Public AI Evaluation Snapshots as Archives: Observability, Decisions, and Audit Gates

Yanan Long [1]

Abstract

Public AI evaluations are often read as final leaderboards, yet the public record is a selective, dated sequence shaped by reporting rules, benchmark revisions, and missingness. We reconstruct repeated public archives for LiveBench and Open LLM Leaderboard v2, use LMArena as a preference stress test, and include GAIA and tau-bench as limited agentic pilots. A constructed final top-10 record under a fixed pool of 1,000 candidate systems can match incompatible histories: the same terminal tail model implies times of 23.03 or 75.13 to reach within 0.05 of the ceiling, depending on the path. Repeated snapshots can also change the action favored by stylized Bayesian loss functions. Fixed audit gates reject stronger claims for the candidate model: it fails synthetic recovery, real objective-archive prediction, preference transfer, and uncertainty calibration. The result is an auditable protocol for reconstructing public evaluation histories, identifying when final records are insufficient, and falsifying unsupported frontier claims.

1. Introduction

Public evaluation reporting for modern AI systems is selective and time-indexed. A leaderboard, benchmark release, arena slice, or agentic task report usually exposes a ranked or top-k subset while hiding attempted, unpublished, revised, scaffolded, or under-threshold systems. Agentic reports may also aggregate over prompts, tools, planners, judges, and environment settings without exposing the execution trace. A terminal ranking is therefore a lossy endpoint of a repeated public history.

This paper studies that history as an archive object. Repeated snapshots retain report timing, benchmark version, visibility rule, and the metadata needed to tell whether an apparent change is evidential or merely a reporting artifact. We focus on public AI evaluation histories reconstructable from source-native releases, including objective benchmark records, preference leaderboards, and aggregate agentic-result reports. The contribution is an archive contract, an observability boundary, and an adjudication protocol for repeated public evaluation records under selection and benchmark drift.

The boundary question sits at the intersection of selective inference and winner's-curse correction (Zrnic and Fithian, 2025), operational frontier estimation (Liu et al., 2022; Einmahl and He, 2025), and plateau identification in cure-style settings (Jackson et al., 2026; Yuen and Musta, 2026). What distinguishes the present paper is its evidential framing: repeated selected records are the primary object, archive construction is part of the scientific claim, and stronger model statements must pass fixed audit gates rather than borrowing credibility from a terminal rank.

We therefore separate archive-level evidence from model endorsement. Repeated public histories can be reconstructed, graded, compared across observation regimes, connected to Bayesian decision diagnostics, and used to adjudicate candidate inference methods. The selection-aware frontier model evaluated here is a stress-test object; its failures show that the protocol can reject attractive but unsupported frontier-inference claims rather than laundering them into leaderboard narratives.

The evidence spans two primary objective archives, one preference stress test, and two agentic applicability pilots. GAIA (Mialon et al., 2024) enters as a secondary agentic pilot and tau-bench (Yao et al., 2025) as an agentic stress-test pilot; both are aggregate public-result histories, not full agent-trace observability evidence. We first normalize repeated snapshots into a graded archive, then state a constructive observability boundary between repeated traces and terminal records, use Bayesian posterior expected-loss diagnostics to test action sensitivity, and apply fixed audit gates to a candidate selection-aware frontier model. The result is not a model-victory claim: the current candidate does not earn recovery, predictive, transfer, or calibration claims.

2. Related Work

AI evaluation has repeatedly exposed the limits of terminal scores and static leaderboards. Prior work argues that final

[1]StickFlux Labs. Correspondence to: Yanan Long <ylong@uchicago.edu>.

Preprint. June 6, 2026.

test numbers should be accompanied by search and validation traces (Dodge et al., 2019), that leaderboard rank can diverge from user utility (Ethayarajh and Jurafsky, 2020), and that small benchmark differences are often statistically fragile (Card et al., 2020). Efforts to repair benchmarks push in the same direction by emphasizing validity, headroom, dynamic data collection, diagnostic leaderboards, holistic multi-metric reporting, contamination checks, sensitivity audits, and explicit saturation analysis (Bowman and Dahl, 2021; Kiela et al., 2021; Liu et al., 2021; Liang et al., 2023; Sainz et al., 2023; Alzahrani et al., 2024; Akhtar et al., 2026). Continuously updated benchmarks such as LiveBench (White et al., 2025) and Open LLM Leaderboard v2 (Open LLM Leaderboard, 2024) address some contamination and maintenance issues, but they still require source-native histories before they can support temporal archive claims.

The statistical concern is also related to adaptive data analysis, reliable leaderboard mechanisms, post-selection inference, and winner's-curse correction (Dwork et al., 2015; Blum and Hardt, 2015; Roelofs et al., 2019; Berk et al., 2013; Andrews et al., 2024; Zrnic and Fithian, 2025). Those literatures explain why selected frontier records should not be interpreted as ordinary fixed estimates. Our setting differs because the organizer usually does not expose a controlled feedback mechanism or complete submission pool. The archive problem is therefore retrospective: reconstruct source-native public traces, mark what is missing, and test candidate inference methods against future observations rather than treating the observed winner as an unbiased target.

Several baseline families are natural comparators for such an archive. Item-response and latent-ability models for evaluation leaderboards treat examples as unequally informative and separate item difficulty from system skill (Lalor et al., 2016; Rodriguez et al., 2021; Vania et al., 2021). Scaling-law and capability-extrapolation work provides a different frontier baseline, while also warning that aggregate metrics can induce or hide apparent discontinuities (Kaplan et al., 2020; Hoffmann et al., 2022; Srivastava et al., 2023; Schaeffer et al., 2023). Preference leaderboards inherit Bradley-Terry and Elo-style assumptions (Bradley and Terry, 1952; Elo, 1978), and modern Arena-style evaluations add prompt, population, judge, and release-timing confounds (Zheng et al., 2023; Chiang et al., 2024). This is why our protocol treats latent-effect/factor, terminal-history, rolling-frontier, and Bradley-Terry / Elo methods as comparators instead of assuming a selection-aware model should win.

