panel layout, but both panels now show all 13 benchmarks in the active suite, including the two agentic-calibration outliers (Agentic Tasks, Function Calling) excluded from the main-body forest and panel-b regression. Visual conventions identical to the main-body figure. Direction of Δg agrees across LLM and heuristic modes on 12 of 13 benchmarks; the lone disagreement is Adversarial Robustness, within sampling noise of zero in both modes. Panel (b)’s regression is fit on all 13 points; the two outliers anchor the high- $|\Delta g|$ end and tighten the slope substantially relative to the 11-benchmark main-body version.

Table 18: Fraction of seeds in which the absolute gap $|g|$ widens relative to `public_only` at the same seed, per benchmark. Under `private_only`, widening is near-deterministic for the benchmarks that widen and absent for most others; under the `iid_holdout` null (same weights, private reporting) every benchmark sits near a coin flip. Computed by `scripts/plots/paper/holdout_widening_table.py`.

Benchmark	LLM Sonnet 4.6 ($N=10$)		Heuristic ($N=50$)	
	<code>private_only</code>	<code>iid_holdout</code>	<code>private_only</code>	<code>iid_holdout</code>
General Capability	0.00	0.40	0.00	0.44
Coding Evaluation	0.00	0.70	0.00	0.42
Safety Evaluation	0.00	0.70	0.08	0.50
Instruction Following	1.00	0.70	1.00	0.68
Scientific Reasoning	0.00	0.40	0.04	0.56
Clinical Reasoning	1.00	0.50	1.00	0.52
Adversarial Robustness	0.00	0.60	0.76	0.46
Hard Coding	0.00	0.60	0.04	0.46
Agentic Tasks	0.00	0.50	0.00	0.50
Advanced Math	0.00	0.40	0.00	0.50
Function Calling	0.00	0.50	0.00	0.52
Long Context	0.00	0.60	0.62	0.62
Legal Reasoning	1.00	0.60	0.96	0.40

Suite dependence of the widening set. Across the nine heuristic benchmark sequences of Appendix F.5 (30 seeds each, a separate batch from the 50-seed runs of Table 18), Instruction Following widens in every seed of every sequence, including S4, whose round-0 suite adds a creative-writing benchmark. Clinical and Legal Reasoning widen in 97% and 93% of these seeds under the default suite (S0; 100% and 96% in the 50-seed batch) but in none under S6, whose round-0 suite adds a domain-expertise benchmark; under S3, which adds a hard-knowledge benchmark, they widen in 70% and 87%. The knowledge-centered wideners

therefore follow suite composition, while the communication-centered one does not within the suites tested. Computed from `output/paper/holdout_widening_by_roster.csv`.

Robustness to early benchmark introductions. A rule-based stall trigger introduces a scheduled benchmark early in about two thirds of runs, and all 13 benchmarks are active by round 36 in every run (Appendix B.3). Splitting the runs into those with no early introduction and those with at least one leaves the LLM shrink count at 10 of 13 in both subsets. In heuristic mode the count is 9 of 13 and 8 of 13, and the single benchmark that differs is Long Context, whose $|g|$ changes by less than 0.001 between the two conditions.

Per-benchmark and per-condition reliability. Figure 8 pools 87,617 market-leader scoring rows across all five privacy conditions and plots each as a point in (benchmark score, matched consumer satisfaction) space; deviation below the diagonal indicates the benchmark overpromises versus consumer experience. Panel (a) colors points by benchmark, and its legend reports the per-benchmark overpromise rate; panel (b) colors by privacy condition. Reliability varies by benchmark, from low single digits to the mid-teens of overpromise rate, which is consistent with the per-dimension structure in §5. Reliability also varies by privacy condition, because the `public_only` cloud has the highest overpromise rate while the `baseline` and `private_only` clouds sit closer to the diagonal.

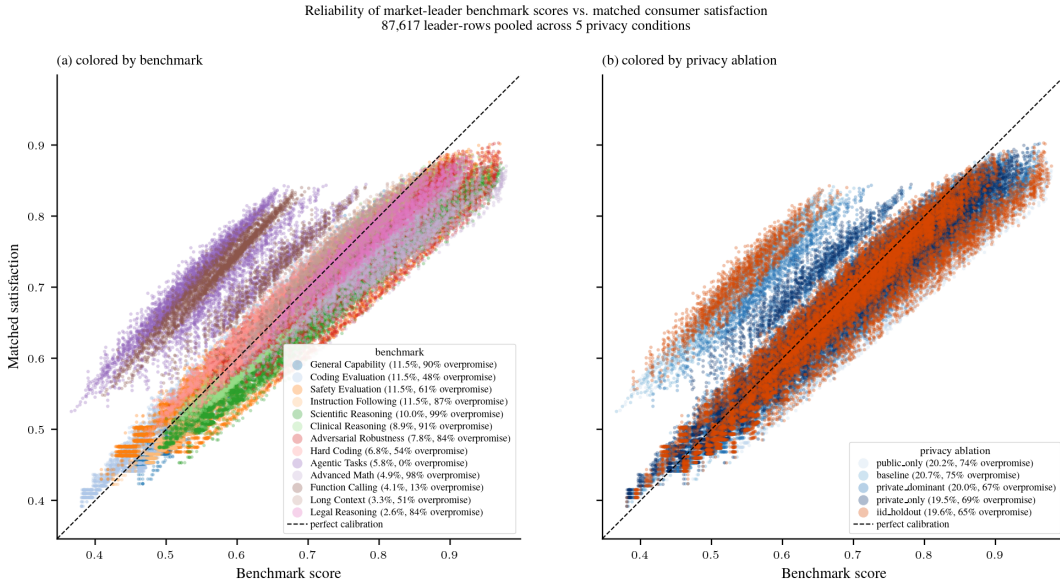


Figure 8: Heuristic privacy ladder: market-leader benchmark-score vs. matched consumer-satisfaction reliability, pooled across 87,617 leader-rows from $N=50$ seeds per condition $\times$ 5 privacy conditions. Diagonal = perfect calibration; below-diagonal = benchmark overpromises. (a) colored by benchmark with per-benchmark overpromise rate; (b) colored by privacy condition with per-condition overpromise rate. The reliability is consistent with the aggregated cloud in Figure 3 and the dimensional structure of §5.

Per-provider reliability. Figure 9 renders the same scatter but colored by market-leader identity, restricted to whichever provider holds the leader role in each (seed, round). The reliability pattern is preserved across the providers that appear as leaders; the figure’s purpose is to rule out the possibility that the aggregate calibration drift is driven by a single dominant provider.

Robustness check 1: randomized benchmark-type assignment. The calibrated baseline places Safety Evaluation and Scientific Reasoning in the `partial` slot by design (Table 4). To isolate the privacy mechanism from this assignment, we re-run the heuristic layer with benchmark-type assignments shuffled per seed: each of

