ANYTIME-VALID LLM LEADERBOARDS VIA BENCHMARK-WEIGHTED AND BLOCK-FACTORIZED e-PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) leaderboards compare model capabilities by ranking models according to their mean performance on fixed benchmarks. However, variability in evaluation outcomes across runs may produce unsupported claims of model superiority on the benchmark, a risk compounded by leaderboard updates. In this paper, we propose BB-EDGE (Benchmark-Weighted and Block-Factorized e-processes for Directed Graph Evaluation), a principled framework that represents an LLM leaderboard as a directed graph whose edges certify pairwise mean-performance advantages, with anytime-valid family-wise error rate (FWER) control. Concretely, for each direction, BB-EDGE constructs an empirical-Bernstein e-process by factorizing evidence over protocol-defined blocks and assigning stakes proportional to the corresponding block weights, then applies direct e-Holm across these e-processes to certify directional advantages as edges. Theoretically, we characterize weight-proportional linear stakes under heterogeneous benchmark-average nulls and prove anytime FWER control under arbitrary within-block and cross-pair dependence. BB-EDGE further supports anytime-valid Top-k certification and simultaneous rank intervals. Extensive experiments on synthetic data and four real-world benchmarks demonstrate that BB-EDGE maintains anytime FWER control while achieving high efficiency.

1 INTRODUCTION

Large language model (LLM) leaderboards have become the dominant means of comparing LLM capabilities. They rank models by mean performance on a fixed benchmark under a standardized protocol and are often updated over successive evaluation runs (Biderman et al., 2024; Madaan et al., 2024). Due to stochastic factors such as sampled decoding (Gonzalez et al., 2025; Du et al., 2026) and LLM-based judging (Chen et al., 2024; Divekar & Majumder, 2026), LLM responses and their evaluation scores vary across runs. Consequently, a higher empirical mean does not by itself establish that one model has higher expected performance than another (Ye et al., 2024; Vashurin et al., 2025; Neuhof & Benjamini, 2026). Treating such score differences as definitive evidence of model superiority can lead to insufficiently supported claims, thereby undermining the credibility of leaderboards. This highlights the need for *anytime-valid* LLM leaderboards whose model comparisons remain statistically guaranteed across all updates.

Several approaches represent leaderboards as confidence graphs (Wang et al., 2024b), quantify ranking uncertainty through rank intervals (Li et al., 2026; Neuhof & Benjamini, 2026), or support model selection through Top-k identification (Fan et al., 2025). However, these methods target a fixed evaluation budget and do not provide anytime-valid guarantees under repeated monitoring or data-dependent stopping. Other work has developed online evaluation procedures for evaluating individual LLMs (Hsu & Shekhar, 2026; Zhou et al., 2026b) and inferring pairwise human preferences (Chiang et al., 2024; Gu et al., 2026), typically under an independent and identically distributed item stream. Simulations show that an i.i.d. item-stream procedure can falsely certify benchmark-average advantages when item-level model differences are heterogeneous despite a near-zero benchmark-average difference, or when observations within multi-turn dialogues are dependent. Same-model comparisons also reveal false certifications on real-world benchmarks.

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

In this paper, we develop a principled framework for efficiently constructing anytime-valid confidence graphs that certify pairwise model advantages as edges while controlling the family-wise error rate (FWER), i.e., the probability of ever reporting a false edge across all replicates. Constructing such confidence graphs presents two key challenges: arbitrary dependence within multi-turn dialogues or multi-step trajectories, and heterogeneity in item-level model advantages. To address both, we propose Benchmark-weighted and Block-Factorized e-processes for Directed Graph Evaluation (**BB-EDGE**). In particular, BB-EDGE factorizes empirical-Bernstein evidence across protocol-defined blocks and uses weight-proportional stakes to preserve the heterogeneous benchmark-average null. It then applies direct e-Holm to certify mean-performance advantages with anytime FWER control. We further provide a theoretical analysis showing that BB-EDGE's weight-proportional stakes are sufficient and necessary for nonpositive linear drift over the full benchmark-average null, and that the resulting procedure guarantees anytime-valid FWER control.

Extensive experiments on both synthetic and real-world datasets validate both the anytime validity and efficiency of BB-EDGE. The simulation results demonstrate that BB-EDGE maintains anytime-valid FWER control while achieving the highest power among FWER-controlling baselines across the examined scenarios. In real-world evaluations, we apply BB-EDGE to ten LLMs across four benchmarks, the single-turn benchmarks MMLU-Pro (Wang et al., 2024a) and GSM8K (Cobbe et al., 2021), the multi-turn dialogue benchmark MT-Bench (Zheng et al., 2023), and ACEBench (Chen et al., 2025). For example, on GSM8K, BB-EDGE certifies 41 of the 45 pairwise comparisons (91.1%) after just 2 replicates, using 67% fewer replicates than the strongest competing baseline. Moreover, BB-EDGE extends to anytime-valid Top-k certification and simultaneous rank intervals enabling certified model selection and ranking uncertainty quantification.

Our contributions are summarized as follows:

1. We introduce BB-EDGE, an anytime-valid and efficient framework for certifying LLM leaderboards. It uses block-factorized empirical-Bernstein e-processes with weight-proportional stakes under heterogeneous block means and arbitrary within-block dependence, while preserving the benchmark-average estimand.
2. We theoretically show that proportionality is necessary and sufficient among nonnegative stake allocations for nonpositive linear drift over the full benchmark-average null. Combining the resulting e-processes with direct e-Holm yields anytime FWER control across all directions and completed replicates.
3. We conduct extensive experiments on synthetic and real-world datasets, showing that BB-EDGE controls anytime FWER while achieving high power. We further extend BB-EDGE to anytime-valid Top-k certification and simultaneous rank intervals.

2 Related Work

LLM Leaderboards. Existing work on LLM leaderboards largely focuses on benchmark construction and evaluation protocols, spanning single-turn knowledge and reasoning tasks (Peng et al., 2024; Zhu et al., 2024), multi-turn dialogues (Zheng et al., 2023), and multi-step trajectories (Chen et al., 2025; Nakash et al., 2026). Evaluation outcomes may also vary across repeated runs because of sampled decoding (Gonzalez et al., 2025; Du et al., 2026) and stochastic LLM-based judging (Chen et al., 2024; Divekar & Majumder, 2026). Nevertheless, common practice summarizes each model by a single average score under a fixed protocol and reports the resulting ranking (Ye et al., 2024; Frick et al., 2025; Wang et al., 2025b; Zhou et al., 2026a), without quantifying which pairwise performance advantages are statistically supported.

Model Ranking. A separate line of work develops statistical methods for model ranking and uncertainty quantification. Some methods replace the raw mean score with alternative ranking criteria designed to improve reliability (Wei et al., 2024; Hariri et al., 2026). Others provide inferential guarantees using Holm-adjusted paired one-sided t-tests for simultaneous pairwise comparisons (Chandrahas, 2026) or conformal methods for rank uncertainty quantification (Neuhof & Benjamini, 2026). These methods generally target inference at a fixed evaluation budget and therefore do not protect against repeated examination of the ranking. Related work also ranks models using estimated pairwise preference probabilities (Chatzi et al., 2024; Haghtalab et al., 2026; Huang et al.,

2026; Menendez et al., 2026), with extensions to online preference data (Chiang et al., 2024; Gu et al., 2026). This estimand differs from that of objectively scored benchmarks such as MMLU-Pro (Wang et al., 2024a) and GSM8K (Cobbe et al., 2021), where models are evaluated against fixed ground truth and ranked by their mean task scores.

Anytime-Valid Inference. Sequential data collection is common in many real-world LLM evaluation (Wang et al., 2025a; Singh et al., 2026). Anytime-valid procedures preserve statistical validity under continuous monitoring and data-dependent stopping (Ramdas et al., 2023; Xu et al., 2024). Recent work applies e-processes to reduce the number of items required for evaluating individual LLMs (Hsu & Shekhar, 2026; Zhou et al., 2026b) or to infer pairwise preference probabilities from human preferences arriving online (Gu et al., 2026). These methods formulate evaluation as a stream of i.i.d. items. This formulation does not directly accommodate fixed benchmarks with heterogeneous item-level model advantages, where one model may outperform another on some items but underperform it on others, nor arbitrary dependence within multi-turn dialogues or multi-step trajectories. In contrast, this paper targets pairwise mean-performance differences on a fixed benchmark and controls the probability of reporting any false edge across all model pairs and replicates.

3 Statistical Framework for Anytime-Valid LLM Leaderboards

We consider a sequential evaluation setting in which LLMs are evaluated on a fixed benchmark across multiple replicates. In this section, we formalize this setting, represent LLM leaderboards through certified edges, and outline our goal of achieving *anytime-valid* and *efficient* certification.

Sequential Evaluation for LLM Leaderboards. An LLM leaderboard compares L candidate models on a fixed benchmark $\mathcal{B} = \{x_1, \ldots, x_N\}$, where each x_i denotes an evaluation item. All models are evaluated under a standardized protocol π that specifies the prompts, decoding configuration, available tools, grading rules, and dataset structure.

In the sequential setting, the benchmark is evaluated over successive complete replicates $r = 1, 2, \ldots$, where each replicate evaluates every model on all N items in $\mathcal{B}$. Let $S_{i,r}^{(\ell)} \in [0,1]$ be the evaluation score of model ℓ on item i in the r-th complete replicate. We define the mean performance of model ℓ on $\mathcal{B}$ as

$$\theta_\ell^{\mathcal{B}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} S_{i,r}^{(\ell)} \middle| \mathcal{B}, \pi\right].$$

where the expectation is taken over the randomized execution of the fixed protocol π. Leaderboards rank models by the empirical mean $\hat{\theta}_{\ell,r} = \frac{1}{rN}\sum_{s=1}^{r}\sum_{i=1}^{N} S_{i,s}^{(\ell)}$, often computed from a single run ($r = 1$), which is an unbiased estimate of $\theta_\ell^{\mathcal{B}}$; thus $\theta_\ell^{\mathcal{B}}$ is precisely the quantity that leaderboards implicitly estimate. However, this estimate varies across runs, so single-run rankings can change from one run to the next. For example, on GSM8K, Figure 13 shows that the single-run ranking differs from the 20-replicate ranking in 45% of replicates (9 of 20), with all reversals occurring among three closely matched models. We therefore need to certify comparisons of $\theta_\ell^{\mathcal{B}}$ with statistical guarantees rather than relying on point estimates, and the margin τ introduced below lets practitioners require practically meaningful advantages rather than merely nonzero ones. Comparisons of $\theta_\ell^{\mathcal{B}}$ concern performance on $\mathcal{B}$ under π, not generalization to unseen items.

Ranking via Certified Edges. Let $\overrightarrow{\mathcal{P}} = \{(a,b) \in [L]^2 \mid a \neq b\}$ denote the ordered pairs of distinct models, where $[L] = \{1, \ldots, L\}$. For each $(a,b) \in \overrightarrow{\mathcal{P}}$, define the mean-performance difference on $\mathcal{B}$ as $\Delta_{a,b}^{\mathcal{B}} = \theta_a^{\mathcal{B}} - \theta_b^{\mathcal{B}}$. For a prespecified margin $\tau \in [0,1)$, each direction $e = (a \to b)$ is tested via

$$H_e^{(\tau)} : \Delta_{a,b}^{\mathcal{B}} \leq \tau \qquad \text{against} \qquad \Delta_{a,b}^{\mathcal{B}} > \tau.$$

Rejecting $H_e^{(\tau)}$ certifies the edge $a \to b$, indicating that model a outperforms model b on the fixed benchmark $\mathcal{B}$ by more than τ. If neither direction is certified, the pair remains *unresolved*, that is, the evidence is insufficient to establish that either model outperforms the other by more than τ. The certified edge set $\mathcal{E}$ assembles into a directed comparison graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the L models, which serves as the leaderboard. A directed path $a \rightsquigarrow b$ of length d certifies $\Delta_{a,b}^{\mathcal{B}} > d\tau$, so the graph reports a partial order in which every reachability relation is statistically supported.

