Bayesian Frontier-Evaluation Testing as Decision Under Repeated Top-k Reporting

Anonymous Authors[1]

Abstract

Frontier evaluation is a sequential decision problem under selective reporting. Public evidence often arrives as repeated top-k snapshots rather than full score distributions, so the primitive object is not rejection of a null hypothesis but a frontier-evaluation action chosen under asymmetric loss. We formulate Bayesian testing for this setting by combining a selection-aware frontier model with hypotheses defined as regions of a latent state space. The mathematical core is an observability result: repeated snapshots can constrain a common frontier trajectory, whereas terminal-only archives can agree on the selected terminal law while leaving action-relevant timing unidentified. Synthetic studies illustrate the decision layer while supporting only bounded claims. A repeated-snapshot protocol study covers 225 fitted records and obtains $219/225$ correct binary horizon decisions, while a posterior-decision analysis covers 374 fitted records and shows substantially lower frontier-path loss for the repeated-snapshot model. These results are deliberately limited: plateau-time interval coverage remains weak, and the matched terminal-only comparator ties the repeated-snapshot model on the binary horizon decision in the full matched comparison.

1. Introduction

A frontier evaluation is not only a score but an input to decisions about whether the current evaluation record remains informative, whether the suite is stale and should be refreshed, whether meaningful headroom remains, whether stronger safeguards should be activated, whether access should be restricted, or whether more evidence should be collected. These choices are difficult because public evidence is usually filtered: repeated top-k records reveal selective extremes of an evolving process rather than random samples from the full score distribution, so this paper treats the setting as Bayesian testing by decision theory (Berger, 1985). Rather than making rejection of a null the primitive object, we choose a frontier-evaluation action under asymmetric losses and selective observation, with hypotheses entering as regions of latent frontier state space, such as open headroom, near plateau, stale evaluation, or threshold-relevant unsafe capability, and public top-k records inducing a posterior over those regions from which posterior expected loss selects actions.

The observation regime matters because repeated snapshots preserve temporal information that terminal-only archives discard. If time-to-near-plateau is a derived functional of the latent frontier-gap path, then two histories can agree on the terminal selected record while implying different posterior uncertainty over current headroom, different beliefs about whether the evaluation suite is stale, and different Bayes-optimal actions. The setup is adjacent to winner's-curse correction under selective reporting (Zrnic & Fithian, 2025), operational frontier estimation (Liu et al., 2022; Einmahl & He, 2025), and plateau-identification questions in cure-style settings (Jackson et al., 2025; Yuen & Musta, 2026), but the target here is frontier-evaluation action under repeated selective reporting rather than any one of those ingredients in isolation. The manuscript therefore makes three conservative contributions: it gives the frontier-functional exposition needed for the decision problem, defines Bayesian frontier-evaluation testing under selective reporting by combining region-valued hypotheses with a frontier-evaluation loss, and presents synthetic evidence from a repeated-snapshot protocol study, a posterior-decision analysis, and a matched terminal-only comparison. We do not claim that the current fitted model solves frontier-evaluation decisions, validates operational governance, or dominates terminal-only evidence on binary action quality.

2. Observation Regimes and Frontier Functionals

For stratum g at reporting time t, let m_{gt} be the latent eligible pool and let Y_{gti} be the score of candidate $i \in \{1, \ldots, m_{gt}\}$. The primary observation is a repeated sequence of selective snapshots rather than the full vector of candidate scores:

$$\mathcal{D}_g^{\text{snap}} = \{(t, y^{\text{top}}_{gt,[1:k_{gt}]}, S_{gt}, a_{gt}, x^{\text{obs}}_{gt}) : t \in \mathcal{T}_g\}. \quad (1)$$

Here $y^{\text{top}}_{gt,[1:k_{gt}]}$ is the ordered top-k_{gt} vector, S_{gt} is the selected identity set when identities are available, x^{obs}_{gt} contains reported covariates, and a_{gt} denotes auxiliary informa-