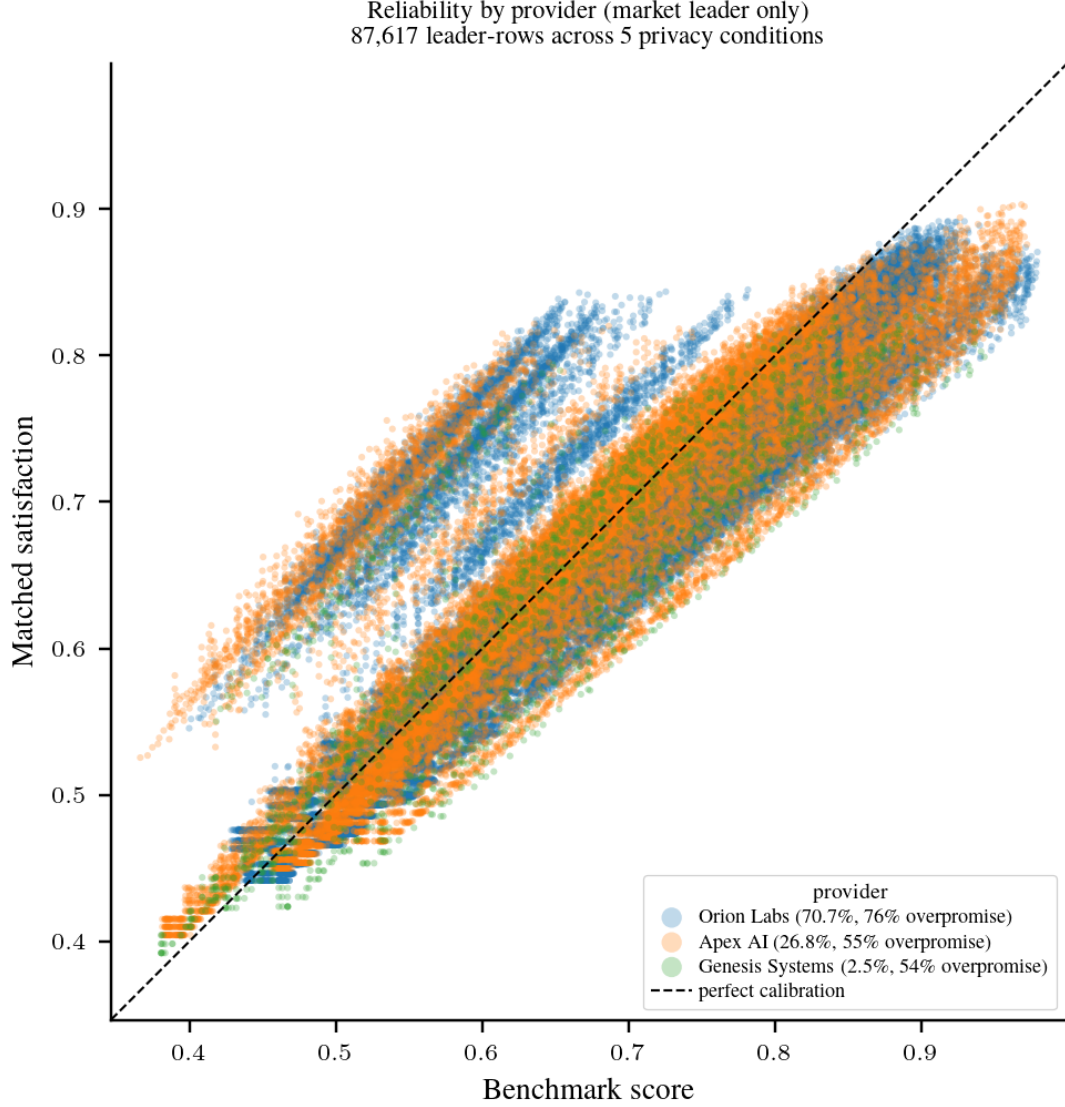


Figure 9: Heuristic privacy ladder, market-leader scoring rows colored by provider identity ($N=50$ seeds per condition). The reliability pattern (deviation from the diagonal) holds across the providers that hold the market-leader position; the aggregate finding is not driven by a single provider’s heuristic response.

the 13 benchmarks is assigned **public** / **partial** / **private** with the calibrated ratio (8:3:2), with assignments deterministic in the seed so paired comparisons remain valid. The randomized-assignment condition yields a per-benchmark gap compression of ≈ 0.013 to 0.017 , which is consistent with the monotone shift along the perturbation axis and which separates the privacy effect from the aggregate **partial**-against-**baseline** comparison, where assignment and privacy effects are pooled.

Robustness check 2: within-benchmark counterfactual. A second check fixes the calibrated baseline and re-scores each **partial** benchmark under its public dimension weights (rather than its holdout weights), holding the provider capability vector constant. The comparison holds the reporting cadence and the benchmark identity fixed and varies only the weight asymmetry. Across 30 seeds the within-benchmark gap coefficient is ≈ 0.014 per benchmark, which sits within the range of the randomized-assignment coefficient above. The check gives the cleanest isolation of Channel 1 (weight distance) from Channel 2 (reporting lag) and Channel 3 (observation noise).

Robustness check 3: scoring-layer versus capability decomposition. The third check asks how much of Δg comes from scoring the same capabilities under different weights, and how much from capabilities evolving differently under privacy. For each of the $N=50$ matched heuristic seed pairs (`public_only`, `private_only`) and each benchmark, we re-score end-of-run capability without observation noise. With $\mathbf{C}_{\text{priv}}$ and $\mathbf{C}_{\text{pub}}$ the end-of-run capability matrices of the two runs, the scoring-layer effect is $S = \text{score}(\mathbf{C}_{\text{priv}}, \mathbf{w}_{\text{holdout}}) - \text{score}(\mathbf{C}_{\text{priv}}, \mathbf{w}_{\text{public}})$ and the capability effect is $K = \text{score}(\mathbf{C}_{\text{priv}}, \mathbf{w}_{\text{public}}) - \text{score}(\mathbf{C}_{\text{pub}}, \mathbf{w}_{\text{public}})$. Matched satisfaction does not depend on the scoring weights, so S passes into Δg one-for-one. The noiseless reconstruction matches the logged Δg to within 0.001 per benchmark.

S is significant at 95% on all 13 benchmarks and has the sign of Δg on 13 of 13. K is indistinguishable from zero on 11 of 13 (Advanced Math and Scientific Reasoning are marginally positive). As a share of summed absolute effects, the scoring layer carries 0.963 of Δg (0.926 without the two agentic-calibration outliers). End-of-run capability differs little between the two conditions (reasoning +0.006, every other dimension inside its confidence interval): in heuristic mode privacy changes how the same capabilities are scored, and barely changes where capabilities end up.

The identity $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$ must nonetheless be evaluated at realized capability. When we evaluate it at round-0 capability, taken as the mean of the six configured vectors, it matches the sign of Δg on 10 of 13 benchmarks and misses Safety Evaluation, Clinical Reasoning and Instruction Following. Mean capability rises from 0.36 to 0.58 over the run and tilts across dimensions, and two of the three stable wideners change sign along the way. The decomposition covers heuristic runs only. Computed by `scripts/plots/paper/holdout_rescoring.py`.

F.3 Qualitative check: LLM reasoning about the benchmark-type label

The heuristic planner cannot read the `public`, `partial` or `private` label (Appendix F.1), so the cross-mode agreement in Figure 3 is evidence that label-reading does not produce the gap outcome. This subsection asks the companion qualitative question, which is whether LLM providers read the labels at all. They do, consistently, and in terminology that matches the mechanism (“K=3 lag”), and the label-reading sits on top of the structural effect.

Sampling rule. We take the round-20 planning entry at seed 42 under Sonnet for three providers that span the archetype range (Orion Labs, a frontier lab; Apex AI, safety-oriented; OpenCore, open-source). For each provider we report its planning reasoning under `public_only` beside its reasoning under `private_only`, matched in seed and round. We inspected no other excerpts before selecting these.

Orion Labs, round 20.

- *public_only*: “Market share up 2.5pp, \$115M raised this month, first to 70% on Safety Evaluation and Scientific Reasoning. . . R&D stays at current levels — we’re achieving strong gains across General Capability (+0.014), Safety (+0.022), and Adversarial Robustness (+0.019).”
- *private_only*: “All scores are flat due to the K=3 lag on private benchmarks — this is expected and not alarming. Our market share grew to 67.1%, confirming our strategy is working.”

Apex AI, round 20.

- *public_only*: “The critical development this month is Orion Labs breaking 70% on Safety Evaluation — our primary differentiator — while outpacing us on General Capability, Coding, and Clinical Reasoning. Their 65M in fresh capital signals sustained acceleration.”
- *private_only*: “All private benchmarks showing +0.000 is expected given the K=3 reporting lag — our Month 17–19 investments haven’t surfaced yet. This means we should stay the course strategically rather than panic-pivot.”

OpenCore, round 20.

- *public_only*: “Orion Labs is pulling away dramatically — massive funding, breaking 70% on Safety and Scientific Reasoning. . . Our Hard Coding score (0.516) is actually our strongest competitive position — we lead Genesis, Mirage, Apex, and Spark there.”

- *private_only*: “All scores are flat due to the $K=3$ lag — the investments from months 17–19 should start showing up in months 20–21. The competitive gap vs. Orion Labs and Apex AI is significant and widening.”

What the excerpts show. All three providers, under *private_only*, open their reasoning by naming the $K=3$ lag mechanism using near-identical phrasing (“ $K=3$ lag”, “reporting lag”, “months 17–19 investments”) across independent planning calls. None of the matched *public_only* excerpts cite benchmark-type labels; they focus on competitor scores, funding rounds, and per-dimension R&D gains. This is direct qualitative evidence for the $\approx 32\%$ per-round privacy-framing rate reported quantitatively in Appendix E.7, and it is what Appendix F.1 identifies as LLM-specific behavior: strategic reasoning over structure the heuristic planner cannot see.

The cross-mode agreement claim in Figure 3 is that label-reading does not produce the gap outcome, because heuristic providers, which cannot read labels, produce the same per-benchmark ladder. It does not claim that LLM providers ignore the labels. The excerpts above fit that reading, because LLM providers reason about the mechanism explicitly while their portfolio decisions, visible in Figure 3 and Table 21, land in the same region as heuristic-planner decisions under the same condition.

F.4 Structural ablation: removing incidents

We ran structural ablations in LLM mode (Claude Sonnet 4.6), each of which removes one ecosystem component. These runs use an all-public benchmark suite, so we compare each run with the *public_only* run at the same seed. One ablation gives the same direction in every run we have, and we report that one. Without the incident stream (*no_incidents*), Orion Labs ends with 78% to 82% of the market in four runs at seeds 42, 43 and 44, against 59% to 74% in the matched runs, and HHI is higher by 0.09 to 0.23. The result agrees with Appendix C.8, where the timing of the first critical incident separates the baseline runs.

The other ablations remove the regulator, the media actor, the funders or the open-source provider, collapse consumers into one segment, or equalize initial capability. Their effects on concentration were small, or they changed direction across seeds or across repeated runs of the same seed, so we draw no conclusion from them. LLM sampling is not deterministic, and one run per condition does not fix the sign of an effect of this size.

F.5 Null-model capability test

This subsection asks how much of the per-benchmark Δg pattern (§5.1) is recoverable from initial conditions through a growth model fit on simulation runs. The structural identity $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$ holds with $r > 0.99$ when $\mathbf{c}$ is taken from the simulation’s 40-round end-of-run capability vector. That fit is a tautology, because $\mathbf{c}$ is itself a simulation output. We therefore ask whether the simulation’s trajectory adds predictive content over a null fit on round-0 inputs.