Agentic benchmarks extend evaluation from base models to system configurations involving tools, browsing, code execution, environments, simulators, and judges. GAIA and tau-bench are central examples for general-assistant and tool-agent-user interaction evaluation (Mialon et al., 2024; Yao et al., 2025); SWE-bench, WebArena, AgentBench, and ToolLLM illustrate the broader move toward execution-grounded and tool-mediated tasks (Jimenez et al., 2024; Zhou et al., 2024; Liu et al., 2024; Qin et al., 2024). These sources motivate the archive contract's metadata requirements. They do not turn the current GAIA or tau-bench pilots into main evidence for the candidate frontier model.

3. Evaluation-Trace Archive

The archive layer is a first-class contribution because no real-data evidence is allowed to count until deterministic source normalization is complete. Each source record stores the public source, snapshot unit, timestamp field, score fields, score orientation, rank handling, duplicate policy, missingness summary, and inclusion grade. We convert scores to a single higher-is-better orientation and collapse duplicates to one canonical record per source, benchmark, timestamp, system, and task group, preferring explicit rank, then better rank, then higher score. Timestamps must be source-native or explicitly flagged as derived. Any source that fails schema, timestamp, orientation, duplicate, top-k reconstruction, or role-specific minimum-history checks is excluded from the corresponding evidence role rather than silently disappearing into downstream tables.

Table 1 provides the compact source-validation readout used in the main text. LiveBench (White et al., 2025) and Open LLM Leaderboard v2 (Open LLM Leaderboard, 2024) are the primary objective archives, and LMArena (Chiang et al., 2024) serves as a preference stress test. GAIA is included as a secondary agentic pilot with 463 snapshots, 3,353 systems, and 11,784 diagnostic rows, while tau-bench is included as an agentic stress-test pilot with 10 snapshots, 27 systems, and 27 diagnostic rows. LiveCodeBench, HELM Capabilities (Liang et al., 2023), and SWE-bench Verified (Jimenez et al., 2024) remain excluded because the public histories used here did not provide the versioned source tables needed for temporal reconstruction. Those exclusions, like the non-primary roles assigned above, are part of the evidence package rather than after-the-fact omissions. For richer agentic benchmarks, the same contract should additionally record tool budget, environment version, scaffold identity, retry policy, judge version, and human-intervention policy.

Archive validation protocol. The inclusion grades are operational rather than descriptive. `main` denotes eligibility for the primary objective rolling-origin backtest, `stress-test` denotes use only in a source-specific stress regime, `secondary` denotes an archive-applicability pilot outside the main objective evidence, and `excluded` denotes a source retained in the manifest but not counted in

Table 1. Compact source-validation readout for the public evaluation archive. Roles define how each source can be used in the manuscript evidence: primary objective prediction, preference stress testing, agentic applicability, or explicit exclusion.

Source	Public role	Validated scale	Use
LiveBench	Primary objective archive	94 snapshots; 195 systems	Primary rolling-origin objective prediction evidence.
Open LLM Leaderboard v2	Primary objective archive	262 snapshots; 4,484 systems	Primary rolling-origin objective prediction evidence.
LMArena	Preference stress test	152 snapshots; 365 systems	Separate Arena-style stress test; not an objective archive.
GAIA	Secondary agentic pilot	463 snapshots; 3,353 systems; 11,784 diagnostic rows	Archive-applicability evidence for aggregate agentic histories.
tau-bench	Agentic stress-test pilot	10 snapshots; 27 systems; 27 diagnostic rows	Archive-applicability evidence for tool-use submission histories.
LiveCodeBench; HELM Capabilities; SWE-bench Verified	Excluded	Not counted in the evidence baseline	Public histories used here lacked the versioned source tables needed for temporal reconstruction.

the evidence baseline. To qualify as `main`, an archive must have at least 5 validated snapshots, at least 10 distinct canonical systems, at least 3 eligible one-step folds, and at least 1 eligible two-step fold; `stress-test` and `secondary` sources must have at least 3 validated snapshots. Eligible fold counts are deterministic functions of the validated snapshot history, with one-step folds equal to $n_{\text{snap}} - 3$ and two-step folds equal to $n_{\text{snap}} - 4$ when positive. Together with the row-level checks above and the manifest release pointers in Appendix A, these thresholds make the archive claim inspectable rather than merely asserted.

4. Evaluation-Trace Observability Regime

For evaluation source b at reporting time t, we treat the public snapshot as a selected record: it contains a selected set S_{bt} with observed scores or preference summaries $\{y_{bti} : i \in S_{bt}\}$, a reported cutoff size k_{bt} when the source exposes one, and auxiliary archive metadata a_{bt}. For simple model leaderboards, i indexes a model submission; for richer agentic evaluations, it may index a system configuration that includes the base model, prompt, tools, memory, planner, environment policy, and judge. The repeated-snapshot archive is therefore

$$D_b^{\text{snap}} = \{(t, S_{bt}, y_{bti} : i \in S_{bt}, k_{bt}, a_{bt})\}_{t \in \mathcal{T}_b}. \quad (1)$$

The metadata a_{bt} is part of the evidence, not bookkeeping: it records benchmark version, source timestamp, score orientation, rank handling, duplicate policy, task slice, and any available evaluator or environment information.

A terminal-only archive keeps only the final selected public record,

$$D_b^{\text{term}} = (t_T, S_{bT}, y_{bTi} : i \in S_{bT}, k_{bT}, a_{bT}). \quad (2)$$

This makes the missingness contrast explicit. The repeated archive retains a time-indexed sequence of selected measurements together with their public reporting context, whereas the terminal archive compresses that sequence to a single selected cross-section. As a result, terminal records can support terminal rank or terminal-score summaries, but they discard the temporal evidence needed for claims about how headroom changed before the final report.