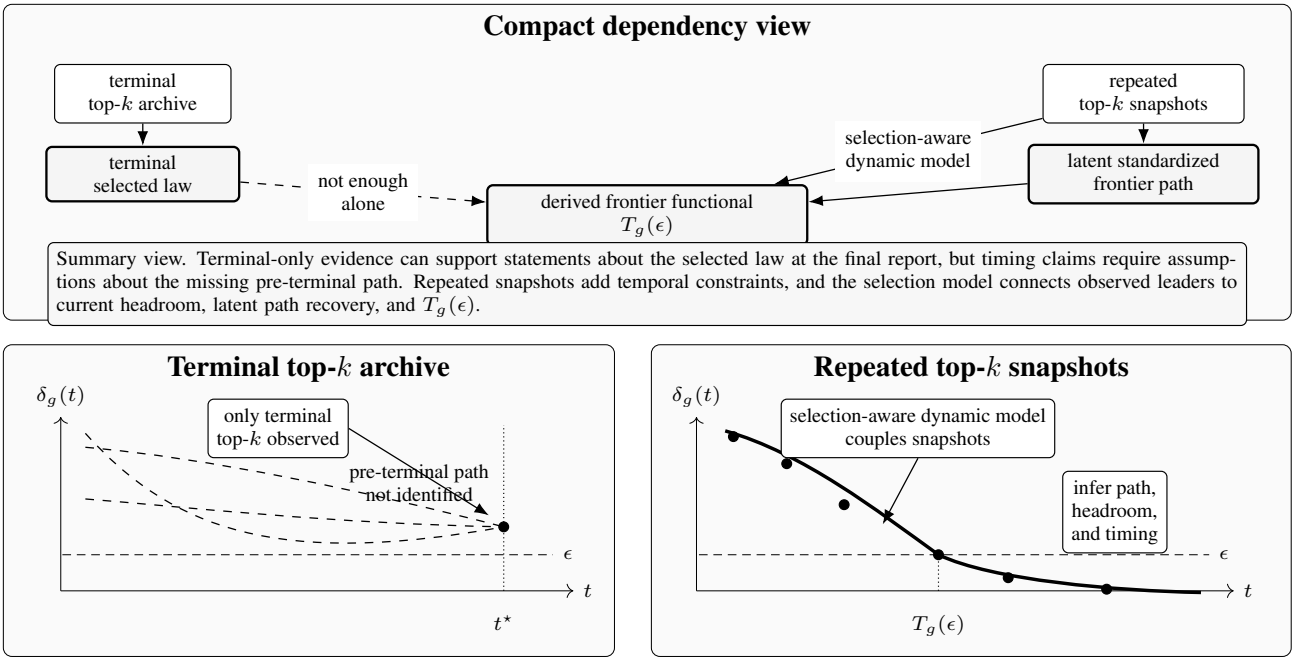

Figure 1. Observation regimes for action-facing frontier assessment. Repeated top-k snapshots retain temporal information that terminal-only archives discard, changing posterior uncertainty over frontier functionals and downstream frontier-evaluation actions.

tion about the candidate pool or reporting rule. This notation makes the missingness structure explicit: the archive reports the leaders and some context about the pool, while the scores of the unreported candidates remain latent. A weaker terminal-only archive keeps only the final selected record,

$$\mathcal{D}_g^{\text{term}} = \{(y^{\text{top}}_{g,[1:k_g]}, \tau_{g,[1:k_g]}, S_g, a_g, x_g^{\text{obs}})\}. \tag{2}$$

The distinction is substantive rather than cosmetic. Repeated snapshots expose a sequence of selective measurements generated under a common latent path, so they can carry information about how the frontier moved between reporting times. Terminal-only archives collapse that path to one selected cross-section and therefore remove precisely the temporal structure needed for plateau-timing claims.