Method. For each (sequence, rung, seed) run we extract: (i) round-0 capability vector $\mathbf{c}_0(p, d)$ per (provider, dimension); (ii) round-0 R&D / safety / product portfolio $\pi(p)$ per provider; (iii) end-of-run capability $\bar{\mathbf{c}}_T(p, d)$ averaged over the last 5 rounds. We fit a per-(provider, dimension) Ridge regression $\bar{\mathbf{c}}_T(p, d) \approx f_{p,d}(\mathbf{c}_0(p, d), \pi_\ell(p))$ with two features: the round-0 capability scalar in the same dimension and the portfolio fraction at the lever ℓ that drives growth in that dimension ($\ell=\text{rd}$ for non-safety dims, $\ell=\text{safety}$ for safety). We then propagate the null-predicted capability into a null Δg via the same identity above and compare the resulting null prediction against the observed Δg via 5-fold CV by seed. Both null and sim predictions are evaluated against the same observed Δg (segment-weighted gap from *rounds.jsonl*, paired *private_only* – *public_only* per matched seed).

Cross-sequence LLM result (primary). Table 19 and Figure 10 report pooled 5-fold CV metrics on the cross-sequence LLM dataset (Claude Sonnet 4.6, 9 (*seq, rung*) cells: the S0 baseline composition across the full five-condition ladder plus the s5_aligned and s8_agentic sequences at the *public_only*/*private_only* endpoints (5+2+2), with 3–10 seeds per cell, for 62 runs and $n=208$ paired-prediction cells). The null model alone explains 97.4% of Δg variance. The simulation’s contribution over and above the null is +0.2 pp r^2 (+0.001 to +0.005 across folds, no fold flips the sign).

Table 19: Null-vs-sim Δg prediction on the LLM cross-sequence dataset (Claude Sonnet 4.6, $n=208$ paired cells, 5-fold CV by seed). The null model is fit on round-0 capability + portfolio per (provider, dim); the sim model uses the simulation’s 40-round end-of-run capability. Both feed the same $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$ identity.

Model	r	r^2	RMSE from $y=x$	MAE
Null (round-0 initial conditions)	0.987	0.974	0.0097	0.0072
Sim (40-round end-of-run capability)	0.988	0.976	0.0091	0.0070
Sim – Null	+0.001	+0.002	–0.0006	–0.0002

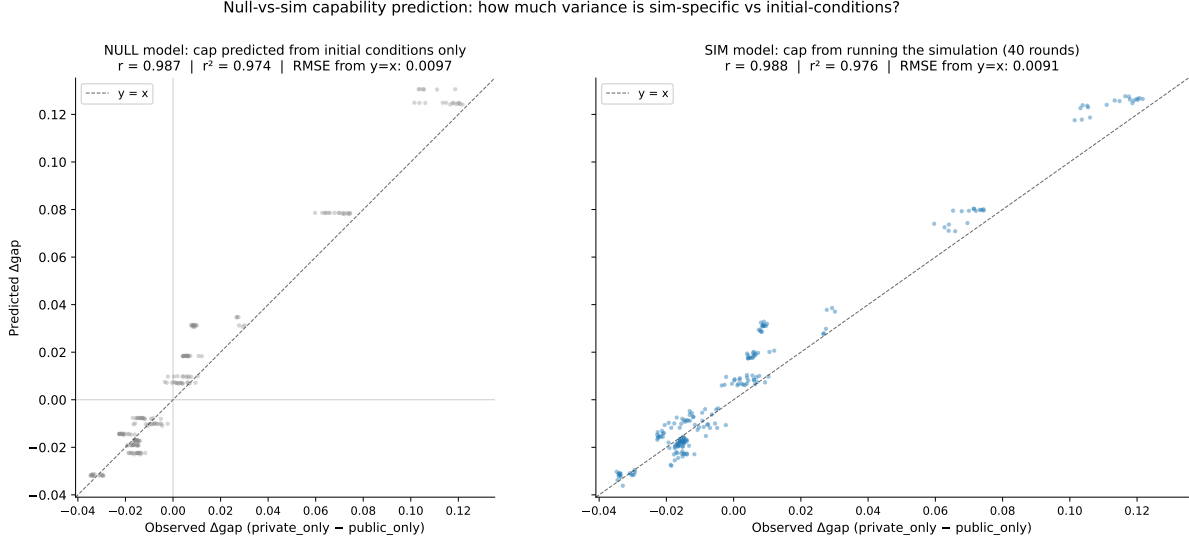


Figure 10: Null vs. sim prediction of Δg on the cross-sequence LLM dataset ($n=208$). Left: null model fit on round-0 initial conditions. Right: sim model using the simulation’s 40-round end-of-run capability. Both panels track the $y=x$ diagonal closely, and both achieve $r^2 \approx 0.97$. The simulation’s 40-round trajectory adds +0.2 pp r^2 over the fitted null.

For comparison, the same procedure on the heuristic-mode dataset ($n=3000$, S0–S8 sequences, 30 seeds) produces null $r^2=0.880$ and sim $r^2=0.922$ ($\Delta r^2 = +0.042$). Heuristic providers’ stochastic safety efficiency, random breakthroughs, and noise-driven focus updates produce capability variance that initial conditions do not capture; LLM capability paths are more predictable from initial conditions than heuristic ones. LLM planning therefore adds less trajectory variance than noisy rule-based planning, the reverse of what strategic reasoning might be expected to produce.

Path A: variance decomposition under capability ablation. The null model’s central signal is provider-level heterogeneity in round-0 capability. As a non-parametric stress test, we compare cross-provider variance of mean per-(provider, benchmark) gap between (i) regular baseline ($n=10$ Sonnet seeds with heterogeneous initial caps) and (ii) the same condition under the `initial_uniform_capability` structural ablation, which flattens all providers to the population-mean capability vector at $t=0$ ($n=2$ Sonnet seeds; distinct from the later 3-seed batch launched for Path B below). Under the ablation, any cross-provider gap variance must come from simulation dynamics (strategic investment divergence, breakthroughs, regulator targeting), since the null’s main signal source is removed.

The aggregate cross-provider σ^2 collapses from 0.000491 (regular) to 0.000037 (uniform-cap), a ratio of 0.075. The simulation-attributable share of cross-provider gap variance is therefore $\sim 7.5\%$; the null-attributable (initial-cap heterogeneity) share is $\sim 92.5\%$. Per-benchmark ratios (Figure 11, left panel) cluster in 0.01–0.13 for most benchmarks. Two notable exceptions surface as boundary cases: *Adversarial Robustness*

(ratio 0.451, $\sim 45\%$ sim contribution) and *Advanced Math* (ratio 0.185). The Adversarial Robustness boundary is consistent with strategic safety-investment divergence across providers being a source of provider-level differentiation that the fitted null cannot capture. At $n=2$ seeds the per-benchmark ratios are indicative, and the aggregate ratio is the firmer quantity.

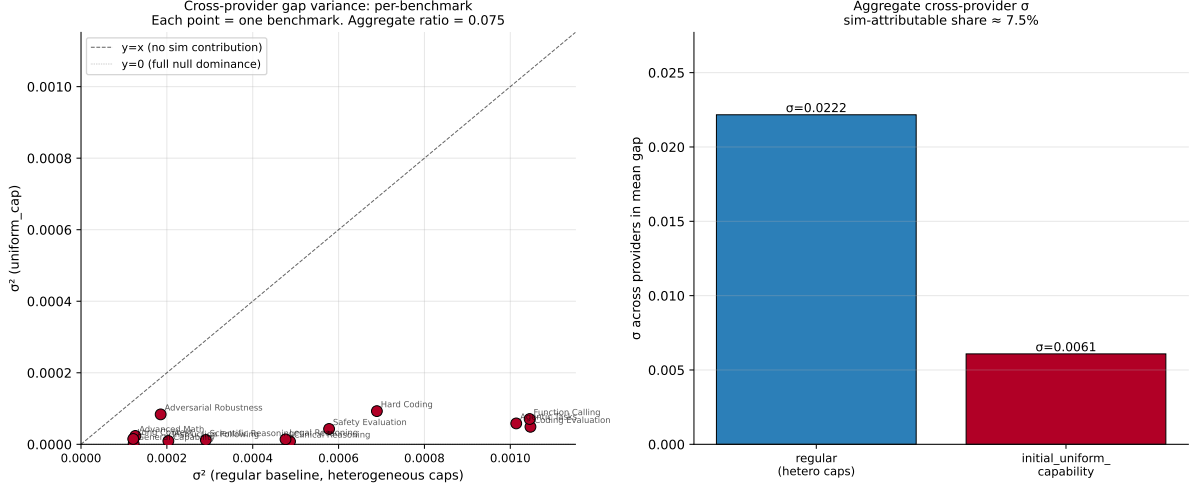


Figure 11: Path A variance decomposition. Left: per-benchmark cross-provider σ^2 under regular baseline (x) vs. `initial_uniform_capability` ablation (y). The ablation has $n=2$ seeds, so each y value is the variance of two runs. Points at $y=0$ would mean full null dominance; points on $y=x$ would mean sim contributes everything. Most benchmarks cluster low along y , with Adversarial Robustness as an outlier. Right: aggregate cross-provider σ^2 comparison (regular vs. uniform-cap), showing the $\sim 13\times$ collapse in variance under ablation (equivalently, $\sim 3.6\times$ collapse in σ). Aggregate ratio $\sigma_{\text{uniform}}^2/\sigma_{\text{regular}}^2 = 0.075$.