For frontier-style questions, the object of interest is not the realized winner in the observed candidate mix but a path functional defined under a fixed reference convention. After fixing the score orientation, the reference scale, and the population or pool size against which the frontier is evaluated, let $F_{bt}^{\star}$ denote the source-normalized score law at time t under that convention. One useful standardized frontier is

$$\varphi_b(t) = (F_{bt}^{\star})^{-1}(1 - 1/m^{\star}), \quad (3)$$

with fixed reference pool size $m^{\star}$. In the reported boundary construction, $m^{\star} = m = 1000$. The corresponding headroom path is

$$\delta_b(t) = u_b - \varphi_b(t), \quad (4)$$

and the timing functional is

$$T_b(\epsilon) = \inf\{t : \delta_b(t) \le \epsilon\}, \quad (5)$$

where u_b is the source-normalized score ceiling and ϵ is a fixed residual-gap tolerance on that same scale. The reported construction uses a unit upper bound and $\epsilon = 0.05$. This is why timing is not an independently observed event: it inherits the information and limitations of whatever trace is used to infer the headroom path. Other evaluation-trace targets include future-observation prediction, benchmark saturation, rank stability, preference drift, judge sensitivity, and action stability under candidate deployment decisions.

Identification boundary. Repeated selected records can make frontier timing an inferential target only when source

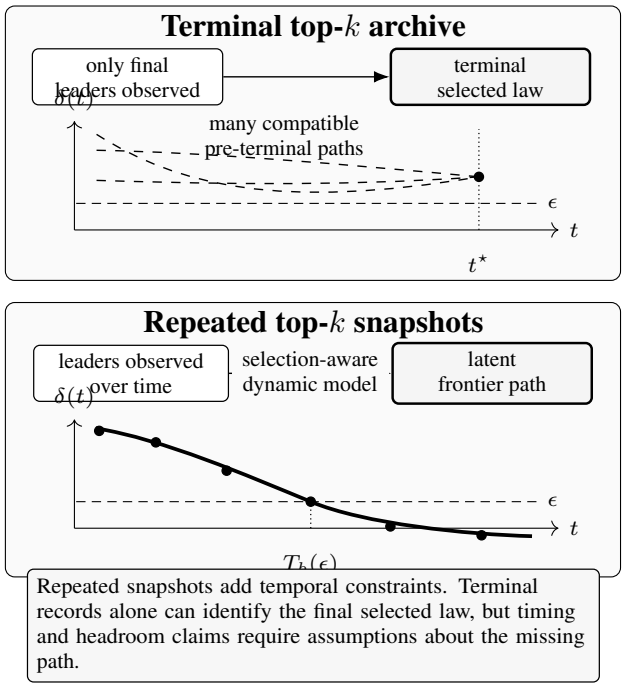

Figure 1. Repeated top-k snapshots constrain a common latent path. Terminal-only archives constrain only the selected terminal law and can leave pre-terminal timing unidentified.

metadata, reporting rules, system identities, score orientation, and an explicit path convention connect observations across distinct reporting times; terminal-only records constrain the selected terminal law but do not, by themselves, identify the pre-terminal headroom path.

Figure 1 shows the qualitative split, and the concrete boundary construction is sharper. The verified terminal-only counterexample uses one archive at $t_T = 10$ with shared $(m, k) = (1000, 10)$ and a shared generalized-Pareto terminal law on the normalized score scale with $(\mu, \sigma, \xi) = (0.8160602794, 0.18, -0.12)$. Path A uses $(\delta_\infty, \delta_0, \nu, \lambda) = (0, 0.5, 1, 0.1)$ and path B uses $(0, 0.2246644821, 1, 0.02)$. Both induce the same terminal gap 0.1839397206 and therefore the same terminal selected likelihood; the verification artifact records maximum absolute log-density difference 0.0. Yet the implied timing targets differ materially: $T_b(\epsilon)$ is 23.03 for path A and 75.13 for path B. This makes the observability claim constructive rather than merely schematic: repeated snapshots define a different information regime, so terminal leaderboards are not a sufficient evidential object for timing and headroom claims.

4.1. Bayesian decision readout

The same observability boundary also has a decision-theoretic reading. The primitive here is an action under posterior loss, not rejection of a null hypothesis. Let θ_b denote the latent evaluation state for source b, including the

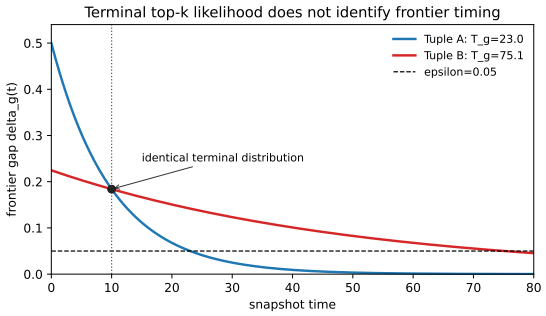

Figure 2. Boundary construction used for the observability claim. The terminal selected likelihood is identical at the observed time, but the compatible frontier-gap paths imply $T_b(\epsilon)$ values of 23.03 and 75.13.

frontier path, pool and reporting mechanism, benchmark adequacy, and any decision-relevant risk variables. A decision maker observes an archive D and chooses an action

$$a \in \mathcal{A} = \{\text{continue}, \text{refresh}, \text{harden}, \text{restrict}, \text{collect}\}. \tag{6}$$

Given a loss $L(a, \theta_b)$, Bayesian posterior decision analysis evaluates actions by posterior risk,

$$\rho(a \mid D) = \mathbb{E}[L(a, \theta_b) \mid D], \quad a^\star(D) \in \arg\min_{a \in \mathcal{A}} \rho(a \mid D), \tag{7}$$

rather than treating a rank, interval, or hypothesis label as the final object (Berger, 1985). In this framing, hypotheses such as open headroom, near plateau, stale evaluation, or unacceptable residual risk are regions of latent state space. A posterior-threshold rule is only a special case: if acting on a region H has false-positive cost c_{10} and not acting has false-negative cost c_{01}, then acting is Bayes-optimal exactly when

$$\Pr(H \mid D) > \frac{c_{10}}{c_{10} + c_{01}}. \tag{8}$$

So the threshold is induced by asymmetric loss, not by a universal significance level. This translation into hypothesis regions is useful only after the action and loss have been fixed; it does not make the archive protocol a generic Bayesian substitute for null-hypothesis significance testing, and it does not validate operational governance decisions.