The frontier is defined against a fixed reference population rather than against the candidate mix that happened to appear at time t. Let $Q_g^\star$ be a reference distribution over candidate descriptors and define the standardized score law

$$F^\star_{gt}(y) = \int P(Y_{gti} \le y \mid x, z_g, t, \vartheta)\, dQ_g^\star(x). \tag{3}$$

This standardization prevents changes in the composition of the candidate pool from being mistaken for frontier progress, and in the bounded-score setting considered here the operational frontier is

$$\varphi_g(t) = \left(F^\star_{gt}\right)^{-1}(1 - m_\star^{-1}), \tag{4}$$

where $m_\star$ is a fixed reference pool size. After rescaling to a bounded unit interval, the frontier gap is $\delta_g(t) = 1 - \varphi_g(t)$, and for tolerance $\epsilon > 0$ the time-to-near-plateau functional is

$$T_g(\epsilon) = \inf\{t : \delta_g(t) \le \epsilon\}. \tag{5}$$

Thus plateau time is not an independently observed event but a derived property of the inferred frontier path, so any claim about $T_g(\epsilon)$ inherits both the information and the limitations of the observed selective archive.

2.1. Selection-Aware Dynamic Frontier Model

The fitted model has two linked components because it must account for top-k selection in each snapshot while connecting those snapshots through a shared frontier-gap path. The selection component conditions on the reporting cutoff rather than treating the reported leaders as ordinary independent draws. Let $c_{gt} = y_{gt,[k_{gt}]}$ denote the reporting threshold, so when selected identities are observed the conditional likelihood is

$$L^{\text{sel}}_{gt} = \left[\prod_{r=1}^{k_{gt}} m_{gt,s_{gt,r}}\big(y_{gt,[r]} \mid u_{gt}\big)\right] \times \left[\prod_{j \notin S_{gt}} M_{gtj}(c_{gt} \mid u_{gt})\right]. \tag{6}$$

Here m_{gti} and M_{gti} are candidate-specific marginal densities and cdfs under the latent score model. The first factor

scores the reported leaders, while the second records the information carried by non-reporting because every unreported candidate must have remained below the cutoff.

The dynamic component models the latent frontier through the gap process rather than through a separate event-time layer. With $\varphi_{g,\infty}$ denoting the long-run frontier, define

$$\delta_g(t) = \varphi_{g,\infty} - \varphi_g(t) \geq 0. \tag{7}$$

We use the parametric family

$$\delta_g(t) = \delta_{g,\infty} + \Delta_{g0} \exp(-\lambda_g t^{\nu_g}), \tag{8}$$

with $\Delta_{g0}, \lambda_g, \nu_g > 0$ and $\delta_{g,\infty} \geq 0$. This parameterization separates initial headroom, closure rate, approach shape, and irreducible residual gap. The synthetic studies use the exponential-gap subfamily with $\nu_g = 1$, which yields the closed-form plateau functional

$$T_g(\epsilon) = \begin{cases} \lambda_g^{-1} \log\left(\dfrac{\Delta_{g0}}{\epsilon - \delta_{g,\infty}}\right), & \epsilon > \delta_{g,\infty}, \\ \infty, & \epsilon \leq \delta_{g,\infty}. \end{cases} \tag{9}$$

The coupling is deliberate because, if plateau time is defined by the frontier gap, a model that predicts terminal scores well but reconstructs the path poorly is not solving the inferential problem posed by repeated frontier records.

2.2. Identification Boundary

The identification claim is about what the reporting regime makes observable, not about generic posterior consistency. Repeated top-k records can identify frontier timing only when the reporting rule, candidate-pool information, score model, and dynamic gap family supply enough structure to connect selected observations across time.

Proposition 1 (Repeated snapshots can identify frontier dynamics)**.** *Suppose the reporting rule and auxiliary pool information are observed for all snapshots, the conditional score law underlying the selection-aware likelihood is identifiable, the standardized reference population is fixed, and the dynamic gap family is injective in its parameters. If the archive contains at least three distinct reporting times spanning a nonzero interval, then repeated top-k snapshots can identify the latent frontier-gap path and the derived functional $T_g(\epsilon)$ up to the stated model class.*

This proposition records the positive side of the boundary because, under an explicit dynamic model, repeated selective observations constrain a common path rather than a single terminal distribution, and that temporal coupling makes plateau timing an inferential target rather than a purely speculative extrapolation.