Path B: held-out null replication under capability ablation. Path A’s variance decomposition is a within-data summary statistic. As a tighter test, we ask whether the null model trained on the 62 heterogeneous-cap LLM runs of the primary dataset above still predicts a held-out test set drawn entirely from a regime where its training-set heterogeneity is removed. The held-out test set is 6 LLM Sonnet runs at `public_only` and `private_only` crossed with seeds $\{42, 43, 44\}$ under the `initial_uniform_capability` ablation. In those runs the cross-provider standard deviation of round-0 capability is 0.000000 in every dimension, and pairing on (seed, benchmark) gives $n=39$ prediction cells.

Trained on the heterogeneous data, the null still predicts the held-out uniform-cap test set at $r^2=0.955$, which is about 2 pp below the heterogeneous-cap reference of 0.974 (Table 20, Figure 12). The simulation’s marginal contribution grows from +0.2 pp on the heterogeneous set to +2.0 pp on the uniform-cap set, a tenfold relative increase that remains small in absolute terms.

Table 20: Path B held-out null replication. The null model is fit on the 62 heterogeneous-cap LLM runs of the primary dataset, then evaluated on a held-out test set of 6 new LLM runs under the `initial_uniform_capability` ablation. The null’s r^2 drops only ~ 2 pp despite test-set initial-capability heterogeneity being structurally zero.

Test condition	n	r_{null}^2	r_{sim}^2	Δr^2 (sim – null)
Reference: heterogeneous caps (S0+s5+s8)	208	0.974	0.976	+0.002
Held-out: uniform caps (<code>initial_uniform_capability</code>)	39	0.955	0.975	+0.020

Triangulation. The three tests measure complementary quantities and point the same way: the pooled cross-sequence LLM comparison ($\Delta r^2 = +0.002$), the Path A variance decomposition (σ^2 ratio 0.075) and

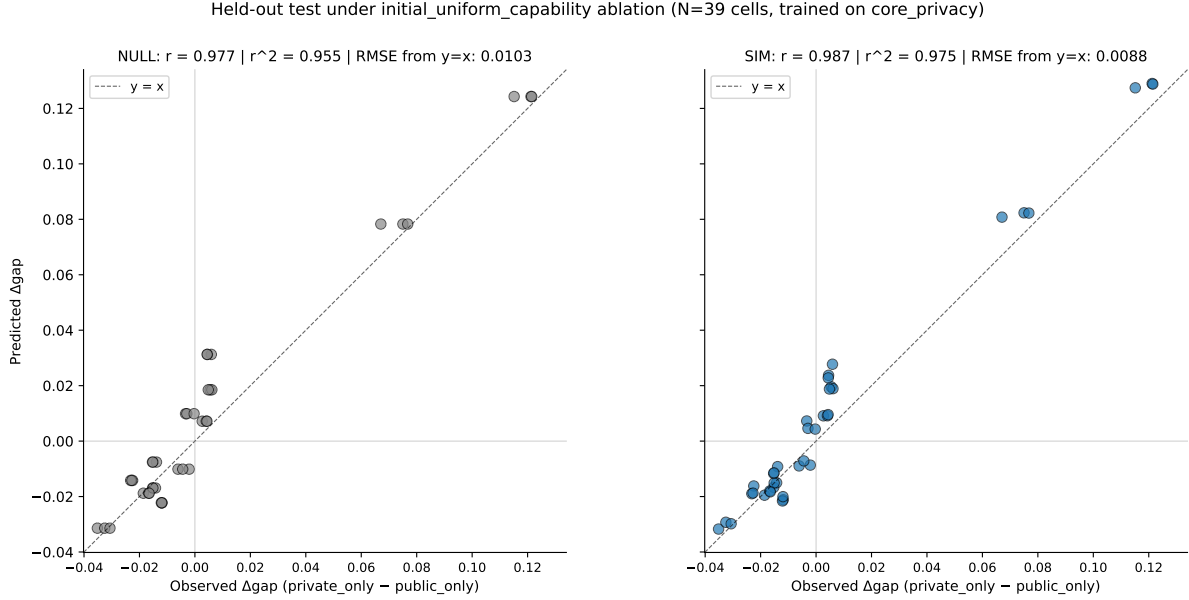


Figure 12: Path B held-out null replication ($n=39$ test cells under `initial_uniform_capability`). Null model is trained on 62 heterogeneous-cap runs and predicts the held-out uniform-cap test set. Both null and sim track the $y=x$ diagonal closely; null $r^2 = 0.955$, sim $r^2 = 0.975$.

the Path B held-out replication ($r^2_{\text{null}} = 0.955$). The predictive content of the privacy mechanism rests on two pieces, the closed-form identity $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$ and a growth model, fit on simulation runs, that predicts end-of-run capability $\mathbf{c}$ from round-0 capability and portfolio. The simulation contributes a small residual on top, and its size depends on where it is measured: it is negligible on the pooled cross-sequence set, about 2 pp on the uniform-cap set, and concentrated on benchmarks such as Adversarial Robustness, where strategic safety-investment divergence is a primary differentiator. Most of the pooled variance lies between benchmarks, where the null and the simulation use the same fixed weight differences, so the pooled fit does not test what the trajectory adds. The identity is fixed by how scores and satisfaction are defined, and it needs realized capability, because evaluated directly on round-0 capability it mis-signs three benchmarks (Robustness check 3, Appendix F.2). The predictability of end-of-run capability from round-0 conditions is an empirical property of these runs, and Path B checks it on held-out runs. The cross-mode sign difference on market concentration under evaluator capture (Appendix H.1, LLM $N=2$ against heuristic $p \approx 0.20$) may mark an outcome that needs the full system, and it is too small to establish.

Caveats. The test-set sample sizes for the structural-ablation robustness checks are small. Path A’s per-benchmark variance ratios rest on $n=2$ uniform-cap seeds, which gives wide confidence intervals on individual benchmarks and leaves the aggregate ratio of 0.075 as the firmer claim, and Path B’s $n=39$ paired cells sit at the low end of what supports an r^2 estimate. Bootstrap confidence intervals on both quantities would tighten the claim, and the directional conclusion does not depend on them. The null model uses two features per (provider, dimension). A richer null with portfolio interactions or round counts could raise r^2_{null} further, which would narrow the simulation’s contribution, so the null we use is the conservative choice for the claim we make.

F.6 Cross-model setup and scope

The Tier 1 privacy ladder consists of five conditions (`public_only`, `baseline`, `private_dominant`, `private_only`, `iid_holdout`) executed at multiple seeds. Sonnet runs cover the full ladder at seeds 42–51 ($n=10$

per condition); Opus and GPT-5.5 each cover the full ladder at seeds 42, 43, and 44 ($n=3$ per condition).² Table 21 lists the runs that entered this appendix as matched triples, which are the same (condition, seed) cell run under all three models, so that per-benchmark gap, market structure and reasoning traces can be compared directly across vendors.

Coverage caveat. The matched-triple structure has $n=3$ per condition for the cross-model comparison (seeds 42, 43, and 44 paired across all three models in every privacy condition), and $n=10$ on Sonnet alone (seeds 42–51). The error bars in Figure 13 therefore rest on $n=10$ in the Sonnet panel and on $n=3$ in the Opus and GPT-5.5 panels. The comparison works at the outcome level, which is enough to test direction and ordering and to detect clear cross-model differences in magnitude, while small differences in a single (condition, seed) cell remain noisy. Where we report a quantitative cross-model difference at a single cell, we report it as directional and we do not test it statistically. Section E.7, by contrast, aggregates over all fifteen matched triples and pools many planning-stage entries per (model, provider) cell, so frame-share comparisons are pooled estimates rather than single-condition reads.

We draw on three windows into each run: end-of-run endpoint metrics (share-weighted final-5-round means), per-benchmark clean-gap trajectories, and the full reasoning transcripts of the six providers. The endpoint and per-benchmark analyses use the same construction as the heuristic privacy-ladder extended results above; the reasoning-frame analysis is presented in Appendix E.7.

F.7 Privacy-ladder robustness

Figure 13 plots the per-benchmark clean gap under each privacy condition, one panel per model, with benchmarks ordered by Sonnet `public_only` descending.

The privacy-condition ordering replicates across all three planners. On the high-gap benchmarks (Advanced Math, Scientific Reasoning, Safety Evaluation, Adversarial Robustness), all three models show the same ladder, with `public_only` (blue) at the largest gap, `private_only` (dark red) compressed nearest to zero, and `private_dominant` (orange) and `baseline` (yellow) between them. The main-body claim that privacy conditions compress the per-benchmark gap toward zero therefore holds under all three planners and across two vendors. On the underpromise-side benchmarks (Instruction Following, Long Context), all three planners produce the same negative `public_only` signature that §5 documents.

GPT-5.5 sits closer to Sonnet than Opus does. The paired per-benchmark view at seed 42 (Figure 14) shows that on most matched conditions the GPT-5.5 dot (green square) and the Sonnet dot (blue circle) sit within a few units of each other, while the Opus dot (red diamond) sits apart, often closer to zero on the high-gap benchmarks and further from zero on the underpromise-side benchmarks. On most rows the within-vendor gap between Sonnet and Opus is as large as the cross-vendor gap between Anthropic and OpenAI. Vendor identity is therefore not the dominant axis of variation in our data.