Anytime Validity and Power. Let $\mathcal{E}_r = \{e \in \overrightarrow{\mathcal{P}} : H_e^{(\tau)} \text{ is rejected at replicate } r\}$ denote the set of directly certified edges at replicate r, R the maximum number of complete replicates, and $\alpha \in (0,1)$ the target level. For $1 \le r \le R$, define the anytime family-wise error rate

$$\mathrm{FWER}_r = \mathbb{P}\left(\exists s \le r,\ \exists e = (a \to b) \in \mathcal{E}_s \text{ such that } \Delta_{a,b}^{\mathcal{B}} \le \tau \,\middle|\, \mathcal{B}, \pi\right). \tag{1}$$

We say that a procedure is *anytime-valid* if $\mathrm{FWER}_r \le \alpha$ for every $r \le R$. It permits the graph to be examined after every completed replicate $r \le R$ while maintaining simultaneous error control over all such monitoring replicates.

To measure efficiency, we also evaluate Power_r, the expected proportion of true edges certified:

$$\mathrm{Power}_r = \mathbb{E}\left[\frac{\left|\left\{e = (a \to b) \in \mathcal{E}_r : \Delta_{a,b}^{\mathcal{B}} > \tau\right\}\right|}{\left|\left\{e = (a \to b) \in \overrightarrow{\mathcal{P}} : \Delta_{a,b}^{\mathcal{B}} > \tau\right\}\right|} \,\middle|\, \mathcal{B}, \pi\right]. \tag{2}$$

Ideally, a practical approach should simultaneously ensure that $\mathrm{FWER}_r \le \alpha$ at every replicate r while attaining as high a Power_r as possible.

4 METHODOLOGY

In this section, we first describe the statistical challenges of fixed-benchmark inference. We then propose BB-EDGE for constructing anytime-valid and efficient leaderboards.

Dependence and Heterogeneity. Certifying mean-performance advantages on a fixed benchmark $\mathcal{B}$ and protocol π presents two key statistical challenges:

- Dependence. Evaluation scores arising from the same multi-turn dialogue or multi-step trajectory may be arbitrarily dependent.
- Heterogeneity. Item-level performance differences may vary across the benchmark, while the directional null constrains only their benchmark-weighted average.

Both features are common in modern LLM benchmarks. For example, Figure 14 (a) shows that Gemma-4-E4B-it and MiMo-7B-RL have opposing item-level advantages on GSM8K, despite a near-zero benchmark-average gap of -0.008. In Section 6.1, we empirically show that either feature can cause item-level online methods to inflate the FWER. To address both, we propose BB-EDGE, which factorizes evidence across protocol-defined blocks and places weight-proportional stakes that preserve the heterogeneous benchmark-average null. We present its construction in three steps.

Design-Determined Resolution. We first partition the fixed benchmark $\mathcal{B}$ into M prespecified blocks $\mathcal{C} = (C_1, \ldots, C_M)$ based on the dependence structure induced by the fixed protocol π. For each direction $e = (a \to b)$, define the item-level score difference $D_{i,r}^e = S_{i,r}^{(a)} - S_{i,r}^{(b)}$, its rescaled version $X_{i,r}^e = \frac{D_{i,r}^e + 1}{2} \in [0,1]$, and the null reference point $\mu_0 = \frac{1+\tau}{2}$. Then the directional null $H_e^{(\tau)} : \Delta_{a,b}^{\mathcal{B}} \le \tau$ is equivalently written as $\frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[X_{i,r}^e \mid \mathcal{B}, \pi\big] \le \mu_0$. For each block C_m, let $w_m = |C_m|/N$ and define the block-level observation $Y_{e,m,r} = \frac{1}{|C_m|}\sum_{i \in C_m} X_{i,r}^e$. These block summaries preserve the original benchmark-level target because

$$\sum_{m=1}^M w_m Y_{e,m,r} = \frac{1}{N}\sum_{i=1}^N X_{i,r}^e. \tag{3}$$

Thus, the partition changes only the evidence resolution, not the estimand. Here, $M = 1$ permits arbitrary within-replicate dependence, $M = N$ enables item-level factorization under conditional independence, and intermediate choices accommodate within-block dependence.

Benchmark-Weighted Block-Factorized e-Process. Let $W_{i,r} = (S_{i,r}^{(1)}, \ldots, S_{i,r}^{(L)})$ collect the scores of all L models on item i, and let $Z_r = (W_{1,r}, \ldots, W_{N,r})$ denote the complete evaluation panel at replicate r. Let $\mathcal{Q}$ collect all pre-inference information, including the directional family, stake grid, mixture weights, and any independent pilot data used to specify them. We define

$\mathcal{F}_r = \sigma(\mathcal{B}, \pi, \mathcal{Q}, Z_1, \ldots, Z_r)$ and, for each direction $e = (a \to b)$, let $\nu_{e,m,r} = \mathbb{E}[Y_{e,m,r} \mid \mathcal{F}_{r-1}]$ denote the conditional mean of block m before replicate r. Under $H_e^{(\tau)}$, the block means need not individually satisfy $\nu_{e,m,r} \le \mu_0$. The null constrains only their benchmark-weighted average

$$\sum_{m=1}^{M} w_m \nu_{e,m,r} = \frac{1 + \Delta_{a,b}^{\mathcal{B}}}{2} \le \mu_0. \tag{4}$$

To preserve the benchmark-weighted constraint in (4), let $\Lambda = \{\lambda_1, \ldots, \lambda_G\} \subset [0, 1)$ be a prespecified stake grid and define $w_\star = \max_{m \in [M]} w_m$,

$$\lambda_{g,m} = \lambda_g \frac{w_m}{w_\star}. \tag{5}$$

The linear drift satisfies $\sum_{m=1}^{M} \lambda_{g,m}(\nu_{e,m,r} - \mu_0) = \frac{\lambda_g}{w_\star}\left(\sum_{m=1}^{M} w_m \nu_{e,m,r} - \mu_0\right) \le 0$ under $H_e^{(\tau)}$. Because the benchmark-average null does not require $\nu_{e,m,r} \le \mu_0$ for each block, naive blockwise e-processes with arbitrary block-specific stakes can induce positive drift under $H_e^{(\tau)}$. Proposition 1 characterizes the linear drift requirement used in our factorized construction: among nonnegative blockwise stake allocations, proportionality to the benchmark weights is necessary and sufficient for this drift to be nonpositive under every admissible vector of block means.

Before replicate r, let $\widehat{Y}_{e,m,r} \in [0, 1]$ be a predictable estimate of $Y_{e,m,r}$. For each $\lambda_g \in \Lambda$, building on the predictable empirical-Bernstein construction of Waudby-Smith & Ramdas (2024), define the block-factorized empirical-Bernstein contribution

$$F_{e,r}^{\mathcal{C}}(\lambda_g) = \exp\left\{\sum_{m=1}^{M}\left[\lambda_{g,m}(Y_{e,m,r} - \mu_0) - \psi_E(\lambda_{g,m})(Y_{e,m,r} - \widehat{Y}_{e,m,r})^2\right]\right\}, \tag{6}$$

where $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$. Let $\rho_1, \ldots, \rho_G$ be prespecified mixture weights satisfying $\rho_g \ge 0$ and $\sum_{g=1}^{G} \rho_g = 1$. The corresponding evidence process is

$$E_{e,r}^{\mathcal{C}}(\lambda_g) = \prod_{s=1}^{r} F_{e,s}^{\mathcal{C}}(\lambda_g), \qquad E_{e,r}^{\mathcal{C}} = \sum_{g=1}^{G} \rho_g E_{e,r}^{\mathcal{C}}(\lambda_g). \tag{7}$$

with $E_{e,0}^{\mathcal{C}}(\lambda_g) = E_{e,0}^{\mathcal{C}} = 1$. Together, (6) and (7) define the evidence process used by BB-EDGE. The proportional stakes ensure nonpositive drift under the heterogeneous benchmark-average null, while the block-specific predictable estimates provide variance adaptation without changing the target hypothesis. Section 5 establishes the validity of this construction.

Confidence Graph Construction. At each completed replicate r, we apply direct e-Holm (Hartog & Lei, 2025) to $\{E_{e,r}^{\mathcal{C}} : e \in \overrightarrow{\mathcal{P}}\}$. Specifically, define $\mathcal{J}_r = \left\{e \in \overrightarrow{\mathcal{P}} : E_{e,r}^{\mathcal{C}} < \frac{1}{\alpha}\right\}$ and the rejection threshold $c_r = \frac{1}{\alpha} + \sum_{e \in \mathcal{J}_r}\left(\frac{1}{\alpha} - E_{e,r}^{\mathcal{C}}\right)$. The set of directly certified edges is

$$\mathcal{E}_r = \left\{e \in \overrightarrow{\mathcal{P}} : E_{e,r}^{\mathcal{C}} \ge c_r\right\}. \tag{8}$$

Direct e-Holm exploits the evidence across the entire directional family and remains valid under arbitrary dependence among the pairwise e-processes.

The resulting confidence graph is $\mathcal{G}_r = (\mathcal{V}, \mathcal{E}_r)$, where an edge $a \to b$ certifies $\Delta_{a,b}^{\mathcal{B}} > \tau$. A directed path $a \rightsquigarrow_r b$ of length d implies $\Delta_{a,b}^{\mathcal{B}} > d\tau$, inducing a partial ranking in which unreachable pairs remain unresolved. The edge sets need not be nested over time, and Section 5 establishes anytime FWER control across all model pairs and replicates. The confidence graph further supports anytime-valid **Top-k certification** and **rank intervals** (Appendices A and B).

5 Theoretical Analysis

We establish validity using nonnegative supermartingales (Howard et al., 2020). The benchmark-level null imposes a constraint on the weight-averaged block means; proportional stakes preserve this constraint and yield a valid e-process for each direction. Direct e-Holm then gives anytime false-edge control for the full graph.

Assumption 1. *The following assumptions hold conditional on the fixed benchmark and protocol:*

1. *Conditional on $(\mathcal{Q}, \mathcal{B}, \pi)$, the complete panels $Z_1, \ldots, Z_R$ are independent and identically distributed across runs.*
2. *Within each run, the block vectors $V_{m,r} = (W_{i,r} \mid i \in C_m)$ are mutually independent conditional on the history $(\mathcal{F}_{r-1}, \mathcal{B}, \pi)$. Their distributions may differ across blocks.*
3. *Let $\mathcal{Q}$ denote the information used to specify the partition, directional family, stake grid, and mixture weights. These objects are $\mathcal{Q}$-measurable, and $\mathcal{Q}$ is independent of the inference panels $Z_1, \ldots, Z_R$ conditional on $(\mathcal{B}, \pi)$. In particular, they are fixed before the inference runs or selected using independent data.*
4. *Each prediction $\widehat{Y}_{e,m,r}$ is $\mathcal{F}_{r-1}$-measurable and lies in $[0, 1]$.*

For a finite budget R, all guarantees are restricted to $1 \le r \le R$. Statements covering every $r \ge 1$ require the same conditions for the entire infinite sequence of inference panels.

Remark 1. *The independence conditions concern the data collection procedure and cannot be verified from the completed score table alone. Different random seeds do not establish independence if runs share a provider state, a cache, or a source of evaluator randomness. If the protocol does not justify independence across the selected blocks, the theorem-backed fallback is the one-block complete-run analysis. Singleton blocks require independent randomization of each item vector, but they do not require the item vectors to be exchangeable or identically distributed.*

Under a true directional null, (4) bounds only the benchmark-weighted average of the block means. Individual blocks may favor either model, which is why unrestricted block-specific stakes can turn a valid benchmark-level null into a different claim.

Proposition 1. *Let $w_m > 0$, $\sum_m w_m = 1$, and $\mu_0 \in (0, 1)$. Suppose a nonnegative stake vector η satisfies*

$$\eta^\top(\nu - \mu_0 \mathbf{1}) \le 0 \quad \text{whenever} \quad \nu \in [0,1]^M \text{ and } w^\top \nu \le \mu_0. \tag{9}$$

Then $\eta = cw$ for some $c \ge 0$. Conversely, every vector of this form satisfies

$$\eta^\top(\nu - \mu_0 \mathbf{1}) = c\, w^\top(\nu - \mu_0 \mathbf{1}) \le 0$$

for every $\nu \in [0,1]^M$ with $w^\top \nu \le \mu_0$.

Thus, proportional stakes are necessary for nonpositive linear drift over the full benchmark-level null, not merely a convenient choice. The proof is given in Appendix D.