Proposition 2 (Terminal top-k archives do not generically identify $T_g(\epsilon)$)**.** *If only one terminal top-k record is observed, then in general there exist distinct combinations of terminal score law, latent pool configuration, and frontier-gap path that induce the same terminal selected likelihood but different values of $T_g(\epsilon)$. Hence time-to-near-plateau is not generically identified from terminal top-k evidence alone without additional path restrictions.*

Figure 2 gives a concrete version of the negative result. Consider one terminal archive at time $t^\star = 10$ with pool size $m = 1000$ and reported archive size $k = 10$. Let the terminal selected-score law be a generalized Pareto tail with common parameters

$$(\mu, \sigma, \xi) = (0.8161, 0.18, -0.12), \tag{10}$$

and compare two exponential gap paths,

$$\begin{aligned} \delta_A(t) &= 0.5 \exp(-0.1t), \\ \delta_B(t) &= 0.2246644821 \exp(-0.02t). \end{aligned} \tag{11}$$

The paths differ dynamically but agree at the only observed time:

$$\delta_A(t^\star) = \delta_B(t^\star) = 0.1839397206. \tag{12}$$

Both tuples therefore imply the same terminal frontier level, $\varphi_g(t^\star) = 1 - 0.1839397206$, and the same terminal score law. Because terminal-only evidence consists solely of the ordered top-k archive at $t^\star$, and because the two tuples also share (m, k, μ, σ, ξ), their terminal order-statistic likelihoods are identical:

$$\begin{aligned} &L_{t^\star}^{\text{top-}k}(y_{[1:k]}^{\text{top}} \mid A) \\ &\quad = L_{t^\star}^{\text{top-}k}(y_{[1:k]}^{\text{top}} \mid B) \end{aligned} \tag{13}$$

for every admissible ordered top-k vector. The example isolates the ordinary failure mode of terminal selected archives: they constrain the terminal law, but not the pre-terminal trajectory from which plateau timing is computed.

Proposition 3 (Why coupling matters)**.** *Assume δ_g is differentiable in a neighborhood of $T_g(\epsilon)$ and that $-\delta_g'(t) \geq \ell_g > 0$ throughout that neighborhood. If an estimator $\widehat{\delta}_g$ satisfies $\sup_t |\widehat{\delta}_g(t) - \delta_g(t)| \leq r_g$, then its induced plateau-time estimator obeys*

$$\left|\widehat{T}_g(\epsilon) - T_g(\epsilon)\right| \leq \frac{r_g}{\ell_g}. \tag{14}$$

This bound gives the formal reason to evaluate frontier-path functionals directly because plateau-time error is controlled by path error scaled by the local slope, so estimating $T_g(\epsilon)$ cannot be assessed separately from estimating the path from which $T_g(\epsilon)$ is derived, and slow approach to the threshold is intrinsically hard because even modest path error can translate into large timing error when the local slope is small.

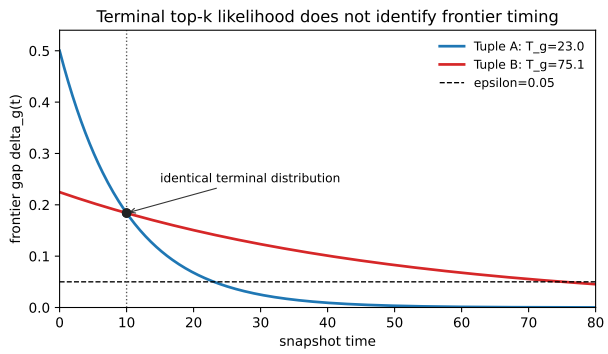

Figure 2. Terminal-only decision non-identification. Both tuples induce the same terminal top-k law at the only observed time $t^\star = 10$, but they imply different frontier-gap trajectories and different crossing times for $\epsilon = 0.05$.