Market structure differs across models, and it does not vary monotonically. The matched-triple endpoint table (Table 21) shows three differences. Opus runs produce lower mean safety allocations across the ecosystem (≈ 0.20 to 0.24 , against 0.26 to 0.30 for Sonnet and 0.26 to 0.32 for GPT-5.5). HHI varies without a consistent ordering across models within a single condition; under `baseline` at seed 43 it is 0.528 for Sonnet, 0.480 for Opus and 0.374 for GPT-5.5. The five-round-mean gap can change sign across models on the same (condition, seed) cell at the third decimal. None of these differences fall within the scope of the privacy-ladder claim, and at $n=3$ per condition they support direction only.

F.8 Path dependence across models

Figure 15 overlays single-provider market-share trajectories across (seed, model) cells under `baseline`, with every other LLM run in the core privacy set drawn in grey. The highlighted lines are seeds $\{42, 43, 44\}$ crossed with the three models, with colour identifying the model and line style identifying the seed, and all nine cells are populated. The figure tracks Orion Labs, the same provider as the incident path-dependence figure (Figure 5, Appendix C.8), so a reader can compare dispersion across seeds and privacy conditions under one model against dispersion across models and seeds under one condition.

²Opus and GPT-5.5 coverage is narrower than Sonnet coverage because each is roughly 1.7 to $2\times$ more expensive per token than Sonnet at published API pricing, so we used Sonnet for breadth and Opus and GPT-5.5 for the paired-seed cross-model check. Appendix E.8 gives the wall-clock and pricing detail.

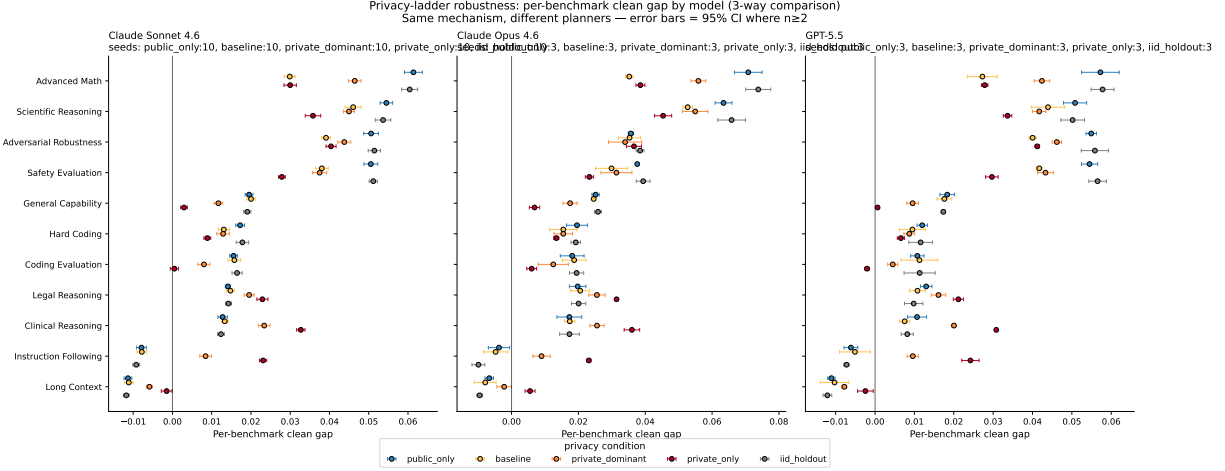


Figure 13: Per-benchmark clean gap by privacy condition, one panel per model. Benchmarks sorted by Sonnet `public_only` mean descending so the panels share a reading axis. Error bars are 95% CIs across seeds ($n=10$ Sonnet seeds 42–51, $n=3$ Opus and GPT-5.5 seeds 42–44). The privacy-condition ordering (public at the largest gap, `private_only` most compressed on high-gap benchmarks) replicates under all three planners.

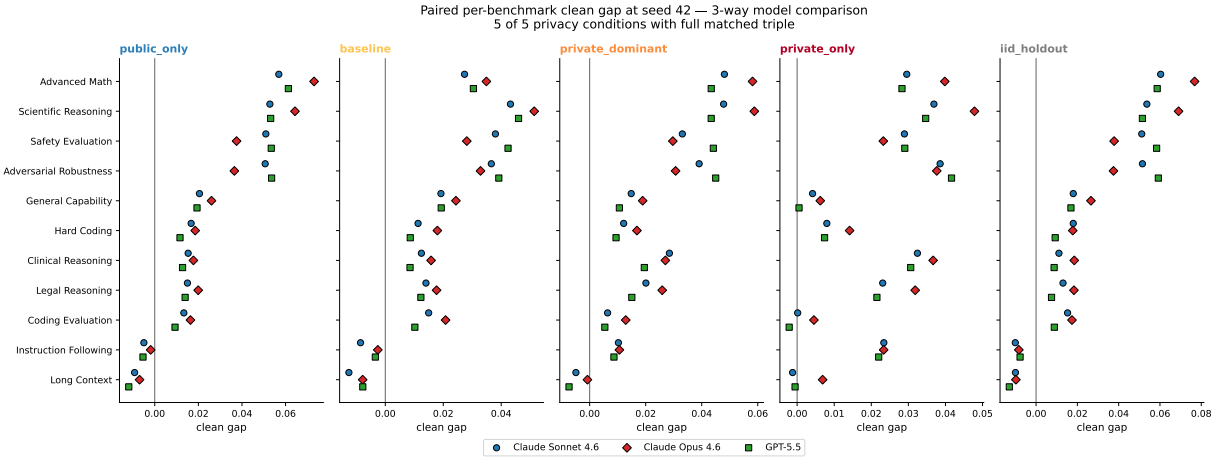


Figure 14: Paired per-benchmark clean gap at seed 42. One panel per privacy condition with a matched 3-way Sonnet+Opus+GPT-5.5 triple; circle = Sonnet, diamond = Opus, square = GPT-5.5. Dots are provider-mean gap per benchmark. All five privacy conditions have a complete 3-way matched triple at seed 42.

Dispersion within a model across seeds (Sonnet at ≈ 0.63 , 0.66 and 0.46 across seeds 42, 43 and 44; GPT-5.5 at ≈ 0.71 , 0.41 and 0.74) and dispersion across models at a fixed seed (at seed 44, Sonnet ≈ 0.46 , Opus ≈ 0.68 , GPT-5.5 ≈ 0.74) are of comparable magnitude, and both span ranges of ≈ 25 to 30 pp in final-round market share. The envelope traced by the full $N=10$ Sonnet ladder and by the three-way matched seeds is consistent with Appendix C.8, where the timing of the first critical incident separates the runs. Opus clusters tightly across seeds (≈ 0.68 to 0.69 on all three) while Sonnet and GPT-5.5 spread widely. Three seeds do not tell us whether that reflects an Opus-specific response to incidents or a small-sample artifact.

What the cross-model robustness check does *not* claim. We do not claim that any of the three models is more or less “correct.” At $n=3$ matched triples per condition we read direction and ordering as supported, and we make no claim about whether small endpoint-level differences in gap or HHI between models at a single (condition, seed) cell are stable, because three seeds per condition is narrow coverage. We

Table 21: Endpoint metrics for the 15 matched 3-way triples: all five privacy conditions at seeds 42, 43, and 44. All values are share-weighted means over the final 5 rounds except *HHI* (market Herfindahl). *gap* = score – consumer satisfaction. The prose above gives the per-model range of mean safety allocation, which we drop from the table to keep it within the page width. **S** = Sonnet 4.6, **O** = Opus 4.6, **G** = GPT-5.5.

Condition	Seed	gap			score			HHI		
		S	O	G	S	O	G	S	O	G
public_only	42	+0.000	+0.001	+0.021	0.705	0.706	0.695	0.514	0.500	0.486
public_only	43	+0.019	+0.071	+0.044	0.685	0.730	0.661	0.321	0.469	0.342
public_only	44	+0.008	+0.043	+0.009	0.689	0.711	0.694	0.540	0.403	0.588
baseline	42	−0.009	−0.008	−0.006	0.657	0.701	0.667	0.410	0.468	0.472
baseline	43	+0.073	+0.043	+0.060	0.708	0.730	0.671	0.528	0.480	0.374
baseline	44	+0.009	+0.022	−0.005	0.659	0.712	0.667	0.294	0.440	0.541
private_dominant	42	+0.030	+0.039	+0.008	0.726	0.718	0.684	0.581	0.380	0.365
private_dominant	43	+0.018	+0.015	+0.017	0.693	0.698	0.661	0.390	0.361	0.320
private_dominant	44	+0.011	+0.035	+0.003	0.664	0.705	0.693	0.268	0.322	0.552
private_only	42	+0.005	+0.013	+0.029	0.690	0.733	0.692	0.332	0.469	0.471
private_only	43	+0.022	+0.041	+0.049	0.698	0.764	0.675	0.399	0.574	0.377
private_only	44	+0.013	+0.050	+0.014	0.668	0.726	0.663	0.272	0.408	0.299
iid_holdout	42	+0.020	+0.027	+0.000	0.694	0.702	0.663	0.503	0.363	0.330
iid_holdout	43	+0.020	+0.082	+0.040	0.664	0.730	0.656	0.299	0.440	0.344
iid_holdout	44	+0.002	+0.016	+0.022	0.655	0.691	0.669	0.262	0.287	0.405