The implemented readout is simply the finite-draw version of that rule. Each fitted posterior provides draws of $T_b(\epsilon)$, latest headroom, near-threshold and stale-evaluation indicators, residual-risk scores, and headroom shortfall. For each loss family and weight profile, the readout computes

$$\widehat{\rho}(a \mid D) = \frac{1}{M} \sum_{m=1}^{M} L(a, \theta_b^{(m)}) \tag{9}$$

for every action and records both the selected action and all per-action risks. The loss profiles are not claimed to be elicited utilities; they are scale-controlled diagnostics. The indicator family uses binary events, the smooth-residual family uses normalized residual and stale-degree scores, and the balanced, safety-heavy, and staleness-heavy profiles vary the relative cost of missed risk, stale evaluation, over-restriction, refresh, hardening, restriction, and additional collection. That is why we report action agreement and regret across profiles rather than a single calibrated policy.

The prior is initialized in the fitted frontier model, not in the decision rule itself. The repeated-snapshot and terminal-only coupled fits use the same dynamic prior family; in the evaluated parameterization, the asymptotic gap is a fraction of ϵ, the initial excess gap is log-normal, the current-headroom fraction is beta-distributed, and the score precision is log-normal. These choices are treated as part of the candidate model under audit. The present evidence does not include a full prior-sensitivity sweep, so no claim below depends on prior-robust posterior timing or operational utility calibration.

We use this layer only as an action-facing diagnostic over synthetic posterior draws, not as an operational policy. The decision readout compares repeated-snapshot and terminal-only posteriors under three loss families and three weight profiles, which in the full matched setting yields $1{,}350$ action comparisons across 150 paired trajectories. Repeated and terminal archives usually agree on the modal action, but the conservative readout still produces loss-sensitive disagreements: in the exact matched summary, the minimum action-agreement settings are 0.82 under noisy plateau-like behavior, 0.76 under candidate-pool growth, and 0.94 under heavy selective reporting. Terminal-under-repeated regret is positive in the disagreement settings and zero in settings with no action disagreement. This supports only the boundary-level claim that the observation regime can matter for downstream Bayes actions, even when a particular fitted frontier model remains under adjudication.

5. Candidate Stress Test and Adjudication Protocol

To demonstrate the archive-and-adjudication workflow end-to-end, we evaluate a plausible selection-aware frontier architecture. The main candidate, `S0`, is a dynamic coupled fit over repeated snapshots with a selection-aware likelihood that conditions on the reporting cutoff instead of treating reported leaders as ordinary uncensored draws. Its ablations are defined before use: `S1` is a static iid fit with neither temporal coupling nor selection correction, `S2` is static but selection-aware, `S3` is dynamic but not selection-aware, `S4` is a deterministic rolling-max heuristic over observed frontier scores, and `S7` is the same coupled selection-aware architecture as `S0` conditioned only on the terminal snapshot. The purpose is not to endorse `S0` in advance. It is to expose an attractive modeling idea to a sequence of auditable claims that can be supported, localized, or rejected.

The comparator set is intentionally explicit and minimal. For objective-archive diagnostics we also include the latent-effect/factor baseline, and for Arena-style preference data we use Bradley-Terry / Elo as the native comparator. Throughout the real-data sections, validation is against future observations only rather than latent frontier truth. Appendix Table 7 records the same candidate labels together with their gate roles.

The adjudication protocol has six components: truth-known synthetic recovery and false-positive control, archive construction and validation, rolling-origin objective backtests, a separate LMArena preference stress test, timing-interval calibration, and reproducibility. Its gate rules are fixed before applying the reported gate summary and are deliberately asymmetric. In truth-known recovery, `S0` must strictly beat `S4` and `S7` on both latest-gap error and finite-$T_b(\epsilon)$ error while not losing to `S1`–`S3`; in the slow-frontier negative control, it must make zero false finite-plateau decisions; in the objective backtest, it must beat both `S7` and `S4` on predictive log score without reversing rank calibration on the primary archives; and in the calibration audit, only diagnostic-admissible `S0` rows count, with posterior pass probability at least 0.8 for the acceptable-calibration region. Appendix Table 7 and Table 3 map those claims to their audit artifacts, and Table 2 summarizes the resulting evidence.

6. Results

We report the results as separate checks rather than collapse them into a single aggregate score, because the protocol is meant to adjudicate distinct claims. Archive validity, observability, decision relevance, agentic applicability, predictive adequacy, preference transfer, and posterior calibration do different jobs and can succeed or fail separately. On the evidence available here, the archive-and-adjudication contribution is supported, while endorsement of the particular selection-aware model remains withheld.

Technical definitions. Several terms recur in this section in a narrow technical sense. Predictive log score is the average held-out predictive log likelihood,

$$\mathrm{LogScore} = \frac{1}{n}\sum_{r=1}^{n} \log p(y_r^{\mathrm{hold}} \mid D_r^{\mathrm{train}}), \qquad (10)$$

Table 2. Main archive-and-adjudication readout. Positive rows establish what repeated public snapshots add; diagnostic rows show that the protocol localizes unsupported model claims instead of hiding them.