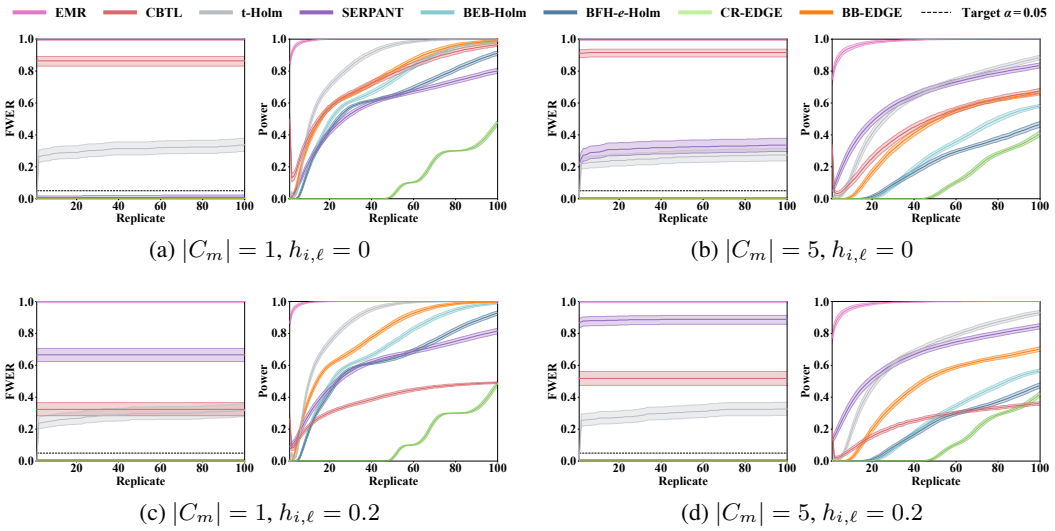
Theorem 1. *Suppose Assumption 1 holds. For every true directional null, the process $\{E^{\mathcal{C}}_{e,r}\}_{r=0}^R$ in (7) is a nonnegative supermartingale. Let $\mathcal{N}_0 = \{e = (a \to b) \in \overrightarrow{\mathcal{P}} : \Delta^{\mathcal{B}}_{a,b} \le \tau\}$ denote the set of true directional nulls. If the edge sets $\mathcal{E}_r$ are obtained by direct e-Holm, then*

$$\mathbb{P}\left(\exists r \le R,\ \exists e \in \mathcal{N}_0 :\ e \in \mathcal{E}_r \middle| \mathcal{B}, \pi\right) \le \alpha. \tag{10}$$

On the complementary event of probability at least $1 - \alpha$, every direct edge is a true margin claim, every path $a = v_0 \to \cdots \to v_d = b$ implies $\Delta^{\mathcal{B}}_{a,b} > d\tau$, and the graph is acyclic.

The complete proof is given in Appendix D.3. The displayed FWER bound conditions on the benchmark and protocol, averaging over any independent pilot used to fix the design. It does not require independence among the directional e-processes. The theorem applies to every resolution fixed in advance and justified by the evaluation design. One block gives the robust complete-run graph, singleton blocks give the item-factorized graph when item-level independence is justified, and intermediate partitions provide the corresponding intermediate resolutions. The resolution cannot be selected after observing which graph reports more edges.

Validity alone does not determine certification speed. For an alternative $e = (a \to b)$, let $\delta_e = (\Delta^{\mathcal{B}}_{a,b} - \tau)/2 > 0$. The conditional expected log increment of grid component g is $a_{e,g,r} = \frac{\lambda_g}{w_\star}\delta_e - \sum_{m=1}^M \psi_E(\lambda_{g,m}) \mathbb{E}\Big[(Y_{e,m,r} - \widehat{Y}_{e,m,r})^2 \mid \mathcal{F}_{r-1}\Big]$. The first term captures the benchmark-average advantage, while the second accounts for blockwise prediction error. Theorem 4 shows that a positive-weight grid component with persistently positive average conditional log growth yields eventual certification.



(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$

(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

Figure 1: Empirical FWER_r and Power_r (with 95% confidence intervals) versus replicate r. Panels (a)–(d) correspond to the i.i.d., dependence-only, heterogeneity-only, and combined dependence-and-heterogeneity settings, respectively. All settings use $\tau = 0$, $N = 100$, and $L = 10$. The dashed line marks the target level $\alpha = 0.05$.

6 Experiments

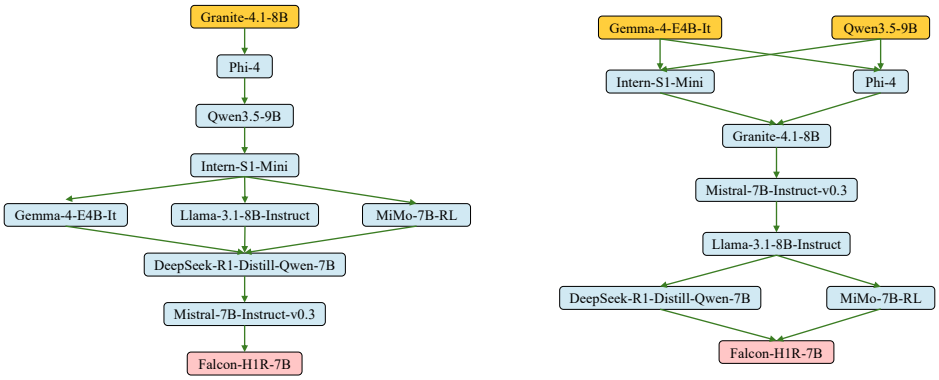
6.1 Synthetic Data Analysis

Data-Generating Process. We simulate $L = 10$ models on a fixed benchmark of $N = 100$ items. Their benchmark-level means are $(\theta_1, \theta_1, \ldots, \theta_{L/2}, \theta_{L/2})$, with $\theta_1, \ldots, \theta_{L/2}$ evenly spaced over $[0.41, 0.59]$. To specify ground-truth fixed-benchmark performance, we define the probability that model ℓ correctly answers item i as $p_i^{(\ell)} = \theta_\ell + c_i + h_{i,\ell}$, where c_i is a fixed item-difficulty offset, equally spaced over $[-0.08, 0.08]$, and $h_{i,\ell}$ is a model-item effect constructing heterogeneous benchmark-average nulls. Both effects are centered so that $\sum_{i=1}^N c_i = \sum_{i=1}^N h_{i,\ell} = 0$ for every ℓ.

We generate the LLMs' synthetic outcomes via a latent Gaussian variable $Z_{i,\ell,r} = \sqrt{\kappa_1}\, U^{\mathrm{block}}_{b(i),\ell,r} + \sqrt{\kappa_2}\, U^{\mathrm{item}}_{i,r} + \sqrt{1-\kappa_1-\kappa_2}\, \varepsilon_{i,\ell,r}$, where U^{block}, U^{item}, and ε are independent standard Gaussian variables, and κ_1 and κ_2 control within-block and within-item cross-model dependence, respectively. We set $\kappa_1 = 0.5$ and $\kappa_2 = 0.1$. Let $S^{(\ell)}_{i,r} = \mathbb{1}\{Z_{i,\ell,r} \le \Phi^{-1}(p_i^{(\ell)})\}$ denote whether model ℓ answers item i correctly at replicate r. Each configuration is repeated 500 times.

Baseline Methods. We compare BB-EDGE with external methods and three internal baselines. External baselines: EMR performs uncorrected empirical-mean comparisons; CBTL (Wang et al., 2024b) combines context-dependent Bradley–Terry–Luce model with a Gaussian multiplier bootstrap; and t-Holm (Chandrahas, 2026) applies Holm's procedure to paired one-sided t-tests. SERPANT (Gu et al., 2026) is an online method for pairwise preferences based on e-processes under item-level independence. We include three internal baselines. BEB-Holm uses the same block observations but performs fixed-time empirical-Bernstein inference followed by Holm's procedure, providing a fixed-time baseline. BFH-e-Holm differs from BB-EDGE only in replacing the empirical-Bernstein penalty with a Hoeffding penalty, isolating the benefit of variance adaptation. CR-EDGE differs from BB-EDGE only in treating each complete replicate as a single block, isolating the benefit of block factorization. Full details of synthetic experiments are in Appendix G.2.

BB-EDGE Controls FWER_r with High Power_r. Figure 1 first examines error control under the fixed-benchmark target. BB-EDGE, BFH-e-Holm, and CR-EDGE maintain empirical anytime FWER_r control across the examined settings. The item-level SERPANT adaptation controls error in the i.i.d. item setting but can report false benchmark-average advantages under item heterogeneity or within-block dependence. This diagnostic illustrates the consequence of applying item-stream



(a) GSM8K (b) ACEBench

Figure 2: Certified comparison graphs at replicate $r = 20$ on GSM8K and ACEBench, respectively. An edge $A \to B$ indicates that model A is certified to outperform model B on the benchmark under the evaluation protocol. Pairs not connected by a directed path remain unresolved.

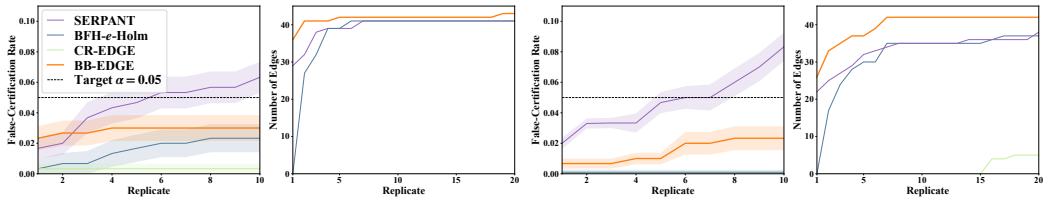
monitoring to the fixed-benchmark target. Additionally, three offline baselines EMR, CBTL, and t-Holm significantly exceed the target level. BEB-Holm is included as a fixed-time reference; its pointwise behavior does not constitute an anytime guarantee. Figure 1 also presents that BB-EDGE exhibits faster power growth than CR-EDGE and BFH-e-Holm in the examined settings. For example, under item-level heterogeneity, BB-EDGE reaches 90% power within 56 replicates, compared with 95 replicates for BFH-e-Holm. These comparisons isolate the effects of block factorization and empirical-Bernstein variance adaptation, respectively. The conditional log-growth expression in Section 5 identifies how blockwise prediction error enters evidence accumulation.

Ablation Study. We conduct a series of ablation studies under diverse experimental conditions. First, BB-EDGE continues to perform well with a larger model set ($L = 20$; Figure 6), across benchmark sizes $N \in \{50, 500\}$ (Figures 7–8), and under prespecified margins $\tau \in \{0.01, 0.03\}$ (Figures 9–10). Second, Figure 11 compares BB-EDGE with BB-EDGE-PI (plug-in stake), AsympCS-Bonf (non-factorized asymptotic confidence sequences), and SERPANT-R (SERPANT evaluated at completed replicates): BB-EDGE-PI performs similarly, whereas AsympCS-Bonf and, under within-block dependence, SERPANT-R exceed the target level. Third, Figure 12 varies the within-block correlation from 0.1 to 0.7 (panel (a)) and the block size from 1 to 10 (panel (b)); BB-EDGE controls FWER_r with the highest power, since the weight-proportional stakes require prespecified independent blocks but neither equal sizes nor equal correlations (Theorem 1).

6.2 Real Data Analysis

Models and Datasets. We evaluate BB-EDGE on four benchmarks: knowledge reasoning on MMLU-Pro (Wang et al., 2024a); mathematical reasoning on GSM8K (Cobbe et al., 2021); multi-turn dialogues on MT-Bench (Zheng et al., 2023); and multi-turn tool calling on ACEBench (Chen et al., 2025). We use accuracy for MMLU-Pro, exact match for GSM8K, AST-based accuracy for ACEBench, and the LLM-judge score for MT-Bench (Zheng et al., 2023) as the evaluation metrics. For MT-Bench, we use MiniMax-M3 (Lai et al., 2026) as the LLM judge and rescale its ratings to $[0, 1]$. We evaluate ten open-source LLMs: DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), Falcon-H1R-7B (Chaabane et al., 2026), Gemma-4-E4B-it (Team et al., 2026), Granite-4.1-8B (Padhi et al., 2024), Intern-S1-Mini (Bai et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), MiMo-7B-RL (Xiaomi et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Phi-4 (Abdin et al., 2024), and Qwen3.5-9B (Qwen Team, 2026a). Details of the datasets, prompt templates, block sizes, and other protocol configurations are provided in Appendix G.