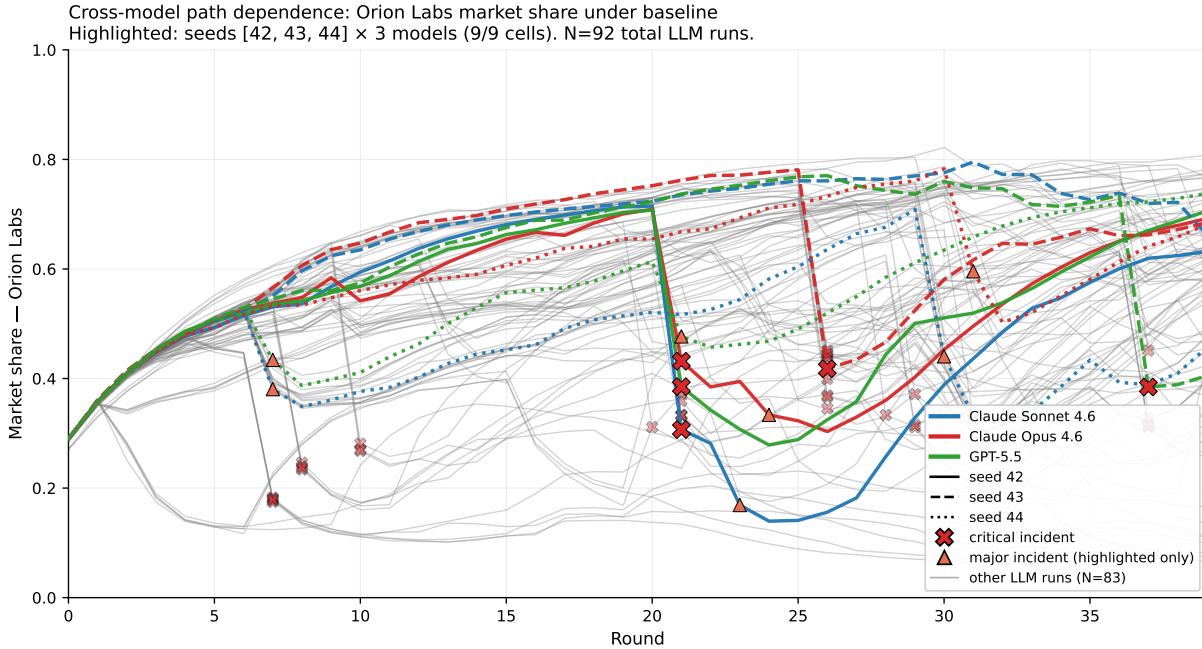


Figure 15: Cross-model market-share path dependence under **baseline** for Orion Labs. Highlighted: seeds {42, 43, 44} × three models (color = model, linestyle = seed); all nine cells are populated. Background grey: every other full-40-round LLM run in the core privacy set, across all conditions, seeds and models. Critical incidents marked on every line; major incidents on highlighted lines only. Figure 5 shows the same provider under one model across five privacy conditions; this figure holds the condition fixed and varies the model.

do claim that two further frontier LLMs replicate the privacy-mechanism finding of §5 at the behavioral level, because the condition ordering and the sign of gap compression on high-gap benchmarks hold across Sonnet, Opus and GPT-5.5. At the reasoning level the claim is partial, because both Anthropic planners invoke the K -lag mechanism explicitly in matched planning rounds while GPT-5.5 reaches the same operational conclusion with much less explicit reasoning over the mechanism (Appendix E.7). Two systematic differences remain, a lower mean safety allocation under Opus and a stepwise fall in the incident-framing rate from Sonnet to Opus to GPT-5.5. Both are consistent with reasoning styles that differ by vendor and by model family, and neither shows the privacy mechanism failing to generalize.

F.9 Methodological caveats

LLM seed coverage. With ten paired seeds per condition under the primary Sonnet ladder, and three paired seeds each under Opus and GPT-5.5 for cross-vendor coverage (Appendix F.7), we can compare matched runs and report directional agreement, while variance estimates on continuous outcomes (gap magnitude, HHI, per-provider trajectories) stay wider than the heuristic layer’s. Wherever the main text uses an LLM-derived number, its denominator is a within-run event count, such as an incident-response rate, or a matched-seed paired difference. We do not report an inter-run mean with a tight confidence interval from the LLM layer. Distributional claims rest on the heuristic $N=50$ layer throughout.

Paired-seed design. LLM runs within a comparison (e.g., `baseline` vs `private_only`) use the same set of paired seeds $\{42, 43, \dots, 51\}$ (ten under Sonnet; the first three for Opus and GPT-5.5 cross-vendor coverage). This controls the simulation’s internal random streams (incident draws, noise, market dynamics) across paired conditions, reducing inter-seed variance in paired differences. It does *not* control the LLM’s server-side sampling stochasticity: even at temperature zero, API responses are not perfectly deterministic. We rely on the $N=50$ heuristic layer for the residual server-side-noise contribution to variance estimates.

Fallback handling. An LLM call whose response fails JSON validation is retried with the same prompt, up to three attempts in total. When all three fail, the actor takes a fail-safe action for that round: a provider keeps the portfolio it chose in the previous round, a funder splits its allocation evenly, and the regulator falls back to its heuristic plan (Appendix E.3). The simulation prints a line for every call that reaches the fail-safe, and we read those lines from the run console, because the saved run files do not record a fallback count. In the LLM privacy-ladder runs, fewer than 2% of calls per run reached the fail-safe. We re-queued any run with a higher rate instead of including it in the paper layer.

G Limitations

This appendix gives the full treatment of the limitations that the main text summarizes. Appendix G.1 expands the four numbered modeling-choice limits of §5.3, Appendix G.2 lists the per-actor action-space choices we did not implement, and Appendix G.3 sets out the methodological bounds of GABMs.

G.1 Depth on the four main-body modeling-choice limits

§5.3 states four foundational modeling choices on which the privacy-ladder mechanism rests; this subsection expands each with material that did not fit in the main body.

1. Privacy operates through weight perturbation in this simulator. The simulation’s privacy mechanism decomposes into three channels (Appendix C.4): a cosine-distance perturbation $\cos \theta$ between public and holdout dimension weights, a $K=3$ reporting lag, and observation noise scaling with $\sqrt{\text{samples} \times h}$. The cosine channel is a *composite proxy* aggregating at least three mechanistically distinct phenomena: (i) item-level contamination and memorization of public splits during training [Haimes et al., 2024, Xu et al., 2024]; (ii) “training on the test task” effects [Dominguez-Olmedo et al., 2025]; and (iii) adversarial holdout construction, where holdout items are sampled to test a skill mix that differs from the category label. A result attributed to weight asymmetry may be driven mainly by any one of these, or by their interaction, and the

simulator does not resolve them separately. Two further holdout-design phenomena lie outside the composite cosine abstraction. The first is item-level randomization within the same dimension distribution, which draws fresh items from the same task family without changing the dimension-weight mix, so it acts as observation noise and not as weight redistribution. The second is distribution shift away from the training distribution, where holdout items are out of distribution while the nominal dimension-weight target is unchanged. Within the cosine channel, the `iid_holdout` null ($\cos \theta = 1$) isolates cosine effect from lag and noise, but does not split the *scoring-layer* effect (same trained capability, scored under different weights) from the *R&D-direction* effect (cosine perturbation drifts inferred weights, providers re-target). A within-benchmark rescoring (fixed end-of-run capability vectors scored under both the public and the holdout weights of the same benchmark) identifies the scoring-layer share in heuristic mode at 0.93–0.96 of Δg (Appendix F.2); LLM runs are outside that decomposition (§5.3, item 1).

2. The six-dimension ontology. The dimensions (reasoning, coding, knowledge, safety, communication, agentic) are anchored to HELM- and LMSYS-style benchmark taxonomies and to where industry investment separates into distinct R&D pathways, such as domain corpora, instruction tuning, RLHF and tool-use trajectory RL. The choice of ontology could change the result in two ways. A *flatter* ontology, which collapsed reasoning and knowledge into a single general-capability axis, would erase the unevenness between the dimensions that receive investment and the dimensions that receive credit, and the holdout weight perturbation acts on that unevenness. Under a one-dimensional ontology the holdout cosine perturbation reduces to scoring noise. A *finer-grained* ontology, which separated mathematical reasoning from causal reasoning, or long-context coherence from instruction following, could move specific benchmarks from one side of the population mean to the other, which would move those benchmarks between the shrinking and widening sets. The mechanism holds within any ontology that has at least one asymmetry between a high-investment and a low-investment dimension, and it does not survive an arbitrary re-coordinatization of the capability space.