3. Bayesian Testing as Decision for Frontier Evaluation

We use Bayesian decision theory in the standard sense of a statistical decision problem with a state space, an action space, and a loss function (Berger, 1985, Chs. 1–4). In our notation, the latent state is $\theta \in \Theta$, the observed data are D, the available actions form $\mathcal{A}$, and the loss of taking action a in state θ is $L(a, \theta)$. The likelihood $p(D \mid \theta)$ and prior $\pi(\theta)$ determine the posterior

$$\pi(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{\int p(D \mid \vartheta)\pi(\vartheta)\, d\vartheta}. \tag{15}$$

Berger's posterior decision analysis then evaluates an action by averaging its loss under the posterior rather than by treating the estimate, interval, or test label as the final object. We write this posterior risk as

$$\rho(a \mid D) = \mathbb{E}[L(a, \theta) \mid D] = \int L(a, \theta)\pi(\theta \mid D)\, d\theta, \tag{16}$$

and choose a Bayes action

$$a^\star(D) \in \arg\min_{a \in \mathcal{A}} \rho(a \mid D). \tag{17}$$

This formulation keeps the manuscript's notation while following the conditional Bayes principle: once the data are observed, the relevant distribution for decision making is the posterior, and the selected action minimizes posterior expected loss (Berger, 1985, §§ 1.5.1 and 4.4.1).

A hypothesis is a region $H \subset \Theta$, not necessarily a primitive null model. Its posterior probability is

$$\mathbb{P}(H \mid D) = \int_H \pi(\theta \mid D)\, d\theta. \tag{18}$$

For finite action problems, Berger emphasizes that the actions often determine which hypotheses matter: the analyst faces possible actions first, and the state-space partition matters because it changes the loss of those actions (Berger, 1985, § 4.4.3). A classical-looking threshold test is therefore a special case of a decision rule. Suppose the choice is between acting as though H is true, a_1, and acting as though it is false, a_0. If a false positive costs c_{10} and a false negative costs c_{01}, while correct decisions have zero loss, then

$$\rho(a_1 \mid D) = c_{10}\mathbb{P}(H^c \mid D), \tag{19}$$

$$\rho(a_0 \mid D) = c_{01}\mathbb{P}(H \mid D). \tag{20}$$

The action a_1 is chosen exactly when

$$\mathbb{P}(H \mid D) > \frac{c_{10}}{c_{10} + c_{01}}. \tag{21}$$

The threshold is a consequence of the loss ratio rather than a universal significance level. When the cost of a miss is much larger than the cost of overacting, the appropriate posterior-probability threshold can be far below one half.

The same framework covers the option of gathering more evidence. In Berger's preposterior formulation, experiment choice is made before the future data and posterior are observed, so the decision maker balances the cost of sampling against the expected reduction in decision loss (Berger, 1985, Ch. 7). If protocol e has cost $C(e)$ and would generate future data D' from the posterior predictive distribution, then collecting that evidence is preferred to acting immediately when

$$C(e) + \mathbb{E}_{D' \mid D, e}\left[\min_{a \in \mathcal{A}} \mathbb{E}[L(a, \theta) \mid D, D']\right] < \min_{a \in \mathcal{A}} \rho(a \mid D). \tag{22}$$

Thus "collect more data" is not an informal deferral. It is itself a Bayes action, justified only when its cost plus expected post-evidence risk improves on the current Bayes risk.

We now instantiate these definitions for frontier evaluation under selective top-k reporting by letting θ_t denote the latent frontier state at time t. We keep the state abstract enough to include the frontier trajectory, candidate-pool pressure, reporting parameters, and decision-relevant quantities such as capability, safeguards, exposure, and evaluation adequacy, and write the frontier gap as a functional

$$g_t = G(\theta_t, t), \tag{23}$$

with the sign convention that larger g_t means more remaining headroom. For tolerance $\epsilon > 0$, the derived time-to-near-plateau is

$$T_g(\epsilon) = \inf\{t : g_t \le \epsilon\}. \tag{24}$$

Thus plateau timing is not a primitive label because it is inherited from the latent path and the observation model that supports inference about that path.