Results. We present two representative examples—GSM8K and ACEBench—and defer the MMLU-Pro, MT-Bench, and API results to Appendix F due to space constraints. Figure 2 presents the confidence graphs for GSM8K and ACEBench, while Tables 3 and 5 report exact match and



(a) GSM8K (b) ACEBench

Figure 3: Same-model null diagnostic and certification efficiency on GSM8K and ACEBench after r replicates. Left: mean false-certification rate across 100 random 10-versus-10 splits of the same 20 observed replicates; Monte Carlo interval across random splits, conditional on the observed replicates. Right: number of distinct model pairs resolved by reachability in the current confidence graph after each complete replicate. At most $\binom{10}{2} = 45$ pairs can be resolved.

AST-based accuracy, respectively. On GSM8K, the confidence graph contains 10 nodes and is acyclic. Granite-4.1-8B occupies the top position and is certified above Phi-4, followed by Qwen3.5-9B and Intern-S1-Mini. Gemma-4-E4B-It, Llama-3.1-8B-Instruct, and MiMo-7B-RL remain mutually unresolved, while all three are certified above DeepSeek-R1-Distill-Qwen-7B. The graph then forms a strictly resolved chain from DeepSeek-R1-Distill-Qwen-7B through Mistral-7B-Instruct-v0.3 to Falcon-H1R-7B. Additionally, when we increase the prespecified margin from $\tau = 0$ to $\tau = 0.01$, Figure 16 shows that the edge from Granite-4.1-8B to Phi-4 disappears because their exact-match difference, 0.0071, does not exceed the required margin. On ACEBench, Gemma-4-E4B-It and Qwen3.5-9B have the highest mean scores, but BB-EDGE does not certify either model as better than the other. Both models are certified above Intern-S1-Mini and Phi-4, followed by a descending chain ending with Falcon-H1R-7B.

Figure 3 reports a same-model null diagnostic on GSM8K and ACEBench. For each model, we randomly split its 20 observed replicates into two groups of 10 and compare them as pseudo-models. Both groups have the same expected score on every item, so any directional certification is false. Across these splits, the SERPANT adaptation produces more false certifications than BB-EDGE. Because this diagnostic reuses the same 20 replicates, it does not estimate repeated-evaluation FWER or isolate item-level heterogeneity; Section 6.1 examines the latter through simulations. Meanwhile, Figure 3 also presents that BB-EDGE exhibits substantially faster gains under small evaluation budgets. Under the prespecified singleton-block analysis for GSM8K, BB-EDGE therefore uses 66.7% fewer queries than BFH-e-Holm (six replicates, 79,140 queries), the fastest compared baseline with an anytime FWER guarantee for the same benchmark-average hypotheses. Appendix F.3 reports that BB-EDGE certifies Top-k sets with fewer replicates than the baselines. For example, on ACEBench, BB-EDGE certifies the Top-4 set after 4 replicates, whereas BFH-e-Holm requires 17 and CR-EDGE issues no certificate within 20 replicates. Appendix F.4 reports rank intervals, which are narrower for BB-EDGE and indicate less uncertainty about model rankings.

7 Conclusion

This paper proposes BB-EDGE, an anytime-valid and efficient method for certifying model mean-performance advantages on fixed-benchmark leaderboards. Our approach begins by partitioning the benchmark according to the evaluation protocol. We then construct block-factorized empirical-Bernstein e-processes with weight-proportional stakes and apply direct e-Holm to certify mean-performance advantages. Furthermore, we provide a theoretical analysis demonstrating that BB-EDGE controls the probability of ever reporting a false edge across all model pairs and replicates at level α. Extensive experiments on both synthetic and real-world benchmarks demonstrate that BB-EDGE achieves anytime-valid FWER control with the highest statistical power.

Limitation. BB-EDGE's anytime-valid guarantees are conditional on the fixed benchmark and protocol, and do not extend to out-of-distribution tasks. Developing item-level anytime-valid inference that accommodates heterogeneous item effects and cross-item dependence is left for future work.

AI USE STATEMENT

We used a large language model solely for improving the fluency and readability of the manuscript. The model was not involved in research ideation, experimental design, data analysis, or result interpretation. All scientific contributions and substantive content were solely produced by the authors.

REPRODUCIBILITY STATEMENT

We provide comprehensive technical details in the paper, including the methodology (Section 4), theoretical proofs (Appendix D), and experimental setup (Appendix G). The code is also provided in the supplementary material.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Iheb Chaabane, Puneesh Khanna, Suhail Mohmad, Slim Frikha, Shi Hu, Abdalgader Abubaker, Reda Alami, Mikhail Lubinets, Mohamed El Amine Seddik, Hakim Hacid, et al. Falcon-h1r: Pushing the reasoning frontiers with a hybrid model for efficient test-time scaling. *arXiv preprint arXiv:2601.02346*, 2026.

Shreyas K Chandrahas. evalci: A python library for statistically rigorous comparison of language model evaluations. *arXiv preprint arXiv:2607.04429*, 2026.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel G Rodriguez. Prediction-powered ranking of large language models. *Advances in Neural Information Processing Systems*, pp. 113096–113133, 2024.

Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Yuefeng Huang, Xiangcheng Liu, Wang Xinzhi, and Wu Liu. ACEBench: A comprehensive evaluation of LLM tool usage. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12970–12998, 2025.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Abhishek Divekar and Anirban Majumder. Precise: Reducing the bias of llm evaluations using prediction-powered ranking estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 39929–39938, 2026.

Shaoshuai Du, Penghao Liang, Yixian Shen, Chuanqi Shi, Hang Zhang, and Lun Wang. On the stability of prompt ranking in large language model evaluation. *arXiv preprint arXiv:2606.24381*, 2026.

Jianqing Fan, Zhipeng Lou, Weichen Wang, and Mengxin Yu. Ranking inferences based on the top choice of multiway comparisons. *Journal of the American Statistical Association*, 120(549): 237–250, 2025.

Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015. doi: 10.1214/EJP.v20-3496.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, and Ion Stoica. Prompt-to-leaderboard: Prompt-adaptive LLM evaluations. In *Forty-second International Conference on Machine Learning*, 2025.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do repetitions matter? strengthening reliability in llm evaluations. *arXiv preprint arXiv:2509.24086*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Runzhe Gu, Wenguang Sun, Bowen Gang, and Xintao Xia. Anytime-valid inference for online ranking of large language models. In *Forty-third International Conference on Machine Learning*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Nika Haghtalab, Ariel D Procaccia, Han Shao, Serena Lutong Wang, and Kunhe Yang. Pluralistic leaderboards. *arXiv preprint arXiv:2606.02547*, 2026.

Mohsen Hariri, Amirhossein Samandar, Michael Hinczewski, and Vipin Chaudhary. Don't pass@ k: A bayesian framework for large language model evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 148539–148579, 2026.

Will Hartog and Lihua Lei. Family-wise error rate control with e-values. *arXiv preprint arXiv:2501.09015*, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models. *arXiv preprint arXiv:2607.17409*, 2026.

Jenny Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. In *International Conference on Learning Representations*, volume 2026, pp. 90685–90721, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, et al. Minimax sparse attention. *arXiv preprint arXiv:2606.13392*, 2026.

Jiachun Li, David Simchi-Levi, and Will Wei Sun. Low rank for rank: Uncertainty-aware task-specific llm ranking under sparse pairwise comparisons. *arXiv preprint arXiv:2605.29395*, 2026.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

Angel Rodrigo Avelar Menendez, Yufeng Liu, and Xiaowu Dai. Prompt-dependent ranking of large language models with uncertainty quantification. *arXiv preprint arXiv:2603.03336*, 2026.

Itay Nakash, George Kour, and Ateret Anaby Tavor. Efficient agent evaluation via diversity-guided user simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 1627–1648, 2026.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. *arXiv preprint arXiv:2606.08679*, 2026.

OpenAI. GPT-5.6 Luna Model. `https://developers.openai.com/api/docs/models/gpt-5.6-luna`, 2026. Accessed: 2026-09-09.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian. *arXiv preprint arXiv:2412.07724*, 2024.

Qiwei Peng, Yekun Chai, and Xuhong Li. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pp. 8383–8394, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL `https://qwen.ai/blog?id=qwen3.5`.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026b. URL `https://qwen.ai/blog?id=qwen3.8-flash-next`.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel Dsouza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. *Advances in Neural Information Processing Systems*, 38, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025.

Ganghua Wang, Zhaorun Chen, Bo Li, and Haifeng Xu. Cer-eval: Certifiable and cost-efficient evaluation framework for llms. *arXiv preprint arXiv:2505.03814*, 2025a.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024a.

Zebin Wang, Yi Han, Ethan X Fang, Lan Wang, and Junwei Lu. Confidence diagram of nonparametric ranking for uncertainty assessment in large language models evaluation. *arXiv preprint arXiv:2412.05506*, 2024b.

Zongqi Wang, Tianle Gu, Chen Gong, Xin Tian, Siqi Bao, and Yujiu Yang. From rankings to insights: Evaluation should shift focus from leaderboard to feedback. *arXiv preprint arXiv:2505.06698*, 2025b.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:39501–39521, 2024.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. *Advances in Neural Information Processing Systems*, 37:60110–60132, 2024.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026a.

Zhijian Zhou, Zesheng Ye, Zhaorun Chen, Bo Li, and Feng Liu. Celeus: Certifiable and efficient llm evaluation via e-processes. *arXiv preprint arXiv:2606.20820*, 2026b.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *International Conference on Learning Representations*, volume 2024, pp. 18091–18128, 2024.

APPENDIX

A ANYTIME-VALID INFERENCE FOR TOP-k CERTIFICATION

When the number of models is large, we may care only about identifying the Top-k models, motivating a more efficient procedure that tests only the relevant cross-set comparisons. Specifically, write $a \rightsquigarrow_r b$ if model b is reachable from model a in the current confidence graph $\mathcal{G}_r = (\mathcal{V}, \mathcal{E}_r)$. A set $\mathcal{T} \subseteq \mathcal{V}$ with $|\mathcal{T}| = k$ is certified as a Top-k set at replicate r if

$$a \rightsquigarrow_r b \qquad \text{for every } a \in \mathcal{T} \text{ and } b \in \mathcal{V} \setminus \mathcal{T}.$$

On the no-false-edge event in (1), every path from a to b has length $d \geq 1$ and satisfies $\theta_a^{\mathcal{B}} - \theta_b^{\mathcal{B}} > d\tau \geq \tau$. Hence Theorem 1 implies that, with probability at least $1 - \alpha$, every Top-k certificate issued from the full graph by replicate R satisfies $\min_{a \in \mathcal{T}} \theta_a^{\mathcal{B}} > \max_{b \notin \mathcal{T}} \theta_b^{\mathcal{B}} + \tau$. This conclusion permits searching the certified graph for a set; it does not permit selecting a new testing family from the same confirmatory data. No ordering is asserted among the models within $\mathcal{T}$ or within $\mathcal{V} \setminus \mathcal{T}$; the certificate concerns only the between-set comparison.

Our procedure consists of three steps: selecting a candidate set using independent pilot replicates, testing the corresponding cross-set directions, and certifying the candidate set using confirmatory evidence. Fix a pilot sample size $R_0 \geq 1$. For pilot run s, let $Z_s^{(0)} = (S_{i,s}^{(\ell,0)} : 1 \leq i \leq N,\ 1 \leq \ell \leq L)$ contain the scores of all models on all benchmark items, and write $\mathcal{D}_0 = (Z_1^{(0)}, \ldots, Z_{R_0}^{(0)})$. These pilot data are independent of the subsequent confirmatory runs $(Z_1, \ldots, Z_R)$ conditional on $(\mathcal{B}, \pi)$. The pre-inference information in Assumption 1 comprises $\mathcal{D}_0$ and all design choices fixed before confirmatory sampling.