Protocol question	Readout	Manuscript consequence
Can public histories be reconstructed as traces?	Supported	LiveBench and Open LLM Leaderboard v2 become primary objective archives; LMArena is a preference stress test; GAIA is a secondary agentic pilot; tau-bench is an agentic stress-test pilot.
Do repeated traces add information?	Supported	Terminal-only top-k evidence can match the selected likelihood while leaving plateau timing materially different, with compatible $T_b(\epsilon)$ values of 23.03 and 75.13.
Can traces change action-facing readouts?	Supported diagnostic	In synthetic posterior comparisons, repeated and terminal observation regimes produce loss-sensitive posterior-action disagreements under specified Bayesian decision losses, but not operational policy evidence.
Can the archive contract stage aggregate public agentic-result histories?	Supported pilot	GAIA and tau-bench show that the archive contract can stage aggregate public agentic histories, while also exposing missing scaffold and tool metadata.
Can the protocol reject unsupported models?	Supported diagnostic	The candidate method does not earn truth-known recovery, primary-archive prediction, preference transfer, or calibrated timing uncertainty.

Table 3. Claim-to-artifact reproducibility map for the headline manuscript evidence.

Claim	Main artifact	Driver / readout	Audit use
Truth-known recovery	Synthetic gate summary and machine-readable companion.	Frontier metric summary and gate readout.	Synthetic recovery regimes, slow-frontier negative control, and $0/3$ gate readout.
Archive validation	Archive validation table.	Archive build summary.	Source grades, fold counts, duplicate policy, and exclusion reasons.
Observability boundary	Terminal-boundary verification and Figure 2.	Boundary evidence index.	Shared terminal likelihood and differing $T_b(\epsilon)$ values.
Objective backtest	Objective headline table and paired bootstrap intervals.	Rolling-origin manifest.	Primary-archive predictive diagnostics and comparator rule.
Preference stress test	Arena headline table and paired bootstrap intervals.	Arena stress summary.	LMArena stress-test diagnostics against `S7` and `BT/ELO`.
Decision diagnostic	Exact observation-regime summary.	Exact decision-readout driver.	Repeated-vs-terminal action agreement and regret summaries.
Calibration audit	SBC report and calibration failure table.	Coverage-by-regime and diagnostic-pathology summaries.	Acceptable-calibration posterior probabilities and interval pathologies.

so higher is better. Continuous ranked probability score (CRPS) is the proper score

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbf{1}\{y \le z\}\right)^2 \, dz, \qquad (11)$$

so lower is better. For the Normal predictive distributions used in the backtest metric, if $F = \mathcal{N}(\mu, \sigma^2)$ and $u = (y - \mu)/\sigma$, then

$$\mathrm{CRPS}(F, y) = \sigma\Big[u\big(2\Phi(u) - 1\big) + 2\phi(u) - \frac{1}{\sqrt{\pi}}\Big]. \qquad (12)$$

Item response theory (IRT) motivates one common latent-ability form

$$\Pr(Y_{iq} = 1) = \sigma(a_q(\theta_i - \beta_q)), \quad \sigma(x) = \frac{1}{1 + e^{-x}}, \qquad (13)$$

where system ability θ_i, item difficulty β_q, and discrimination a_q parameterize the response. In this paper, the objective-archive comparator is a shrinkage latent-effect/factor baseline that separates system/model, family, task, and time effects, for example through a linear predictor of the form

$$\eta_{mfqt} = \mu + \alpha_m + \gamma_f + \delta_q + \tau_t, \qquad (14)$$

with partial pooling on those components. We use it as a diagnostic comparator for objective archives, not as an endorsed frontier model. The terminal-history baseline conditions only on the final coupled evidence, while the rolling-frontier heuristic extrapolates a simple time-indexed frontier without the candidate model's selection-aware likelihood.

For the preference stress test, LMArena denotes leaderboard rating snapshots derived from Arena-style preference evaluations rather than an objective archive. Bradley-Terry uses

$$\Pr(i \succ j) = \sigma(\eta_i - \eta_j), \qquad (15)$$

and Elo is a rating and update variant built on the same latent-strength idea, so Bradley-Terry / Elo serves here as the native comparator for preference-style ratings rather