Hypotheses are subsets of latent state space rather than primitive classical nulls, and useful frontier-evaluation examples include open headroom, near plateau, stale evaluation, imminent near-plateau crossing, and unsafe residual risk:

$$H_{\text{open}}(\epsilon) = \{\theta_T : g_T > \epsilon\}, \tag{25}$$
$$H_{\text{near}}(\epsilon) = \{\theta_T : g_T \le \epsilon\}, \tag{26}$$
$$H_{\text{stale}}(b_{\min}) = \{\theta_T : B_T < b_{\min}\}, \tag{27}$$
$$H_{\text{soon}}(\epsilon, h) = \{\theta_T : T_g(\epsilon) \le T + h\}. \tag{28}$$

The posterior probability of a hypothesis is $\mathbb{P}(H \mid D) = \int \mathbf{1}\{\theta \in H\} p(\theta \mid D)\, d\theta$, but the main object is the Bayes action

$$\begin{aligned} a^\star(D) &= \arg\min_{a \in \mathcal{A}} \rho(a \mid D), \\ \rho(a \mid D) &= \mathbb{E}[L(a, \theta) \mid D]. \end{aligned} \tag{29}$$

The action set used in the simulation studies is

$$\mathcal{A} = \{\text{continue}, \text{refresh}, \text{harden}, \tag{30}$$
$$\text{restrict}, \text{collect}\}. \tag{31}$$

We use a frontier-evaluation loss decomposition

$$L(a, \theta) = L_{\text{miss}} + L_{\text{over}} + L_{\text{stale}} + L_{\text{eval}} + L_{\text{cred}}, \tag{32}$$

where L_{miss} penalizes missed dangerous capability or missed need for stronger mitigation, L_{over} penalizes unnecessary restriction or over-mitigation, L_{stale} penalizes continuing to rely on a saturated or low-headroom evaluation suite, L_{eval} accounts for evaluation cost and delay, and L_{cred} penalizes false reassurance or false certification. This is the main bridge to hypothesis testing: the action threshold is induced by asymmetric loss, not by a universal significance level.

Ordinary posterior-threshold tests appear as special cases. Let a_1 denote hardening, a_0 denote continuing, and let $H_{\text{unsafe},d}$ be the region where residual risk in domain d exceeds an acceptable level. If

$$L(a_1, \theta) = \begin{cases} 0, & \theta \in H_{\text{unsafe},d}, \\ c_{\text{over},d}, & \theta \notin H_{\text{unsafe},d}, \end{cases} \tag{33}$$

$$L(a_0, \theta) = \begin{cases} c_{\text{miss},d}, & \theta \in H_{\text{unsafe},d}, \\ 0, & \theta \notin H_{\text{unsafe},d}, \end{cases} \tag{34}$$

then hardening is chosen iff

$$\mathbb{P}(H_{\text{unsafe},d} \mid D) > \frac{c_{\text{over},d}}{c_{\text{over},d} + c_{\text{miss},d}}. \tag{35}$$

The threshold is set by the loss ratio. In that sense the testing rule is Bayesian and action-induced, not a generic null-rejection recipe.

The collect-more-evidence action is ideally governed by preposterior value of information. Define

$$R^\star(D) = \min_a \rho(a \mid D), \tag{36}$$
$$V(D, D') = \min_a \mathbb{E}[L(a, \theta) \mid D, D']. \tag{37}$$

Then evidence source e is worth collecting when

$$C(e) + \mathbb{E}_{D' \mid D, e}[V(D, D')] < R^\star(D). \tag{38}$$

The synthetic analysis does *not* implement this full posterior-predictive calculation and instead evaluates repeated-snapshot fits under three concrete interventions: one additional snapshot, larger reported k, and harder tasks.

4. Synthetic Decision Studies

The empirical component consists of three synthetic analyses summarized in Table 2. The first varies the repeated-snapshot protocol, the second evaluates decisions from posterior draws rather than scalar summaries, and the third compares a repeated-snapshot coupled model with a terminal-only coupled comparator on paired trajectories. The purpose is to test the decision formulation under controlled observation regimes, not to validate an operational deployment rule.