3. Public-benchmark-driven competitive contexts only. The instantiated ecosystem is leaderboard-driven and competitive across firms, it has a single jurisdiction, and it has six providers with no entry and no exit. We do not know how the mechanism engages in three other classes of context. The first is FDA-style gatekeeping, as in medical-device classification under 21 CFR 860 or EMA conditional approval, where mandatory pre-deployment evaluation under regulator-controlled holdouts replaces the leaderboard signal and clinicians stand between the model and the consumer. The second is defense procurement, where evaluation is bilateral, classified and tied to specific use cases, so no public leaderboard exists for the holdout mechanism to act on. The third is enterprise business-to-business sales with an internal proof of concept, where buyers run their own evaluations against deployment-specific data, which partly decouples adoption from public benchmark scores. That correlation between buyer-side internal benchmarks and the public set decides whether the decoupling dampens or amplifies the effect, and we do not estimate it. Internal-deployment evaluation infrastructure, such as Responsible Scaling Policies, Preparedness frameworks and the Frontier Safety Framework [Anthropic, 2023, OpenAI, 2023, Google DeepMind, 2024], dates from 2023 onward, sits beside the public-benchmark channel, and adds pressure on R&D direction that the simulation does not model.

4. Model-defined consumer satisfaction. Consumer satisfaction starts from $\mathbf{c}_p^\top \mathbf{w}_s$, with both $\mathbf{c}_p$ (provider capability) and $\mathbf{w}_s$ (segment need weights) held by the simulation, and incident and cost terms then adjust it (Appendix B.7). The per-benchmark gap uses the capability term alone. We assign each segment a need-weight vector over the six capability dimensions from sector-level evidence on what that profession uses AI for and what it is concerned about, together with occupation-level adoption rates in Bick et al. [2024] (Appendix C.1). Two issues remain open. First, the evidence fixes which needs dominate each profile, and the exact values are a judgment we make from it. A survey instrument that elicits relative weights directly from users, through forced choice, conjoint analysis or willingness to pay, would test that judgment, and no such instrument exists for AI capabilities today. Second, the shrink and widen counts depend on whose satisfaction defines the gap. Need weights cancel out of each benchmark’s shift, $\Delta g = \mathbf{c} \cdot (\mathbf{w}_{\text{holdout}} - \mathbf{w}_{\text{public}})$, and enter only the starting gap, through matched satisfaction. The starting gap’s sign decides whether a shift widens or shrinks the gap. Under uniform need weights ($\mathbf{w}_s = \mathbf{1}/d$ for every segment) matched satisfaction

reduces to mean capability, every starting gap rises, and one classification of 13 changes (Long Context; heuristic, $N=50$, end-of-run capability held fixed). Against safety-heavy segments alone four change: Coding Evaluation, General Capability and Hard Coding widen, and Instruction Following shrinks. For consumers whose needs equal a benchmark’s loadings, the public score tracks their satisfaction, so any shift larger than scoring noise widens their gap.

G.2 Per-actor action-space choices not implemented

The simulation’s actor space is a simplification. Each actor has a richer real-world strategy space than the simulation implements, and Table 22 lists what we left out of each one. The omissions do not affect the specific question we ask in §5.1. For the adjacent questions that the ecosystem frame can host, each entry marks a place where the architecture would need an extension before the question could be studied.

Table 22: Per-actor action-space choices: scope modeled in the simulation vs. extensions not implemented. Each row lists the strategy levers exercised by the actor alongside concrete real-world levers the architecture would need to extend to host an adjacent question.

Actor	Modeled	Not modeled
Model providers	• R&D / safety / product portfolio (sums to 1.0) • Per-benchmark focus levels and weight beliefs • Public communications (announcements, press releases) • Optional best-of-N submissions (evaluator-capture) • Six-dim capability vector with sqrt-budget capability gains	• Distillation or model merging across providers • Internal (non-published) benchmark suites that gate release decisions • Training-data vendor relationships and selective contamination of public benchmarks • Multi-model portfolios (one provider shipping several model sizes simultaneously) • Internal red-team findings that never become observable incidents
Consumers	• 51 segments (archetype $\times$ use-case) • Need weights assigned from sector-level evidence on AI use and concerns, with occupation-level adoption data • Switching cost + post-switch cooldown; dynamic enterprise-share growth • Incident-weighted satisfaction with sector-matched $2\times$ penalty • Media-attention-driven exploration churn	• Multi-provider routing (one consumer using several models in a portfolio) • Internal use-case-specific evaluation by organizational consumers (private benchmarks they trust above the public leaderboard) • Regulatory-mandated provider choice for specific deployment contexts (healthcare-grade certification, government procurement)

Table 22 (continued).

Actor	Modeled	Not modeled
Evaluator	• 22-benchmark pool with three modes: fixed sequence, randomized pool, dynamic LLM-driven • Three benchmark types (public, partial with $h=0.3$, private with $h=1.0$) and a $K=3$ reporting lag • Saturation detection from score deltas; hand-authored holdout weights per benchmark • Optional best-of-N + early-access toggles (evaluator-capture)	• Coalitions or standards bodies (HELM-like consortia, MLCommons) • Multiple competing evaluators with different weightings on the same measurement axis • Benchmark retraction or integrity audits after release • Data-vendor / data-labeling supply chains feeding the benchmark • Conflict-of-interest disclosure requirements and evaluator reputation dynamics (track record affecting which scores are trusted)
Regulator	• Single regulator with a graduated lever ladder (voluntary commitment $\rightarrow$ advisory $\rightarrow$ disclosure $\rightarrow$ audit $\rightarrow$ sanction) • Three parameter sets (US light-touch, EU precautionary, balanced); every reported run uses balanced • Per-lever cooldowns in heuristic mode, a three-round gap on the LLM path; incident-rate and concentration triggers • Mandatory safety floor under audit, sanction or emergency investigation	• Industry-specific gatekeeping regulators (FDA for healthcare AI, FAA for aviation AI) • Multiple jurisdictions with mobility or arbitrage between them • Whistleblower channels surfacing hidden conduct • Safe-harbor provisions and self-certification pathways

Table 22 (continued).

Actor	Modeled	Not modeled
Funders	• Four types (VC, corporate, government, foundation) with type-specific cooldowns, informed by the cadence of AI funding events between 2020 and 2026 • \$193B capital pool with a geometric 7% per month availability curve • Per-funder identity blocks and peer-funder visibility (LLM mode) • Allocation against leaderboard, market shares, regulator interventions, recent funding history	• Compute-credit financing (cloud-provider equity-for-credits arrangements, e.g., Microsoft–OpenAI, AWS–Anthropic) • Strategic vs. purely financial corporate-investor distinction • Exit events (acquisition, IPO) and their behavioral feedback into remaining providers
Media	• Single MediaActor (Tech-Press) with multi-round narrative state inertia (OPTIMISM / SKEPTICISM / CRISIS) • Per-provider attention vector and sentiment; headline budget per round • Weighted coverage of leaderboard movement, incidents, regulatory actions, provider announcements • Rule-based (not LLM-driven)	• Differentiated media types (trade press vs. mainstream vs. research press) • Social-media amplification (TikTok, Twitter/X, Reddit) and the influencer economy • Engagement-driven coverage (algorithmic feedback between virality and what gets covered next)

Table 22 (continued).

Actor	Modeled	Not modeled
Incident generation	• Probabilistic generation (base rate 20%/round, modulated by safety allocation) • Four-level severity (minor / moderate / major / critical) with exponential decay, so an incident’s weight halves in about two rounds • Six categories (healthcare, security, bias, safety, misinformation, misuse) with a sector-matched 2× consumer penalty • Per-provider attribution; incident probability scales with market share and market size, and the penalty an incident imposes on satisfaction scales as $(1 + \text{market_share}^2)$	• Specific failure-mode taxonomy separable from severity (prompt injection vs. hallucination vs. bias) • Whistleblower disclosures and academic exposés generating reputational incidents without direct user harm • Regulatory discovery of hidden failures during audit • Socioeconomic and environmental harms (labor displacement, compute-footprint externalities) and human-computer interaction harms (over-reliance, automation bias), discussed against the AI Risk Repository taxonomy in Appendix C.6

G.3 Methodological bounds of GABMs

The outputs are qualitative. The simulation outputs candidate dynamics, which are directional patterns, an ordering of effects, and the conditions under which a mechanism fires. It does not output point predictions. Specific magnitudes, such as gap sizes, incident counts and market shares, are properties of the parameterization, and we do not present them as claims about real markets. A finding such as “privacy moves scoring credit toward or away from the dimensions where provider R&D has concentrated capability” is the kind of output this work targets. A finding of the form “privacy reduces the gap by $X\%$ ” is outside what it can support.

The validation surface is multi-dimensional. Validation of GABMs is fundamentally harder than validation of classical simulations because the validation surface multiplies across four partially-independent dimensions:

- **Parameter space.** The parameters differ in how much empirical grounding each one has, and Appendix C rates every assumption as calibrated, anchored or stipulated. The core drivers are anchored to external data: the private-benchmark reporting lag $K=3$ to Epoch AI benchmark cadence, the target cosine of 0.95 to retro-holdout score inflation [Haimes et al., 2024], the target cosine of 0.85 to within-family correlations we compute from Epoch AI data [Epoch AI, 2024], the 2023 capability vectors to Maslej et al. [2024], the incident base rate to the AI Incident Database, and the consumer need-weight vector to sector-level evidence on AI use and to occupation-level adoption data [Bick et al., 2024]. Peripheral parameters, such as decay rates, threshold magnitudes and scale factors, are modeling choices, some of them chosen for numerical stability. A sensitivity analysis over the full joint parameter space is computationally intractable, so we rely on cross-condition robustness and paired-seed matching in place of exhaustive sweeps.
- **Prompt space.** Every LLM actor decision rests on a prompt that is one of many possible framings of the same information. The PIMMUR audit (Appendix E.4) rules out the demand characteristics and coaching patterns that it covers, and it does not establish that the observed dynamics are invariant to the prompt. We do not know whether other demand characteristics remain. The neutral framing also bounds

the holdout result. Provider prompts state how each benchmark type is scored and never present holdouts as something to exploit, so the result describes providers that were not prompted to target holdout weights. A provider instructed to infer and chase holdout weights could change which benchmarks widen.