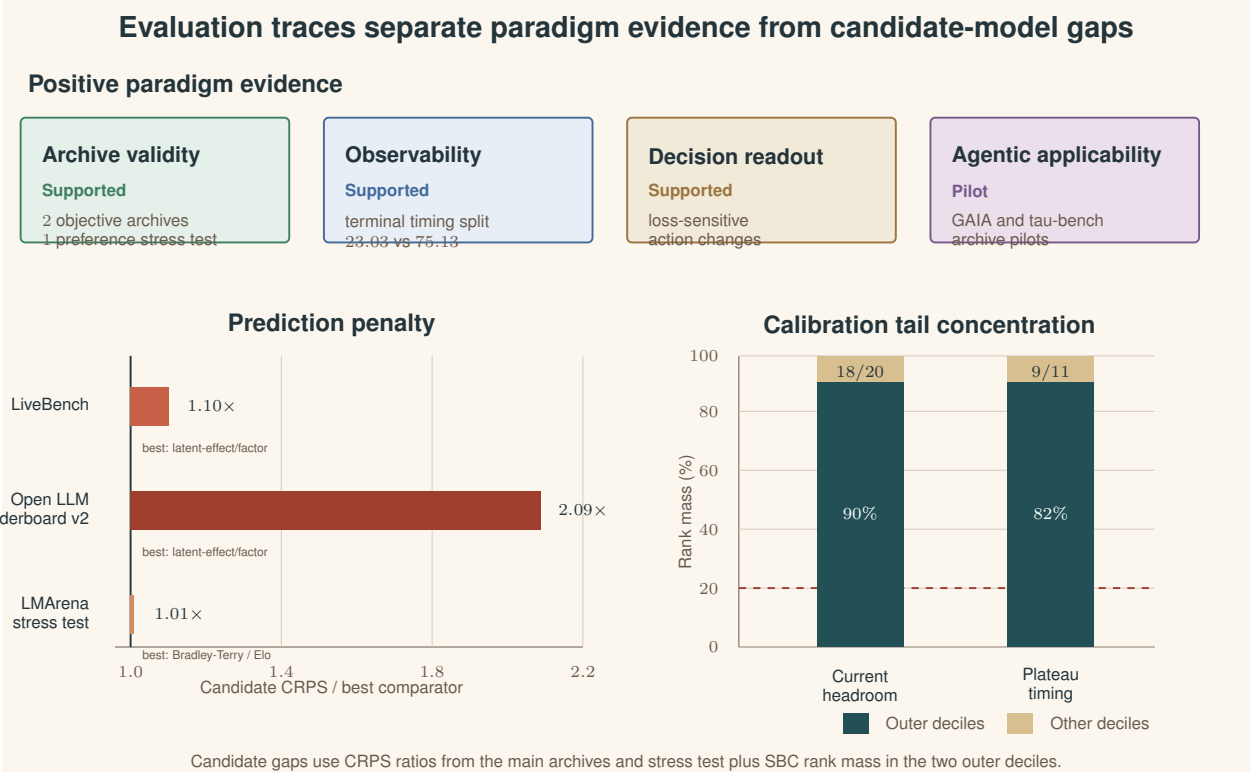

Figure 3. What the trace-adjudication protocol separates beyond terminal tables. The positive evidence is archive validity, observability, decision relevance, and agentic applicability; the candidate-model gaps are predictive shortfalls and posterior calibration failure.

than as an endorsed frontier model. Top-k recall is

$$R_k(y, \hat{y}) = \frac{|S_k(y) \cap S_k(\hat{y})|}{k}, \tag{16}$$

where $S_k(\cdot)$ is the reported top-k set. Rank calibration is the Spearman correlation between observed and predicted descending ranks,

$$\mathrm{RankCal}(y, \hat{y}) = \rho_S\big(r^{\downarrow}(y),\, r^{\downarrow}(\hat{y})\big)\,. \tag{17}$$

Paired cluster-bootstrap intervals report primary-minus-comparator metric differences under resampling over fold clusters.

Simulation-based calibration (SBC) uses the posterior rank or midrank of the data-generating truth among posterior draws,

$$R = \sum_{m=1}^{M} \mathbf{1}\{\theta^{(m)} < \theta^{\star}\} + \frac{1}{2}\sum_{m=1}^{M} \mathbf{1}\{\theta^{(m)} = \theta^{\star}\}, \tag{18}$$

which is uniform when the posterior is calibrated. If these SBC ranks are binned into B bins with counts $n_1, \ldots, n_B$, the Dirichlet rank-bin posterior is

$$p \mid n \sim \mathrm{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_B). \tag{19}$$

The Kolmogorov-Smirnov and chi-square summaries used later are Bayesian posterior discrepancy readouts over those rank-bin probabilities,

$$D_{\mathrm{KS}}(p) = \max_{1 \le b \le B} \left| \sum_{j=1}^{b} p_j - \frac{b}{B} \right|, \tag{20}$$

$$D_{\chi^2}(p) = \sum_{b=1}^{B} \frac{(p_b - 1/B)^2}{1/B}, \tag{21}$$

rather than null-hypothesis significance testing claims.

We state the candidate-model gate outcomes once here to keep the framing explicit. Truth-known recovery is $0/3$, objective-archive prediction is $0/2$, the Arena stress test shows a preference-transfer gap, and posterior calibration fails for both current headroom and finite timing. Tables 4 and 5 provide the predictive details, and Figure 3 summarizes what the trace-adjudication protocol separates beyond terminal tables.

6.1. Truth-known admissibility

The truth-known check shows why the protocol needs fixed model gates before real-archive performance can be interpreted as evidence about latent frontiers. Across controlled

settings for closing gaps, noisy plateau-like behavior, and changes in the candidate pool, the candidate method does not meet the requirement to improve on the terminal and rolling comparators while avoiding losses to reduced variants. The slow-frontier negative control behaves correctly, so the method does not hallucinate a false plateau there. Synthetic evidence is therefore not an optional appendix at the archive level; it is the first place a frontier-inference method must earn the right to make stronger real-data claims.

6.2. Primary archive prediction

The objective-archive check then tests future-observation prediction on the two validated primary archives: LiveBench and Open LLM Leaderboard v2. LMArena is deliberately kept out of this check because it is a preference stress test, and excluded sources remain visible in Appendix A rather than disappearing from the analysis.

This prediction check is a positive result for the archive protocol even though it is not a positive result for the candidate method. On LiveBench, the candidate trails both the terminal-history baseline and the latent-effect/factor diagnostic baseline on predictive log score and CRPS. On Open LLM Leaderboard v2, it is close to the terminal-history baseline on log score but still worse on both log score and CRPS, while the latent-effect/factor baseline is much stronger. The rolling-frontier heuristic is weaker than the candidate on these aggregate metrics, but the adjudication rule requires the candidate to beat the required comparators on enough primary archives, and it does not. The contribution-level result is that repeated public traces support rolling-origin prediction tests with explicit comparators rather than a terminal leaderboard impression.

Paired cluster-bootstrap intervals sharpen the same diagnosis. Against the latent-effect/factor baseline, the candidate-minus-baseline intervals are negative for log score ($[-3.00\times 10^{12}, -2.17 \times 10^{11}]$) and positive for CRPS ($[3.24, 3.94]$), so both proper-score summaries point against the candidate. Against the terminal-history baseline, the log-score interval crosses zero ($[-5.56 \times 10^{10}, 9.29 \times 10^{9}]$), but the CRPS interval remains positive ($[0.085, 0.217]$). Table 4 reports the full archive-by-method details including the rolling-frontier heuristic. The objective archives therefore validate the archive testbed while withholding predictive-superiority claims for this candidate, which is exactly the separation the protocol is meant to enforce between archive value and model endorsement.

6.3. Preference-regime stress test

Arena-style preference data defines a different measurement regime: prompt mix, population, judge, release-timing, and access confounds are not interchangeable with objective benchmark scores. LMArena is therefore used as a stress test rather than as a primary objective archive. On the main Arena comparison, the candidate trails both the terminal-history baseline and Bradley-Terry / Elo on predictive log score, CRPS, top-k recall, and rank calibration, and the resampling intervals point the same way. In this snapshot-derived stress test, the terminal-history and Bradley-Terry / Elo readouts coincide numerically, so the duplicate rows show the candidate gap against both naming conventions rather than independent evidence. Preference-transfer claims therefore need source-specific adjudication rather than a generic frontier narrative.