The repeated-snapshot protocol study evaluates the model on three regimes crossed with three observation or task interventions: another snapshot, larger reported k, and harder tasks. Coverage is complete, and across all 225 fits, 219 binary horizon decisions match the simulated state. The six misses are all false negatives in the selection-heavy closing-gap regime: one under another snapshot and five under harder tasks. Apparent-plateau and slow-frontier cells have $75/75$ correct decisions each, while the selection-heavy cells have $69/75$. Frontier path loss is highest in the slow-frontier cells, especially under harder tasks, even though those binary decisions are correct. Finite-$T_g(\epsilon)$ interval coverage is only $3/225$, so this analysis supports protocol coverage and stable fitting, not calibrated plateau-time uncertainty or strong binary decision reliability.

The posterior decision analysis evaluates actions from posterior draws and covers $374/374$ fitted records. Across the nine loss-family and weight-profile settings in the matched comparison, the modal action is regime-sensitive: apparent-plateau and cohort-arrival cells usually choose to continue monitoring for both observation regimes, while selection-heavy cells choose to refresh the evaluation suite in every setting for both regimes. Repeated-snapshot and terminal-only action agreement is high but not perfect at 93.8% in

Table 1. State-space hypotheses and canonical actions. The same posterior can support multiple hypothesis probabilities, and the Bayes action depends on their joint contribution to posterior risk under the chosen loss.

Hypothesis region	Decision meaning	Typical action response	Dominant loss term
$H_{\text{open}}(\epsilon) = \{\theta_T : g_T > \epsilon\}$	Material headroom remains	Continue monitoring or harden safeguards	missed capability
$H_{\text{near}}(\epsilon) = \{\theta_T : g_T \leq \epsilon\}$	Frontier is near plateau	Continue if the suite remains adequate	over-mitigation
$H_{\text{stale}}(b_{\min}) = \{\theta_T : B_T < b_{\min}\}$	Evaluation suite has low headroom or discrimination	Refresh the suite or collect more evidence	evaluation staleness
$H_{\text{soon}}(\epsilon, h) = \{\theta_T : T_g(\epsilon) \leq T + h\}$	Near-plateau crossing is imminent	Refresh, harden, or restrict depending on risk	missed capability and credibility
$H_{\text{unsafe},d} = \{\theta_T : R_d(\theta_T, a_{\text{base}}) > \tau_d\}$	Residual risk in domain d is too high	Harden safeguards or restrict access	missed capability

Table 2. Synthetic analyses. The main-text claims are intentionally narrower than model superiority or operational decision validation. MCMC diagnostics for included fitted records are reported in Appendix A.

Analysis	Coverage	Manuscript use
Repeated-snapshot protocol study	$225/225$ fitted records across three regimes, three protocol interventions, and 25 trajectories per cell	Assesses protocol coverage and numerical stability rather than a decision-superiority claim
Posterior decision analysis	$374/374$ fitted records with posterior decision evaluations	Supports loss/action tables for repeated snapshots versus terminal-only observation
Matched observation-regime comparison	$300/300$ fitted records, with 150 paired repeated-snapshot and terminal-only trajectories over three regimes	Supports posterior action comparison, while binary horizon decisions do not separate the observation regimes

apparent-plateau, 95.6% in cohort-arrival, and 97.8% in selection-heavy after aggregating over loss settings and paired trajectories. This posterior action comparison is useful for locating loss-sensitive decision boundaries, but it is not evidence that repeated snapshots dominate terminal-only archives in binary decision quality. Table 3 makes that limitation concrete: in the matched comparison, the repeated-snapshot model has much lower frontier path loss than the terminal-only comparator in all three regimes, but the binary horizon decision is tied because both baselines are perfect in apparent-plateau and cohort-arrival cells and both miss the same $9/50$ true positives in selection-heavy. The informative signal is therefore not a clean binary decision win but the combination of posterior action disagreement under specified losses and lower repeated-snapshot path loss.

5. Discussion

The main supported claim is that repeated top-k snapshots and terminal-only archives define different action-relevant posterior problems under selective reporting. The mathematical reason is that hypotheses about open headroom, near plateau, and evaluation staleness are regions of a latent path-dependent state space, and the loss-aware decision rule depends on posteriors over those regions. The empirical evidence is narrower but concrete: posterior decision analysis can compare repeated-snapshot and terminal-only observation regimes on matched fitted trajectories, and the repeated-snapshot protocol study remains numerically stable across the considered synthetic interventions.