Step 1: Candidate selection. Compute the pilot mean of model ℓ as

$$\widehat{\theta}_\ell^{(0)} = \frac{1}{NR_0} \sum_{s=1}^{R_0} \sum_{i=1}^{N} S_{i,s}^{(\ell,0)}. \tag{11}$$

Order the models by decreasing $\widehat{\theta}_\ell^{(0)}$, breaking equal pilot means by increasing model index in $\mathcal{V} = \{1, \ldots, L\}$. Let $\mathcal{T}_0$ contain the first k models in this order. The model indices are fixed before the pilot, and $\mathcal{T}_0$ remains fixed throughout the confirmatory phase. Pilot scores determine this candidate set but are not included as observations in the products in (7), which start from one at confirmatory run zero.

Step 2: Cross-set testing. During the confirmatory evaluation, we test whether each model in $\mathcal{T}_0$ outperforms every model outside $\mathcal{T}_0$ by more than the prespecified margin τ. Accordingly, we restrict the directional hypothesis family to

$$\mathcal{H}(\mathcal{T}_0) = \left\{ H_e^{(\tau)} \;\middle|\; e = (a \to b),\ a \in \mathcal{T}_0,\ b \in \mathcal{V} \setminus \mathcal{T}_0 \right\},$$

which contains the $k(L-k)$ cross-set directions required to certify $\mathcal{T}_0$.

Step 3: Top-k certification. After each completed replicate, we update $E_{e,r}^{\mathcal{C}}$ for every $H_e^{(\tau)} \in \mathcal{H}(\mathcal{T}_0)$ and apply direct e-Holm at level α to this family. We certify $\mathcal{T}_0$ as the Top-k set once all hypotheses in $\mathcal{H}(\mathcal{T}_0)$ are rejected; otherwise, the result remains unresolved.

The pilot determines only the targeted hypothesis family, while all certification evidence is obtained from the independent confirmatory replicates. Theorem 2 shows that restricting inference to this pilot-selected family preserves anytime FWER control and yields a valid Top-k certificate, whose proof is given in Appendix D.4.

Theorem 2. *Under Assumption 1, the pilot-targeted procedure with $1 \leq k < L$ satisfies*

$$\mathbb{P}\big(\exists r \in \{1, \ldots, R\},\ \exists (a \to b) \in \mathcal{E}_r \text{ with } \Delta_{a,b}^{\mathcal{B}} \leq \tau \,\big|\, \mathcal{D}_0, \mathcal{B}, \pi\big) \leq \alpha \quad \text{a.s.} \tag{12}$$

With conditional probability at least $1 - \alpha$ given $(\mathcal{D}_0, \mathcal{B}, \pi)$, every certificate issued by replicate R therefore satisfies

$$\min_{a \in \mathcal{T}_0} \theta_a^{\mathcal{B}} > \max_{b \in \mathcal{V} \setminus \mathcal{T}_0} \theta_b^{\mathcal{B}} + \tau.$$

Theorem 2 controls the false Top-k certification probability, denoted by FCP_r in Appendix G.3, for the fixed pilot candidate. The bound holds for almost every pilot outcome and after averaging over the pilot. An incorrect pilot candidate is allowed, but its probability of being certified by replicate R is at most α. This does not bound the error probability conditional on a certificate being issued. Equation (34) and the calculation following it make the distinction explicit. Full-graph certification instead uses Theorem 1 and the path implication above; it is not an application of the reduced-family pilot procedure.

Table 1 reports the Top-4 certification results under the i.i.d., heterogeneity, dependence, and combined settings. Both BB-EDGE variants maintain estimated $\mathrm{FWER}_R = \mathrm{FCP}_R = 0$ throughout. Although several baselines certify earlier, their FWER is substantially inflated; notably, SERPANT's FWER_R increases from 0.006 under i.i.d. items to 0.574–0.896 under heterogeneity or dependence. Relative to the strongest baseline, full-graph BB-EDGE reduces the mean CR by 44.7%–69.9%, while the pilot-targeted BB-EDGE reduces the mean confirmatory CR by 62.3%–83.7% by testing only the $k(L-k) = 24$ cross-set directions.

	$\lvert C_m\rvert = 1, h_{i,\ell} = 0$				$\lvert C_m\rvert = 5, h_{i,\ell} = 0$			
Method	$\mathrm{FWER}_R \downarrow$	$\mathrm{FCP}_R \downarrow$	$\mathrm{CertProb}_R \uparrow$	$\mathrm{CR} \downarrow$	$\mathrm{FWER}_R \downarrow$	$\mathrm{FCP}_R \downarrow$	$\mathrm{CertProb}_R \uparrow$	$\mathrm{CR} \downarrow$
EMR	$1.000_{\pm 0.000}$	$0.096_{\pm 0.295}$	$1.000_{\pm 0.000}$	$1.042_{\pm 0.211}$	$1.000_{\pm 0.000}$	$0.532_{\pm 0.499}$	$1.000_{\pm 0.000}$	$1.052_{\pm 0.240}$
CBTL	$0.796_{\pm 0.403}$	$0.006_{\pm 0.077}$	$1.000_{\pm 0.000}$	$8.222_{\pm 5.473}$	$0.850_{\pm 0.357}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$38.152_{\pm 14.999}$
t-Holm	$0.248_{\pm 0.432}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$9.764_{\pm 2.068}$	$0.236_{\pm 0.425}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$24.176_{\pm 7.386}$
SERPANT	$0.006_{\pm 0.077}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\underline{18.708}_{\pm 5.294}$	$0.386_{\pm 0.487}$	$0.028_{\pm 0.165}$	$1.000_{\pm 0.000}$	$18.802_{\pm 9.705}$
BEB-Holm	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$56.006_{\pm 3.841}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.888_{\pm 0.316}$	$84.968_{\pm 9.812}$
BFH-e-Holm	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.532_{\pm 0.499}$	$96.562_{\pm 4.842}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.266_{\pm 0.442}$	$97.426_{\pm 5.730}$
CR-EDGE	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$75.050_{\pm 3.783}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.944_{\pm 0.230}$	$83.868_{\pm 8.562}$
BB-EDGE	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$23.780_{\pm 2.752}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\underline{46.966}_{\pm 10.143}$
+Pilot	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\mathbf{12.252}_{\pm \mathbf{2.024}}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.920_{\pm 0.272}$	$\mathbf{32.010}_{\pm \mathbf{21.491}}$
	$\lvert C_m\rvert = 1, h_{i,\ell} = 0.2$				$\lvert C_m\rvert = 5, h_{i,\ell} = 0.2$			
Method	$\mathrm{FWER}_R \downarrow$	$\mathrm{FCP}_R \downarrow$	$\mathrm{CertProb}_R \uparrow$	$\mathrm{CR} \downarrow$	$\mathrm{FWER}_R \downarrow$	$\mathrm{FCP}_R \downarrow$	$\mathrm{CertProb}_R \uparrow$	$\mathrm{CR} \downarrow$
EMR	$1.000_{\pm 0.000}$	$0.042_{\pm 0.201}$	$1.000_{\pm 0.000}$	$1.030_{\pm 0.171}$	$1.000_{\pm 0.000}$	$0.516_{\pm 0.500}$	$1.000_{\pm 0.000}$	$1.078_{\pm 0.276}$
CBTL	$0.010_{\pm 0.100}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$100.000_{\pm 0.000}$	$0.196_{\pm 0.397}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$100.000_{\pm 0.000}$
t-Holm	$0.302_{\pm 0.460}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$8.996_{\pm 1.674}$	$0.212_{\pm 0.409}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$20.444_{\pm 6.027}$
SERPANT	$0.574_{\pm 0.495}$	$0.028_{\pm 0.165}$	$0.986_{\pm 0.118}$	$18.554_{\pm 11.314}$	$0.896_{\pm 0.306}$	$0.174_{\pm 0.379}$	$0.932_{\pm 0.252}$	$20.978_{\pm 23.038}$
BEB-Holm	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$79.880_{\pm 4.743}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.552_{\pm 0.498}$	$95.032_{\pm 6.652}$
BFH-e-Holm	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.578_{\pm 0.494}$	$96.524_{\pm 4.410}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.310_{\pm 0.463}$	$97.110_{\pm 5.787}$
CR-EDGE	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$74.624_{\pm 3.350}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.980_{\pm 0.140}$	$82.146_{\pm 7.666}$
BB-EDGE	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\underline{24.072}_{\pm 2.487}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\underline{42.464}_{\pm 8.497}$
+Pilot	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$\mathbf{13.010}_{\pm \mathbf{1.742}}$	$0.000_{\pm 0.000}$	$0.000_{\pm 0.000}$	$0.952_{\pm 0.214}$	$\mathbf{26.788}_{\pm \mathbf{17.537}}$

Table 1: Monte Carlo means and standard deviations of FWER_R, FCP_R, $\mathrm{CertProb}_R$, and the censored certification replicate CR for Top-4 certification. Formal definitions of the four metrics are defined in Appendix G.3. All experiments use $\tau = 0$, $R = 100$, $N = 100$, and $L = 10$. A repetition without certification by R is assigned $\mathrm{CR} = R$ ($R + R_0$ for +Pilot). Among methods with $\mathrm{FWER}_R \leq 0.05$, the smallest and second-smallest mean CR in each setting are shown in bold and underlined, respectively.

B Anytime-Valid Inference for Rank Intervals

Rank intervals quantify ranking uncertainty and avoid reporting a potentially misleading unique ranking when performance differences are not statistically significant. At each completed replicate r, BB-EDGE applies direct e-Holm to the pairwise e-processes to construct the current confidence graph $\mathcal{G}_r = (\mathcal{V}, \mathcal{E}_r)$. In the transitive closure of $\mathcal{G}_r$, let $\mathcal{A}_{\ell,r}$ and $\mathcal{D}_{\ell,r}$ denote the strict ancestors and descendants of model ℓ, respectively. The rank interval of model ℓ is defined as

$$\mathcal{I}_{\ell,r} = \left[1 + |\mathcal{A}_{\ell,r}|\,,\; L - |\mathcal{D}_{\ell,r}|\right]. \tag{13}$$

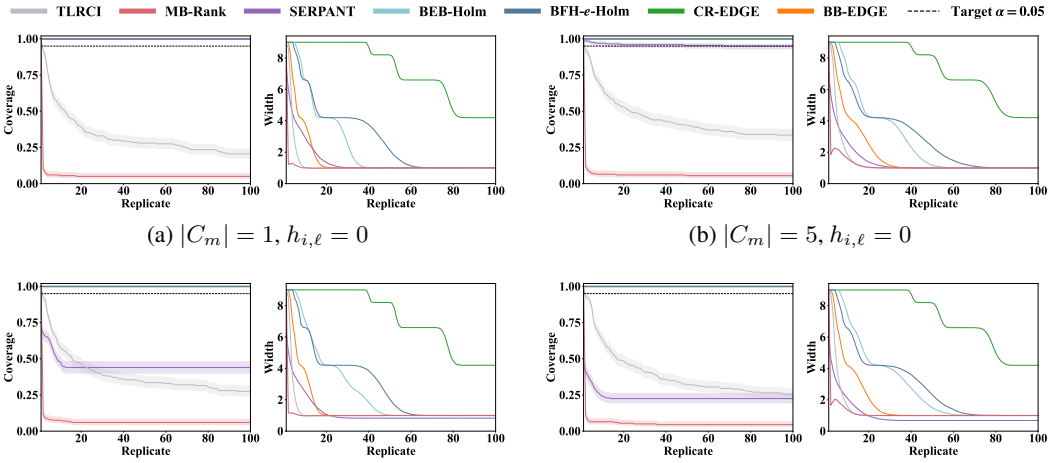
Its lower endpoint counts the models certified to outperform ℓ, while its upper endpoint excludes those certified to be outperformed by ℓ.

Corollary 1. *Define the true benchmark rank of model ℓ as $\mathrm{rank}_B(\ell) = 1 + \sum_{j \neq \ell} \mathbb{1}\{\theta_j^B > \theta_\ell^B\}$ where rank 1 corresponds to the highest benchmark mean. Suppose $\tau \geq 0$ and the conditions of Theorem 1 hold. Then the rank intervals in (13) satisfy*

$$\mathbb{P}(\forall\, 1 \leq r \leq R,\ \forall \ell \in \mathcal{V} : \mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r} \mid B, \pi) \geq 1 - \alpha.$$

The proof, including tied benchmark means, is given in Appendix D.5. If the assumptions hold for the entire infinite sequence, the same bound holds simultaneously for every $r \geq 1$. Thus, with probability at least $1-\alpha$, the intervals simultaneously cover the true benchmark ranks of all models at every completed replicate $1 \leq r \leq R$, conditional on (B, π).