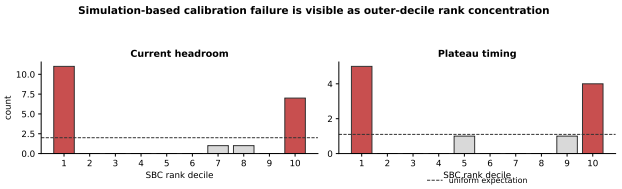

Figure 4. Simulation-based calibration ranks concentrate in the outer deciles: current headroom has $18/20$ admissible ranks with posterior acceptable-calibration probability 1×10^{-5}; finite timing has $9/11$ outer-decile ranks with probability 0.0133.

Candidate-minus-Bradley-Terry / Elo cluster-bootstrap intervals are in the wrong direction for every reported metric: log score $[-0.00793, -0.00410]$, CRPS $[0.0569, 0.1065]$, top-k recall $[-0.00842, -0.00101]$, and rank calibration $[-1.45 \times 10^{-4}, -2.04 \times 10^{-5}]$. Table 5 gives the compact metric readout. The preference stress test therefore withholds transfer claims even though the aggregate differences look numerically modest.

6.4. Posterior uncertainty audit

The calibration audit shows how the protocol handles posterior uncertainty claims. Using the available posterior draws, we compute simulation-based calibration ranks and place a Dirichlet posterior on the 10-bin rank histogram, then ask for posterior mass in an acceptable-calibration region: total-variation distance to uniform at most 0.35, outer-decile mass between 0.10 and 0.30, and overall posterior pass probability at least 0.8. Current-headroom SBC places $18/20$ admissible ranks in the outer deciles and gives posterior acceptable-calibration probability 1×10^{-5}. Finite-timing SBC places $9/11$ admissible ranks in the outer deciles and gives posterior acceptable-calibration probability 0.0133. Figure 4 shows this rank concentration directly.

The KS- and chi-square-style summaries are secondary Bayesian posterior discrepancy checks over rank-bin probabilities, not null-hypothesis significance tail probabilities. They remain mechanistically informative here because $T_b(\epsilon)$ is a thresholded path functional: weak local slope near the threshold can make small posterior shifts move the inferred crossing time sharply, collapse an interval, or push posterior

Table 4. Objective-archive predictive details. Higher log score and lower CRPS are better; candidate gaps are localized against explicit comparators.

Archive	Method	Log score	CRPS	Protocol readout
LiveBench	Latent-effect/factor baseline	-9.67×10^{9}	0.128	Best comparator
LiveBench	Terminal-history baseline	-1.06×10^{10}	0.133	Required comparator
LiveBench	Selection-aware candidate	-1.24×10^{10}	0.141	Candidate gap localized
LiveBench	Rolling-frontier heuristic	-4.24×10^{10}	0.282	Weaker heuristic
Open LLM Leaderboard v2	Latent-effect/factor baseline	-1.28×10^{12}	6.136	Best comparator
Open LLM Leaderboard v2	Terminal-history baseline	-3.95×10^{12}	12.558	Required comparator
Open LLM Leaderboard v2	Selection-aware candidate	-3.99×10^{12}	12.821	Candidate gap localized
Open LLM Leaderboard v2	Rolling-frontier heuristic	-1.19×10^{13}	22.659	Weaker heuristic

Table 5. Preference-stress-test details. Higher log score, top-k recall, and rank calibration are better; lower CRPS is better.

Method	Log score	CRPS	Top-k	Rank cal.
Bradley-Terry / Elo	−4.3190	7.4769	0.8094	0.9958
Terminal-history baseline	−4.3190	7.4769	0.8094	0.9958
Selection-aware candidate	−4.3250	7.5581	0.8047	0.9957

mass onto effectively infinite branches. That pattern is not stable enough to support calibrated timing language for this candidate method.

7. Conclusion

The end-to-end readout is coherent. Repeated public snapshots can be normalized into graded archives; terminal-only evidence has a constructive observability boundary; Bayesian posterior expected-loss diagnostics can depend on the observation regime; and the same archive contract extends to GAIA and tau-bench as aggregate public-result applicability pilots while exposing missing scaffold and tool metadata. Those are the paper's central positive results.

Those positive results do not depend on the candidate model winning. The synthetic test withholds truth-known recovery, the objective backtests withhold predictive-superiority claims on validated public archives, the Arena stress test withholds preference-transfer claims, and the calibration audit withholds strong timing-uncertainty claims. The paper's claim is therefore not that the current candidate model wins, but that public evaluation claims can be tied to a more transparent archive, observability, and adjudication standard.

Limitations

Several limitations are structural rather than accidental. Real archives provide future-observation validation, not latent frontier truth, and candidate-pool reconstruction remains partly assumption-driven. The current candidate model also uses a simple monotone gap family, without demonstrated robustness to power-law, stretched-exponential, or non-monotone frontier dynamics. The Bayesian decision layer uses stylized loss families and synthetic posterior draws, so it should be read as a diagnostic for action sensitivity rather than as a validated governance policy. LiveBench enters as a dated model-judgment aggregate, preference archives are confounded by sampling and access, and excluded archives remain outside the evidence baseline because versioned source tables were unavailable. The GAIA and tau-bench pilots show archive applicability for agentic-style histories, not full agent-trace observability. Richer agentic traces still need tool budget, scaffold identity, environment version, retry policy, judge version, and human-intervention fields.

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of ACL*, 2024.

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

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

Richard A. Berk, Lawrence D. Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of ICML*, 2015.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of NAACL*, 2021.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of EMNLP*, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas 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 *Proceedings of ICML*, 2024.

Richard A. Chechile. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods*. MIT Press, 2020.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP-IJCNLP*, 2019.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2015.