- **Architecture space.** Our choices about which mechanisms to implement and which to abstract away decide specific findings. The capability vector has six dimensions, benchmark orientation is a scalar that blends benchmark-driven and consumer-driven R&D, and safety has its own portfolio lever separate from R&D. The weight-asymmetry mechanism of §5.1 rests on that asymmetric architecture, and a simulation built on a different capability representation could change which benchmarks widen.
- **Model space.** LLM-driven agents may pattern-match training data instead of deliberating over a new situation. The literature documents convergence toward an average persona and underrepresented variance in LLM-agent simulations [Taillandier et al., 2025, Wu et al., 2026]. The behavioral layer in this paper runs Claude Sonnet 4.6 as its primary planner and adds paired robustness checks at matched seeds under Claude Opus 4.6 within the same vendor and under OpenAI GPT-5.5 across vendors (Appendix F.7). The privacy-ladder ordering replicates across all three at the outcome level, while GPT-5.5 invokes the *K*-lag mechanism in its reasoning at roughly half the Anthropic rate. The cross-vendor check weakens the concern that a convergent strategy inside the Claude family reflects model-specific bias, without removing it. Validation across other LLM families, such as Gemini, Llama and Qwen, is the most tractable extension of this work.

External validity rests on face validity and parameter grounding, and no decisive test is available.

The gold-standard test of an ecosystem-level simulation’s external validity would be out-of-sample prediction against real-world market data. That test is not available for this class of question, because the quantities that matter, which are true capability vectors, true consumer satisfaction and a provider’s true strategic rationale, are unobservable in the real world, and the observable proxies, which are market share, benchmark scores and incident counts, are downstream outcomes and not generative inputs. An exogenous-event probe, which injects a capability-release shock at round 25 (Appendix E.6), checks that an injected narrative reaches actor reasoning, and the parameter grounding of Appendix C.2 ties a small set of parameters to real-world referents. Neither substitutes for a decisive external-validity test.

An agent-based model has no natural stopping rule. Every mechanism we add invites a question about the next one. The ecosystem frame bounds that by treating an addition as a new lever on an existing actor or a new channel between actors (§5.4, Appendix H), so the architecture does not have to be rebuilt each time. Where the existing architecture cannot carry a proposed extension, we record the limitation instead of retrofitting the model.

H Extended Case Studies

The main text (§5.4) lists four policy-oriented case studies that the ecosystem frame supports. This appendix presents one worked study, on evaluator capture (Appendix H.1), and sketches two further slots, on media shadow (Appendix H.2) and on benchmark sponsorship (Appendix H.3). Each sketch adds a lever to an actor the simulation already has, the media actor or the evaluator, and neither needs a new actor or a new information flow. We have run neither of the two sketched studies.

H.1 Evaluator capture: cross-mode results

What we model. Evaluator capture refers to the structural risk that an evaluator’s revenue dependence on the providers it evaluates erodes the cross-cutting informational role of the evaluation layer. Real-world anchors include company-run leaderboards such as Scale AI’s SEAL leaderboards (2024) [Scale AI, 2024] and Meta’s private testing of 27 model variants on Chatbot Arena before the Llama 4 release, documented by Singh et al. [2025], whose simulations show that testing 20 variants raises the expected best score by about 50 points. Comparable structural-capture dynamics are documented in adjacent measurement-and-evaluation industries: credit-rating agencies under the pre-2008 issuer-pays model, and the Big-4 audit firms in financial

reporting. In each case the integrity of a measurement layer that multiple actors depend on is conditional on the layer remaining outside the production flow it measures.

Mechanism. When the lever activates (`evaluator_as_company=True`), four structural changes engage:

1. **Best-of- N trial submissions.** Each provider chooses how many submissions to make per round. The condition caps the count at 12, and the LLM planner chooses a number between 1 and 10, so the runs reported below use a cap of 10 (Appendix B.6). Each extra submission costs `fee_per_submission` (0.05 sim units) from the R&D budget. The selection bias from taking the maximum of N trials is $\mathbb{E}[\max(X_1, \dots, X_N)] \approx \mu + \sigma\sqrt{2\ln N}$.
2. **Early access.** Paying providers get blended belief initialization on new benchmarks: their existing noisy prior is moved toward the benchmark’s holdout scoring weights, $(1 - 0.7) \times \text{prior} + 0.7 \times \text{holdout weights}$, falling back to the public weights on public-type benchmarks, which carry no holdout vector.
3. **Discretionary budget.** Each provider gets a parent-company allowance that it can spend on submissions (Genesis and Mirage 2.0; Orion 0.5; Apex 0.3; Spark and OpenCore 0.0).
4. **Evaluator revenue.** The evaluator becomes a company with its own budget, and it collects both the submission fees and allocations from the funders.

Calibration. The submission cap of 12 sits below the 27 variants that Meta tested. The early-access blend factor of 0.7 is a modeling choice, set high enough for the mechanism to show above seed-to-seed noise at $N=30$ in heuristic mode. The per-provider discretionary budgets are modeling choices that follow differences in parent-company resourcing, with Genesis and Mirage backed by a large platform company, Orion partnered with one, Apex funded by venture capital, and Spark and OpenCore standalone. The case study runs at `evaluation_lag=0` instead of the canonical `evaluation_lag=3`, which exposes the extreme of the mechanism with no reporting lag and so gives the largest channel through which best-of- N submissions turn into score inflation.

Cross-mode comparison. We compare matched baseline and `evaluator_capture` runs (40 rounds) in LLM mode at two paired seeds (42, 43) against the heuristic sweep at $N=30$. Paired deltas (`evaluator_capture` minus `baseline`) are summarized in Table 23. The LLM result reverses the sign of the heuristic result on the headline structural outcome: heuristic $\Delta\text{HHI} = +0.041$ (directional, $p \approx 0.20$; concentration *rises* under the business-model condition), LLM $\Delta\text{HHI} = -0.029$ at $N=2$ (concentration *falls*). At these sample sizes the difference in sign is suggestive; neither effect is individually established.

Table 23: Paired deltas (`evaluator_capture` minus `baseline`) in LLM mode at seeds 42 and 43, $N=2$. *Apex share* and *Apex safety* are provider-specific; the remaining columns are ecosystem-level. Δgap uses the per-benchmark gap of \$5, which is the published score minus matched consumer satisfaction, averaged over the 13 benchmarks and the six providers in the final round.

Seed	ΔHHI	$\Delta\text{leader share}$	$\Delta\text{mean safety}$	$\Delta\text{Apex share}$	$\Delta\text{Apex safety}$	Δgap
42	-0.019	-0.032	-0.009	+0.069	-0.022	+0.011
43	-0.039	-0.041	-0.012	+0.093	-0.022	+0.008
Mean	-0.029	-0.037	-0.011	+0.081	-0.022	+0.010

The LLM runs also do not match the heuristic-layer hypothesis that `evaluator_capture` redistributes share toward safety-investing providers. Apex does gain share (mean $\Delta = +0.081$), while its own safety allocation *falls* by 0.022 on average and ecosystem-wide mean safety drops slightly (-0.011). The reallocation traces instead to access to throughput, because the submission cap and the early-access blend of 0.7 toward the true weights benefit whichever provider is best placed to fill the extra leaderboard slots, whatever its safety posture. The paired Δgap carries the same sign on both seeds (+0.011 and +0.008, mean +0.010), so in these two runs the `evaluator_capture` structure widens the distance between the published score and the

satisfaction consumers receive, which is the direction the best-of- N mechanism predicts. Two seeds give a direction and not a magnitude. Read together, the two LLM runs de-concentrate the market slightly, move share toward Apex through a throughput channel, and widen the score-satisfaction gap, and we record the sign reversal on concentration against the $N=30$ heuristic layer as a candidate mode dependence.

H.2 Media shadow

The media’s coverage choices are a governance question distinct from what the evaluator measures. The **MediaActor** already weights leaderboard changes, safety incidents, regulatory escalations and provider announcements when it selects each round’s headlines, and those weights are constants in the code. Shifting the attention budget toward incidents and away from leaderboard movement, or the other way round, would test whether provider investment follows whatever the media covers, independently of what the evaluator publishes.

H.3 Benchmark sponsorship

A different information asymmetry sits upstream of privacy. When a provider co-funds a benchmark, it may learn that benchmark’s weight structure before its competitors do, and the FrontierMath and OpenAI disclosure is the real-world instance. Studying it would need a per-benchmark sponsor tag and a pre-access channel that delivers weight information to sponsors at $t - k$ and to other providers at t , and the simulation implements neither today. The per-submission fee of the evaluator-capture case (Appendix H.1) is open to every provider on the same terms, while sponsorship is selective, because some providers know ahead of time and others do not. Such a study would ask whether co-funding works mainly as a way to extract rent, where the sponsor pays and the sponsor benefits, or as a way to signal commitment, where the sponsor pays and other providers read the signal off its visible behavior.