We generate the synthetic data in Section 6.1, and the evaluation metrics are described in Section G.3. We additionally compare BB-EDGE with two rank-interval baselines, TLRCI (Neuhof & Benjamini, 2026) and MB-Rank (Li et al., 2026). The results show that, among the methods maintaining the desired anytime coverage level, BB-EDGE produces the narrowest rank intervals. For example, under item-level heterogeneity at replicate $r = 15$, BB-EDGE attains an anytime coverage of 1.00 with a mean rank-interval width of 1.633, reducing the width by 62.6% relative to the strongest valid baseline, BEB-Holm (4.369).



(a) $|C_m| = 1, h_{i,\ell} = 0$ (b) $|C_m| = 5, h_{i,\ell} = 0$

(c) $|C_m| = 1, h_{i,\ell} = 0.2$ (d) $|C_m| = 5, h_{i,\ell} = 0.2$

Figure 4: Empirical $\mathrm{Coverage}_r$ and Width_r (with 95% confidence intervals) versus replicate r. Panels (a)–(d) correspond to the i.i.d., dependence-only, heterogeneity-only, and combined dependence-and-heterogeneity settings, respectively. All settings use $\tau = 0$, $N = 100$, and $L = 10$. The dashed line marks the target level $1 - \alpha = 0.95$.

C Anytime-Valid Inference under Updated Benchmarks

C.1 Updated Benchmark Items

As benchmarks are updated with newly added items, the performance target changes. Valid inference must account for this change rather than treating evidence for an earlier version as evidence for the updated benchmark. We propose Epoch-reset BB-EDGE, an extension of BB-EDGE that supports anytime-valid inference under updated benchmarks. Let $\mathcal{K}_{v-1}$ be the sigma-field of all information available before the fresh replicates of epoch v, including its selected design. Epoch-reset BB-EDGE consists of three steps:

Step 1: Benchmark expansion. At update epoch v, let $\mathcal{A}^{(v)}$ denote the set of newly added items. We expand the preceding benchmark to obtain $B^{(v)} = B^{(v-1)} \cup \mathcal{A}^{(v)} = \{x_1, \ldots, x_{N_v}\}$ with $N_v = |B^{(v)}|$.

Step 2: Benchmark Freezing. After expansion, we freeze $B^{(v)}$ throughout epoch v. The frozen benchmark defines the version-specific performance estimand

$$\theta_\ell^{B^{(v)}} = \frac{1}{N_v} \sum_{i \in B^{(v)}} \mathbb{E}\left[S_{i,r}^{(\ell)} \,\middle|\, B^{(v)}, \pi \right], \qquad \Delta_{a,b}^{B^{(v)}} = \theta_a^{B^{(v)}} - \theta_b^{B^{(v)}}.$$

The finite benchmark, finite testing family $\mathcal{H}^{(v)}$, partition, finite stake grid, mixture weights, and budget $\alpha_v \in [0, 1)$ are $\mathcal{K}_{v-1}$-measurable and remain fixed within the epoch.

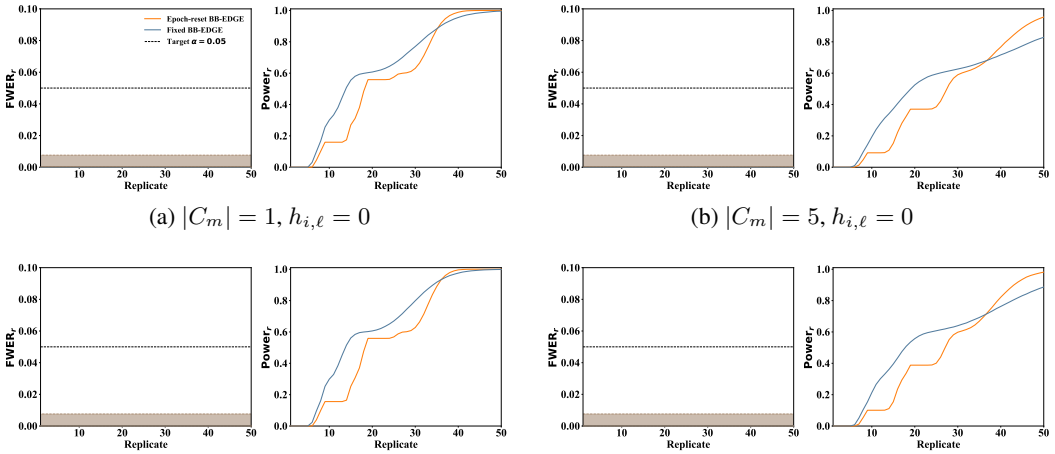
Step 3: e-Process Restart and Cross-Epoch Control. For each $H_e^{(v,\tau)} : \Delta_{a,b}^{B^{(v)}} \leq \tau$, we restart $E_{e,0}^{(v)} = 1$ using fresh replicates on $B^{(v)}$, preserving the supermartingale property under the new estimand, and apply direct e-Holm at level α_v with $\sum_v \alpha_v \leq \alpha$. Every grid component starts at one. Predictions are predictable with respect to $\mathcal{F}_r^{(v)} = \sigma(\mathcal{K}_{v-1}, Z_1^{(v)}, \ldots, Z_r^{(v)})$. Let $\mathcal{E}_r^{(v)}$ denote the edge sets after replicate r. Set $\mathcal{E}_r^{(v)} = \varnothing$ outside completed runs or when $\alpha_v = 0$; an epoch not entered has budget zero. For a finite run budget, the assumptions need to hold only through that budget. A guarantee for every $r \geq 1$ requires them for the entire sequence of fresh panels.

Because each epoch uses a newly initialized e-process constructed for its frozen benchmark, the anytime-validity of BB-EDGE applies separately within every epoch. Allocating a version-specific error budget α_v and requiring the cumulative budget to satisfy $\sum_{v \geq 1} \alpha_v \leq \alpha$ then extends this guarantee across benchmark versions. The resulting bound averages over the adaptive history and covers all reports actually issued. Each claim remains specific to the benchmark version on which it was tested. Theorem 3 formalizes the resulting FWER control, whose proof is given in Appendix D.6.

Theorem 3. *Suppose that, conditional on $\mathcal{K}_{v-1}$, the fresh panels in each epoch satisfy Assumption 1 under the fixed execution law for $(B^{(v)}, \pi)$ defining $\theta_\ell^{B^{(v)}}$. If $\sum_{v \geq 1} \alpha_v \leq \alpha$ almost surely, then Epoch-reset BB-EDGE satisfies*

$$\Pr\Big(\exists v \geq 1,\ \exists r \geq 1,\ \exists e = (a \to b) \in \mathcal{E}_r^{(v)} \text{ such that } \Delta_{a,b}^{B^{(v)}} \leq \tau\Big) \leq \alpha.$$

Following the data-generating process in Section 6.1, we expand the benchmark from 40 to 70 and 100 items at replicates $r = 10$ and $r = 20$, respectively. Epoch-reset BB-EDGE restarts at each update using $\alpha_v = \alpha/3$, whereas fixed BB-EDGE continues on the initial benchmark at level α. Figure 5 shows that both procedures maintain empirical FWER below 0.05 in all four settings. Epoch-reset BB-EDGE has lower power immediately after each reset because it discards previously accumulated evidence and uses a smaller per-version error budget; however, the expanded benchmarks provide more evidence per replicate, allowing it to catch up with and eventually outperform the fixed-benchmark reference, with the improvement being most pronounced under dependence and combined dependence and heterogeneity.



(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$

(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

Figure 5: Empirical FWER_r and Power_r versus replicate r. Panels (a)–(d) correspond to the i.i.d., dependence-only, heterogeneity-only, and combined dependence-and-heterogeneity settings, respectively. All settings use $\tau = 0$, $N = 100$, and $L = 10$. The dashed line marks the target level $\alpha = 0.05$.

C.2 Newly Added Models

BB-EDGE can accommodate newly arriving models without reevaluating the existing models. To allow an unknown number of future model additions, we prespecify an error-spending sequence

$(\alpha_v)_{v\geq 0}$ satisfying $\sum_{v\geq 0}\alpha_v \leq \alpha$, where α_0 is allocated to the initial leaderboard and α_v is allocated to the vth model-addition stage for $v \geq 1$. The procedure consists of two steps.

Step 1: Model addition and panel augmentation When the inclusion of a new model ℓ_v is determined independently of the inference panels, we set $\mathcal{V}^{(v)} = \mathcal{V}^{(v-1)} \cup \{\ell_v\}$. We evaluate ℓ_v on the same benchmark $\mathcal{B}$ and under the same protocol π. According to a pairing rule fixed before inference, its outcomes are combined with the stored outcomes to form $\widetilde{Z}_r^{(v)} = \left(Z_r^{(v-1)}, \{S_{i,r}^{(\ell_v)} : i \in \mathcal{B}\}\right)$, where $r = 1, \ldots, R^{(v)}$ and $R^{(v)}$ is the number of available paired replicates at stage v. The augmented panels are replayed in replicate order, with each predictable quantity at replicate r depending only on the preceding augmented panels.

Step 2: New-direction inference We initialize e-processes only for the newly introduced directional family $\mathcal{H}_{\text{new}}^{(v)} = \left\{H_{\ell_v\to\ell}^{(\tau)}, H_{\ell\to\ell_v}^{(\tau)} : \ell \in \mathcal{V}^{(v-1)}\right\}$ and apply direct e-Holm at level α_v. Let $\mathcal{E}_{r,\text{new}}^{(v)}$ denote the resulting edge set at replicate $r \leq R^{(v)}$. Previously certified edges are retained; their errors are already accounted for by the budgets allocated to the earlier stages.

Corollary 2 establishes FWER control for the combined graph across the initial leaderboard and all subsequent model additions.

Corollary 2. *Let $\mathcal{E}_{r,\text{new}}^{(0)} = \mathcal{E}_r$ denote the edge set of the initial leaderboard, constructed over $R^{(0)}$ replicates at level α_0. Suppose that the initial panels $Z_1, \ldots, Z_{R^{(0)}}$ satisfy Assumption 1. For each $v \geq 1$, suppose that model inclusion, the pairing rule, and all stage-v design choices are fixed independently of the augmented inference panels conditional on $(\mathcal{B}, \pi)$. Suppose further that $\widetilde{Z}_1^{(v)}, \ldots, \widetilde{Z}_{R^{(v)}}^{(v)}$ satisfy Assumption 1 and are processed through the non-anticipating replay described above. If $(\alpha_v)_{v\geq 0}$ is a prespecified error-spending sequence satisfying $\sum_{v\geq 0}\alpha_v \leq \alpha$, then*

$$\mathbb{P}\Big(\exists v \geq 0,\ \exists r \leq R^{(v)},\ \exists e \in \mathcal{E}_{r,\text{new}}^{(v)} \text{ such that } \Delta_e^{\mathcal{B}} \leq \tau \,\Big|\, \mathcal{B}, \pi\Big) \leq \alpha.$$

On the complementary event, all path implications, simultaneous rank intervals, and Top-k certificates derived from the combined graph are valid.

The current experiments allocate the full error budget to the initial leaderboard; the staged allocation above applies prospectively when future model additions are anticipated. The independence condition is satisfied when model additions are driven by external events, such as new model releases, rather than selected using the certified graph itself.

D PROOFS

The fixed-benchmark bounds in Theorem 1 and Corollary 1 condition on $(\mathcal{B}, \pi)$; Theorem 2 additionally conditions on the pilot data $\mathcal{D}_0$. Theorem 3 first controls each epoch conditional on its past and then averages over that past. All evidence processes are monitored at completed-replicate boundaries, using the common filtration containing every model's observations.