John H. J. Einmahl and Yi He. Accurate estimates of ultimate 100-meter records. *arXiv preprint arXiv:2502.04085*, 2025.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco, 1978.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of EMNLP*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, 2022.

Dan Jackson, Michael Sweeting, Robert Hettle, Binbing Yu, Neil Hawkins, Keith Abrams, and Rose Baker. Cure models: What is meant by a survival "plateau," and do experts agree on what constitutes one? *PharmacoEconomics*, 44(1):73–82, 2026.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *Proceedings of ICLR*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts,

and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL*, 2021.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of EMNLP*, 2016.

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

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of ACL-IJCNLP System Demonstrations*, 2021.

Xiaohong Liu, Tony T. Yang, and Yichong Zhang. Quasi-Bayesian inference for production frontiers. *Journal of Business & Economic Statistics*, 2022.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *Proceedings of ICLR*, 2024.

Gregoire Mialon, Clementine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants. In *Proceedings of ICLR*, 2024.

Open LLM Leaderboard. Open LLM Leaderboard v2. Hugging Face collection, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proceedings of ICLR*, 2024.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of ACL-IJCNLP*, 2021.

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

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, and 422 others. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. In *Proceedings of ACL-IJCNLP*, 2021.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *Proceedings of ICLR*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *Proceedings of ICLR*, 2025.

Tsz Pang Yuen and Eni Musta. Testing for sufficient follow-up in survival data with a cure fraction. *arXiv preprint arXiv:2403.16832*, 2026.

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

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *Proceedings of ICLR*, 2024.

Tijana Zrnic and William Fithian. A flexible defense against the winner's curse. *The Annals of Statistics*, 2025.

A. Detailed Evidence

This appendix collects the compact evidence needed to audit the main-text claims without creating a second, low-value appendix section. Table 6 expands the archive-source inventory, Table 7 defines the candidate variants and gate uses, Table 8 summarizes the gate outcomes, and Table 9 gives the agentic applicability pilots.

Table 6. Expanded source inventory for the public evaluation archive.

Source	Public role	Snapshots	Systems	Validation slices	Notes
LiveBench	Primary objective archive	94	195	91	Dated model-judgment aggregate; source-native timestamps; top-k reconstructable.
Open LLM Leaderboard v2	Primary objective archive	262	4484	259	Submission-date snapshots; flagged rows removed; rank reconstructed from score ordering.
LMArena leaderboard snapshots	Preference stress test	152	365	149	Preference archive kept separate from objective headline claims.
GAIA public results	Secondary agentic pilot	463	3353	460	Aggregate agentic-style public rows; dated level scores are usable, scaffold and tool metadata are weak.
tau-bench public submissions	Agentic stress-test pilot	10	27	7	Agentic tool-use submissions with domain, modality, retrieval/voice, and Pass@k metadata; kept outside main objective claims.
LiveCodeBench	Excluded	0	0	0	Versioned source table unavailable in the public histories used here.
HELM Capabilities	Excluded	0	0	0	Versioned source table unavailable in the public histories used here.
SWE-bench Verified	Excluded	0	0	0	Versioned source table unavailable in the public histories used here.

Table 7. Candidate variants and their gate roles.

Label	Construction	Gate role
`S0`	Selection-aware dynamic coupled fit over repeated snapshots.	Candidate baseline in truth-known recovery, objective backtests, decision diagnostics, and calibration audit.
`S1`	Static iid fit without temporal coupling or selection correction.	Reduced-variant non-loss comparator in truth-known recovery.
`S2`	Static selection-aware fit without temporal coupling.	Reduced-variant non-loss comparator in truth-known recovery.
`S3`	Dynamic fit without selection correction.	Reduced-variant non-loss comparator in truth-known recovery.
`S4`	Deterministic rolling-max heuristic over observed frontier scores.	Required strict-win comparator in truth-known recovery and objective backtests.
`S7`	Same coupled selection-aware architecture as `S0`, but conditioned only on the terminal snapshot.	Required strict-win comparator in truth-known recovery; terminal-history comparator in objective backtests and decision diagnostics.
`BT/ELO`	Bradley-Terry / Elo preference comparator applied to Arena-style leaderboard rating snapshots.	Native comparator for the LMArena stress test.

Table 8. Adjudication summary for the evaluated evidence.

Check	Status	Key readout
Truth-known synthetic recovery	not supported	The candidate method passed $0/3$ recovery regimes; only the slow-frontier negative control behaved correctly.
Archive validation	supported	Two objective archives validated as primary evidence; one preference archive validated as a stress test; GAIA validated as a secondary agentic pilot; tau-bench validated as an agentic stress-test pilot.
Objective backtest	not supported	The candidate method passed $0/2$ primary archives under the fixed comparator rule.
Preference stress test	not supported	The candidate method trails both the terminal-history baseline and Bradley-Terry / Elo on the main Arena comparison.
Observability boundary	supported	Terminal-only evidence can match the selected likelihood while yielding $T_b(\epsilon)$ values 23.03 and 75.13.
Bayesian decision readout	supported diagnostic	Synthetic posterior comparisons show loss-sensitive repeated-vs-terminal action differences, but not operational superiority.
Calibration audit	not supported	Simulation-based calibration gives low posterior probability to acceptable calibration for the candidate model; the interval audit finds missed finite-time intervals, degenerate intervals, effectively infinite timing intervals, and numerical instabilities.

Table 9. Agentic archive applicability pilots. These rows demonstrate archive applicability, not candidate-model superiority.

Source	Public role	Snapshots	Systems	Task groups	Result rows	Interpretation
GAIA public results	Secondary agentic pilot	463	3353	8	11784	Completed aggregate public-result applicability pilot; scaffold and tool metadata remain weak.
tau-bench public submissions	Agentic stress-test pilot	10	27	3	27	Completed aggregate public-result applicability pilot over voice Pass@1 slices; richer submission metadata, but not a main objective archive.