Several limitations remain explicit. Clean diagnostics and complete fitted-record coverage do not imply calibrated plateau-time uncertainty because finite-$T_g(\epsilon)$ coverage is poor in the protocol study and remains uneven in the matched comparison. The binary horizon decisions do not support a superiority claim for repeated snapshots because the two observation regimes tie on that outcome and share the same selection-heavy false negatives. The collect-more-evidence rule is specified mathematically but not validated as a full preposterior policy in these experiments, and all empirical evidence is synthetic rather than a certification of real frontier-model governance decisions. The contribution is therefore an action-facing formulation, an observability boundary, and posterior action tables under the reported simulation conditions, not an operationally validated action rule or a broad claim that the current coupled model is decision-quality superior.

6. Conclusion

Bayesian testing for frontier evaluation under selective reporting is better framed as decision theory than as primitive null rejection because repeated top-k snapshots and

Table 3. Matched comparison by regime and observation baseline. Lower frontier loss favors repeated snapshots, but binary horizon decisions do not separate the baselines.

Regime	Observation	Correct	FN	$T_g(\epsilon)$ cov.	Path loss
apparent plateau	repeated snapshots	50/50	0	0.04	0.021
apparent plateau	terminal only	50/50	0	0.08	0.208
cohort arrival	repeated snapshots	50/50	0	0.04	0.065
cohort arrival	terminal only	50/50	0	0.24	0.271
selection heavy	repeated snapshots	41/50	9	0.00	0.046
selection heavy	terminal only	41/50	9	0.22	0.269

terminal-only archives are different observability regimes for the posterior quantities that matter to action. The synthetic studies support the action-facing formulation and the need to reason about observation regimes while showing why stronger claims should wait: binary decision superiority, calibrated $T_g(\epsilon)$ uncertainty, and operational frontier-evaluation validity are not established here.

References

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

Einmahl, J. H. J. and He, Y. Accurate estimates of ultimate 100-meter records. *arXiv preprint arXiv:2502.04085*, 2025. doi: 10.48550/arXiv.2502.04085.

Jackson, C., Hawkins, N., et al. Cure models: What is meant by a survival "plateau," and do experts agree on what constitutes one? *Statistics in Medicine*, 2025.

Liu, X., Yang, T. T., and Zhang, Y. Quasi-bayesian inference for production frontiers. *Journal of Business & Economic Statistics*, 2022.

Yuen, T. P. and Musta, E. Testing for sufficient follow-up in survival data with a cure fraction. *arXiv preprint arXiv:2403.16832*, 2026. doi: 10.48550/arXiv.2403.16832.

Zrnic, T. and Fithian, W. A flexible defense against the winner's curse. *The Annals of Statistics*, 2025. doi: 10.1214/25-AOS2553.

A. MCMC Diagnostics for Included Fits

All HMC-backed fits reported in the main text used the eight-chain diagnostic standard $\hat{R} < 1.01$, bulk ESS at least 800, and zero divergences for accepted fitted records. Table 4 therefore includes only fitted-record families whose numerical results appear in the main text.

The appendix is restricted to fitted records that support numerical claims in the main text. Exploratory checks and unsuccessful fits are not tabulated because their numerical results are not used in the reported comparisons. This keeps the diagnostic summary aligned with the paper's evidentiary boundary.

Table 4. MCMC diagnostics for fitted records included in the main text. ESS denotes the minimum reported bulk ESS unless otherwise stated.

Fit family	Records	Failure status	$\hat{R}$ max	ESS min	Main-text role
Repeated-snapshot protocol study	225	0 failures, 0 divergences	1.0027	5144	Supports the repeated-snapshot protocol study and its $219/225$ binary horizon decisions.
Small matched comparison	24	0 failures, 0 divergences	1.003	4362	Included in the posterior decision analysis.
Medium matched comparison	50	0 failures, 0 divergences	1.003	5492	Included in the posterior decision analysis.
Full matched comparison	300	0 failures, 0 divergences	1.0027	5212	Supports the repeated-snapshot versus terminal-only comparison in Table 3.