D.1 PROOF OF PROPOSITION 1

Proof. Take any $d \in \mathbb{R}^M$ with $w^\top d = 0$. If $d \neq 0$, choose

$$0 < \varepsilon < \frac{\min\{\mu_0, 1-\mu_0\}}{\max_m |d_m|}.$$

Then both vectors $\nu_+ = \mu_0\mathbf{1} + \varepsilon d$ and $\nu_- = \mu_0\mathbf{1} - \varepsilon d$ belong to $[0,1]^M$. Since $w^\top\mathbf{1} = 1$, they satisfy $w^\top\nu_+ = w^\top\nu_- = \mu_0$. Applying (9) to ν_+ and ν_- gives

$$\eta^\top(\nu_+ - \mu_0\mathbf{1}) = \varepsilon\eta^\top d \leq 0, \qquad \eta^\top(\nu_- - \mu_0\mathbf{1}) = -\varepsilon\eta^\top d \leq 0.$$

Since $\varepsilon > 0$, these inequalities imply $\eta^\top d = 0$. This equality also holds for $d = 0$. We have therefore proved

$$w^\top d = 0 \quad \Longrightarrow \quad \eta^\top d = 0, \qquad d \in \mathbb{R}^M. \tag{14}$$

Set $c = (\eta^\top w)/(w^\top w)$ and $h = \eta - cw$. The denominator is positive because every $w_m > 0$, and

$$w^\top h = w^\top \eta - c w^\top w = w^\top \eta - \eta^\top w = 0.$$

Equation (14), applied with $d = h$, now yields

$$0 = \eta^\top h = (cw + h)^\top h = \|h\|_2^2.$$

Consequently $h = 0$ and $\eta = cw$. The nonnegativity of η and positivity of w imply $c \geq 0$. Conversely, if $\eta = cw$ with $c \geq 0$, then for every admissible ν,

$$\eta^\top(\nu - \mu_0 \mathbf{1}) = c(w^\top \nu - \mu_0) \leq 0.$$

This proves both directions, including $c = 0$ and $M = 1$. The necessity concerns precisely the nonpositive linear drift condition in the proposition. □

D.2 The conditional empirical-Bernstein inequality

The following bounded-variable inequality is the ingredient needed for the block factors, which is the predictable empirical-Bernstein construction of Waudby-Smith & Ramdas (2024). We below give its scalar argument, following Fan et al. (2015), to specify the normalization used here.

Lemma 1. *Let $\mathcal{A}$ be a sigma-field, let $Y \in [0,1]$, and write $\nu = \mathbb{E}[Y \mid \mathcal{A}]$. If $p \in [0,1]$ and $u \in [0,1)$ are $\mathcal{A}$-measurable, then*

$$\mathbb{E}\big[\exp\{u(Y-\nu) - \psi_E(u)(Y-p)^2\}\big|\mathcal{A}\big] \leq 1, \qquad \psi_E(u) = -\log(1-u) - u. \tag{15}$$

Proof. First we prove the deterministic inequality

$$\exp\{ux - \psi_E(u)x^2\} \leq 1 + ux, \qquad 0 \leq u < 1, \quad -1 \leq x \leq 1, \tag{16}$$

which is an equality if $u = 0$ or $x = 0$. For $x > 0$, let $z = ux$. The function $f(z) = \log(1+z) - z + z^2/2$ satisfies $f(0) = 0$ and $f'(z) = z^2/(1+z) \geq 0$ for $z \geq 0$. Since $\psi_E(u) = \sum_{j=2}^{\infty} u^j/j \geq u^2/2$, this gives

$$\log(1+ux) \geq ux - u^2x^2/2 \geq ux - \psi_E(u)x^2.$$

For $x = -v$ with $0 < v \leq 1$, absolute convergence and nonnegativity of the series terms give

$$\psi_E(uv) = \sum_{j=2}^{\infty} \frac{u^j v^j}{j} \leq v^2 \sum_{j=2}^{\infty} \frac{u^j}{j} = v^2 \psi_E(u).$$

Hence $\log(1-uv) = -uv - \psi_E(uv) \geq -uv - v^2\psi_E(u)$. As $1 + ux \geq 1 - u > 0$, exponentiation proves (16) in both cases. To apply (16) with $x = Y - p$, write

$$u(Y-\nu) = u(Y-p) + u(p-\nu).$$

Here $u(p-\nu)$ is $\mathcal{A}$-measurable because u, p, and $\nu = \mathbb{E}[Y \mid \mathcal{A}]$ are $\mathcal{A}$-measurable. Thus

$$\begin{aligned}\exp\{u(Y-\nu) - \psi_E(u)(Y-p)^2\} &= \exp\{u(p-\nu)\}\exp\{u(Y-p) - \psi_E(u)(Y-p)^2\} \\ &\leq \exp\{u(p-\nu)\}\{1 + u(Y-p)\}.\end{aligned} \tag{17}$$

In particular, $\exp\{u(p-\nu)\}$ and u can be taken outside a conditional expectation given $\mathcal{A}$. The exponential on the left satisfies

$$0 < \exp\{u(Y-\nu) - \psi_E(u)(Y-p)^2\} \leq \exp\{u(Y-\nu)\} \leq \exp(1),$$

because $\psi_E(u) \geq 0$, $Y - \nu \leq 1$, and $u < 1$. It is therefore integrable. Taking conditional expectations in (17) gives

$$\begin{aligned}&\mathbb{E}\big[\exp\{u(Y-\nu) - \psi_E(u)(Y-p)^2\}\big|\mathcal{A}\big] \\ &\quad\leq \exp\{u(p-\nu)\}\{1 + u(\nu-p)\} = \exp(z)(1-z), \qquad z = u(p-\nu) < 1.\end{aligned}$$

The inequality $\log t \leq t - 1$ for $t > 0$, applied to $t = 1 - z$, gives $z + \log(1-z) \leq 0$. Thus $\exp(z)(1-z) \leq 1$, as required. The argument does not require $p = \nu$. □

D.3 Proof of Theorem 1

Proof. Recall that the inference filtration is

$$\mathcal{F}_r = \sigma(\mathcal{B}, \pi, \mathcal{Q}, Z_1, \ldots, Z_r), \qquad r = 0, 1, \ldots, R,$$

with the convention

$$\mathcal{F}_0 = \sigma(\mathcal{B}, \pi, \mathcal{Q}).$$

Here, $\mathcal{Q}$ collects all information determined before the confirmatory inference runs, including the directional testing family, the benchmark partition, the stake grid, the mixture weights, and any independent pilot data used to select these quantities. Thus, $\mathcal{F}_{r-1}$ represents all information available immediately before the rth complete replicate Z_r is observed. We first work conditionally on $\mathcal{F}_0$. Under this conditional law, the benchmark, protocol, and all design quantities contained in $\mathcal{Q}$ are treated as fixed, while the confirmatory panels $Z_1, \ldots, Z_R$ remain random. We prove the supermartingale property and the corresponding FWER bound conditional on $\mathcal{F}_0$. The resulting conditional bound holds almost surely with respect to the random pre-inference design information. We then use the tower property to average over this initial randomness and obtain the stated guarantee conditional on $(\mathcal{B}, \pi)$.

Fix $e = (a \to b)$ and $1 \le r \le R$. We first prove

$$\sum_{m=1}^{M} w_m \mathbb{E}[Y_{e,m,r} \mid \mathcal{F}_{r-1}] = \frac{1 + \Delta^{\mathcal{B}}_{a,b}}{2}.$$

Write $\mathcal{L}(U \mid \mathcal{A})$ for the conditional distribution of a random variable U given a sigma-field $\mathcal{A}$. Since $\mathcal{F}_{r-1} = \sigma(\mathcal{F}_0, Z_1, \ldots, Z_{r-1})$, Assumption 1(1) and (3) give, respectively,

$$\mathcal{L}(Z_r \mid \mathcal{F}_{r-1}) = \mathcal{L}(Z_r \mid \mathcal{F}_0) = \mathcal{L}(Z_r \mid \mathcal{B}, \pi), \tag{18}$$

where the first equality uses independence across runs conditional on $\mathcal{F}_0$; the second uses independence of $\mathcal{Q}$ and the confirmatory panels conditional on $(\mathcal{B}, \pi)$. Integrating any bounded measurable panel function h against these equal distributions yields

$$\mathbb{E}[h(Z_r) \mid \mathcal{F}_{r-1}] = \mathbb{E}[h(Z_r) \mid \mathcal{F}_0] = \mathbb{E}[h(Z_r) \mid \mathcal{B}, \pi]. \tag{19}$$

Identical distribution across runs in Assumption 1(1) makes the last expectation independent of r.

Recall that $w_m = |C_m|/N$. Section 5, immediately before (4), defines $\nu_{e,m,r} = \mathbb{E}[Y_{e,m,r} \mid \mathcal{F}_{r-1}]$. For each realized partition, (3) holds, and each w_m is $\mathcal{F}_0$-measurable, hence $\mathcal{F}_{r-1}$-measurable. Linearity of conditional expectation and (19) give

$$\begin{aligned}\sum_m w_m \nu_{e,m,r} &= \mathbb{E}\left[\sum_m w_m Y_{e,m,r} \middle| \mathcal{F}_{r-1}\right] = \mathbb{E}\left[\frac{1}{N}\sum_i X^e_{i,r} \middle| \mathcal{F}_{r-1}\right] \\ &= \frac{1}{2} + \frac{1}{2}\mathbb{E}\left[\frac{1}{N}\sum_i (S^{(a)}_{i,r} - S^{(b)}_{i,r}) \middle| \mathcal{F}_{r-1}\right] = \frac{1 + \Delta^{\mathcal{B}}_{a,b}}{2}.\end{aligned} \tag{20}$$

Under $H^{(\tau)}_e$, $\Delta^{\mathcal{B}}_{a,b} \le \tau$, so $(1 + \Delta^{\mathcal{B}}_{a,b})/2 \le (1 + \tau)/2 = \mu_0$.

For a grid point λ_g, each stake $u_m = \lambda_{g,m} = \lambda_g w_m / w_\star$ lies in $[0, 1)$ and is fixed before the run. Decompose the factor as

$$\begin{aligned}F^{\mathcal{C}}_{e,r}(\lambda_g) &= \exp\left\{\sum_m u_m(\nu_{e,m,r} - \mu_0)\right\} \prod_m Q_{m,r}, \\ Q_{m,r} &= \exp\left\{u_m(Y_{e,m,r} - \nu_{e,m,r}) - \psi_E(u_m)(Y_{e,m,r} - \widehat{Y}_{e,m,r})^2\right\}.\end{aligned} \tag{21}$$

Recall the block vectors in Assumption 1(2),

$$V_{m,r} = (W_{i,r} : i \in C_m), \qquad W_{i,r} = (S^{(1)}_{i,r}, \ldots, S^{(L)}_{i,r}).$$

The observation entering $Q_{m,r}$ is

$$Y_{e,m,r} = \frac{1}{|C_m|} \sum_{i \in C_m} \frac{S^{(a)}_{i,r} - S^{(b)}_{i,r} + 1}{2},$$

which is a function of $V_{m,r}$. In contrast, u_m is $\mathcal{F}_0$-measurable, $\widehat{Y}_{e,m,r}$ is $\mathcal{F}_{r-1}$-measurable by Assumption 1(4), and $\nu_{e,m,r} = \mathbb{E}[Y_{e,m,r} \mid \mathcal{F}_{r-1}]$ is $\mathcal{F}_{r-1}$-measurable by the definition of conditional expectation. Thus, after conditioning on $\mathcal{F}_{r-1}$, the parameters u_m, $\widehat{Y}_{e,m,r}$, and $\nu_{e,m,r}$ in $Q_{m,r}$ are fixed. Only $Y_{e,m,r}$ depends on the new run. We do not assume that $Y_{e,m,r}$ is $\mathcal{F}_{r-1}$-measurable. Mutual conditional independence of $V_{1,r}, \ldots, V_{M,r}$ implies mutual conditional independence of $Q_{1,r}, \ldots, Q_{M,r}$. Applying Lemma 1 with $\mathcal{A} = \mathcal{F}_{r-1}$, $Y = Y_{e,m,r}$, $p = \widehat{Y}_{e,m,r}$, and $u = u_m$ gives $\mathbb{E}[Q_{m,r} \mid \mathcal{F}_{r-1}] \le 1$ for each m. Also,

$$\sum_m u_m(\nu_{e,m,r} - \mu_0) = \frac{\lambda_g}{w_\star}\left(\sum_m w_m \nu_{e,m,r} - \mu_0\right) \le 0$$

under the null. Taking conditional expectations in (21) consequently yields

$$\mathbb{E}[F^{\mathcal{C}}_{e,r}(\lambda_g) \mid \mathcal{F}_{r-1}] = \exp\left\{\sum_m u_m(\nu_{e,m,r} - \mu_0)\right\} \prod_m \mathbb{E}[Q_{m,r} \mid \mathcal{F}_{r-1}] \le 1. \tag{22}$$

For each g and m, $0 \le u_m < 1$ implies $0 \le \psi_E(u_m) < \infty$. Since $|Y_{e,m,r} - \mu_0| \le 1$ and $(Y_{e,m,r} - \widehat{Y}_{e,m,r})^2 \le 1$, the exponent in (6) satisfies

$$-\sum_{m=1}^{M}\{u_m + \psi_E(u_m)\} \le \log F^{\mathcal{C}}_{e,r}(\lambda_g) \le \sum_{m=1}^{M} u_m \le M \le N.$$

Both bounds are finite for each initial design. Exponentiating gives $0 < F^{\mathcal{C}}_{e,r}(\lambda_g) \le \exp(N)$. The component product and the mixture in (7) consequently obey, for every finite r,

$$\begin{aligned} 0 < E^{\mathcal{C}}_{e,r}(\lambda_g) &= \prod_{s=1}^{r} F^{\mathcal{C}}_{e,s}(\lambda_g) \le \exp(rN), \\ 0 < E^{\mathcal{C}}_{e,r} &= \sum_{g=1}^{G} \rho_g E^{\mathcal{C}}_{e,r}(\lambda_g) \le \exp(rN) \sum_{g=1}^{G} \rho_g = \exp(rN). \end{aligned} \tag{23}$$

Here $\rho_1, \ldots, \rho_G$ are the mixture coefficients, not the benchmark block weights w_m. Since $\rho_g \ge 0$ and $\sum_g \rho_g = 1$, at least one coefficient is positive. Since N is finite and fixed given the benchmark, (23) implies

$$\mathbb{E}[|E^{\mathcal{C}}_{e,r}(\lambda_g)| \mid \mathcal{F}_0] \le \exp(rN) < \infty, \qquad \mathbb{E}[|E^{\mathcal{C}}_{e,r}| \mid \mathcal{F}_0] \le \exp(rN) < \infty.$$

The same bounds hold conditional on $(\mathcal{B}, \pi)$ by the tower property, which proves the required finite-time integrability under the fixed-benchmark law.

The recursion of $E^{\mathcal{C}}_{e,r}(\lambda_g) = E^{\mathcal{C}}_{e,r-1}(\lambda_g) F^{\mathcal{C}}_{e,r}(\lambda_g)$ and the $\mathcal{F}_{r-1}$-measurability of the preceding product yield

$$\begin{aligned} \mathbb{E}[E^{\mathcal{C}}_{e,r}(\lambda_g) \mid \mathcal{F}_{r-1}] &= E^{\mathcal{C}}_{e,r-1}(\lambda_g)\mathbb{E}[F^{\mathcal{C}}_{e,r}(\lambda_g) \mid \mathcal{F}_{r-1}] \\ &\le E^{\mathcal{C}}_{e,r-1}(\lambda_g). \end{aligned}$$

Assumption 1(3) makes each ρ_g $\mathcal{F}_0$-measurable and keeps it unchanged across runs. As $\mathcal{F}_0 \subseteq \mathcal{F}_{r-1}$ and the sum has $G < \infty$ terms,

$$\begin{aligned} \mathbb{E}[E^{\mathcal{C}}_{e,r} \mid \mathcal{F}_{r-1}] &= \sum_{g=1}^{G} \rho_g \mathbb{E}[E^{\mathcal{C}}_{e,r}(\lambda_g) \mid \mathcal{F}_{r-1}] \\ &\le \sum_{g=1}^{G} \rho_g E^{\mathcal{C}}_{e,r-1}(\lambda_g) = E^{\mathcal{C}}_{e,r-1}. \end{aligned}$$

Moreover, $E^{\mathcal{C}}_{e,0} = \sum_{g=1}^{G} \rho_g = 1$. This proves the directional supermartingale assertion. Zero stakes and zero mixture weights cause no difficulty.

For the FWER claim (10), define

$$\mathcal{N}_0 = \{e = (a \to b) \in \overrightarrow{\mathcal{P}} : \Delta^{\mathcal{B}}_{a,b} \le \tau\}, \qquad F_R = \bigcup_{r=1}^{R} \{\mathcal{E}_r \cap \mathcal{N}_0 \neq \varnothing\}.$$

We will show $\mathbb{P}(F_R \mid \mathcal{F}_0) \le \alpha$ and then average over the initial design to obtain $\mathbb{P}(F_R \mid \mathcal{B}, \pi) \le \alpha$. Membership in $\mathcal{N}_0$ depends only on the tested directional family $\overrightarrow{\mathcal{P}}$ and on the benchmark means $\theta^{\mathcal{B}}_a$ and $\theta^{\mathcal{B}}_b$. The tested directional family is $\mathcal{F}_0$-measurable by Assumption 1(3), while the benchmark means are determined by $(\mathcal{B}, \pi)$. Therefore, for each direction e, $\mathbb{1}\{e \in \mathcal{N}_0\}$ is $\mathcal{F}_0$-measurable, and so is $h_0 = |\mathcal{N}_0|$. If $h_0 = 0$, F_R is empty. For an initial design with $h_0 \ge 1$, write $t = 1/\alpha$ and define

$$A_r = \frac{1}{h_0} \sum_{e \in \mathcal{N}_0} E^{\mathcal{C}}_{e,r}.$$

All directional processes use the same filtration $\mathcal{F}_r$, so

$$\mathbb{E}[A_r \mid \mathcal{F}_{r-1}] = \frac{1}{h_0} \sum_{e \in \mathcal{N}_0} \mathbb{E}[E^{\mathcal{C}}_{e,r} \mid \mathcal{F}_{r-1}] \le \frac{1}{h_0} \sum_{e \in \mathcal{N}_0} E^{\mathcal{C}}_{e,r-1} = A_{r-1}.$$

Hence A_r is a nonnegative supermartingale with $A_0 = 1$. The displayed equality is linearity of conditional expectation for a finite sum; the directional processes may be dependent. The cutoff in (8) is

$$\mathcal{J}_r = \{e \in \overrightarrow{\mathcal{P}} : E^{\mathcal{C}}_{e,r} < t\}, \qquad c_r = t + \sum_{e \in \mathcal{J}_r} (t - E^{\mathcal{C}}_{e,r}).$$

Suppose some $e_\star \in \mathcal{N}_0$ is rejected at time r. Since $c_r \ge t$, this direction does not belong to $\mathcal{J}_r$. For every $e \in \mathcal{J}_r$, the definition of $\mathcal{J}_r$ gives $t - E^{\mathcal{C}}_{e,r} > 0$. Splitting the cutoff sum therefore gives

$$\begin{aligned} E^{\mathcal{C}}_{e_\star,r} \ge c_r &= t + \sum_{e \in \mathcal{N}_0 \cap \mathcal{J}_r} (t - E^{\mathcal{C}}_{e,r}) + \sum_{e \in \mathcal{J}_r \setminus \mathcal{N}_0} (t - E^{\mathcal{C}}_{e,r}) \\ &\ge t + \sum_{e \in \mathcal{N}_0 \cap \mathcal{J}_r} (t - E^{\mathcal{C}}_{e,r}). \end{aligned}$$

Add the e-values in $\mathcal{N}_0 \cap \mathcal{J}_r$ to both sides. Each remaining true-null direction other than $e_\star$ is outside $\mathcal{J}_r$ and thus has e-value at least t. Adding those e-values gives

$$\sum_{e \in \mathcal{N}_0} E^{\mathcal{C}}_{e,r} \ge (1 + |\mathcal{N}_0 \cap \mathcal{J}_r|)t + (h_0 - |\mathcal{N}_0 \cap \mathcal{J}_r| - 1)t = h_0 t. \tag{24}$$

Dividing (24) by h_0 proves the pointwise event inclusion

$$\{\mathcal{E}_r \cap \mathcal{N}_0 \neq \varnothing\} \subseteq \{A_r \ge 1/\alpha\}, \qquad 1 \le r \le R. \tag{25}$$

This is the deterministic implication underlying the direct e-Holm rule of Hartog & Lei (2025, Theorem 4.2).

For completeness, the needed maximal inequality follows by a finite stopping argument. Let $T = \inf\{1 \le r \le R : A_r \ge 1/\alpha\}$, with $T = \infty$ if there is no crossing. For any finite $n \le R$,

$$A_{T \wedge n} = A_0 + \sum_{s=1}^{n} \mathbb{1}\{T \ge s\}(A_s - A_{s-1}).$$

For $1 \le s \le n$, the indicator multiplying $A_s - A_{s-1}$ is

$$\mathbb{1}\{T \ge s\} = \prod_{q=1}^{s-1} \mathbb{1}\{A_q < 1/\alpha\} \in \{0, 1\},$$

where the empty product for $s = 1$ equals one. Since each A_q is $\mathcal{F}_q$-measurable and $\mathcal{F}_q \subseteq \mathcal{F}_{s-1}$ for $q < s$, $\mathbb{1}\{T \ge s\}$ is $\mathcal{F}_{s-1}$-measurable. Applying the tower property to the finite stopped sum gives

$$\begin{aligned} \mathbb{E}[A_{T \wedge n} \mid \mathcal{F}_0] &= 1 + \sum_{s=1}^{n} \mathbb{E}[\mathbb{1}\{T \ge s\} \mathbb{E}[A_s - A_{s-1} \mid \mathcal{F}_{s-1}] | \mathcal{F}_0] \\ &\le 1, \end{aligned}$$

because the indicator is nonnegative and each inner conditional expectation is nonpositive. Nonnegativity of $A_{T\wedge n}$ and the definition of T imply

$$A_{T\wedge n} \geq \alpha^{-1}\mathbb{1}\{T \leq n\}, \qquad \mathbb{P}(T \leq n \mid \mathcal{F}_0) \leq \alpha.$$

By (25), $F_R \subseteq \{T \leq R\}$ when $h_0 > 0$; when $h_0 = 0$, F_R is empty. Therefore $\mathbb{P}(F_R \mid \mathcal{F}_0) \leq \alpha$ almost surely in both cases. Since $\sigma(\mathcal{B}, \pi) \subseteq \mathcal{F}_0$, the tower property gives

$$\mathbb{P}(F_R \mid \mathcal{B}, \pi) = \mathbb{E}[\mathbb{P}(F_R \mid \mathcal{F}_0) \mid \mathcal{B}, \pi] \leq \alpha.$$

This is exactly (10), which averages over any independent initial design selection rather than changing the target. If the assumptions hold for every run, apply this bound for each positive integer n. The events $\{\max_{1\leq r\leq n} A_r \geq 1/\alpha\}$ increase to the event of any finite-time crossing, giving the same bound over all times. This is Ville's argument in the nonnegative-supermartingale framework (Howard et al., 2020).

For the path inequalities, acyclicity, rank coverage, and top-set certificates in Theorem 1, consider

$$\Omega_R = F_R^c = \bigcap_{r=1}^{R} \{\mathcal{E}_r \cap \mathcal{N}_0 = \varnothing\}.$$

We have proved $\mathbb{P}(\Omega_R \mid \mathcal{B}, \pi) \geq 1 - \alpha$. Fix an outcome in Ω_R and a completed run $1 \leq r \leq R$. By the definitions of Ω_R and $\mathcal{N}_0$, each $(a \to b) \in \mathcal{E}_r$ satisfies $\theta_a^{\mathcal{B}} - \theta_b^{\mathcal{B}} > \tau$. For a path $a = v_0 \to v_1 \to \cdots \to v_d = b$ of $d \geq 1$ such edges, summing the d inequalities gives

$$\theta_a^{\mathcal{B}} - \theta_b^{\mathcal{B}} = \sum_{j=0}^{d-1} (\theta_{v_j}^{\mathcal{B}} - \theta_{v_{j+1}}^{\mathcal{B}}) > d\tau. \tag{26}$$

For a directed cycle, $v_d = v_0$, and (26) would give $0 > d\tau \geq 0$, a contradiction. Thus $(\mathcal{V}, \mathcal{E}_r)$ is acyclic. For any distinct vertices linked by a path, $d \geq 1$ and $\tau \geq 0$ imply $\Delta_{a,b}^{\mathcal{B}} > d\tau \geq \tau$, so the derived edge is a true margin claim as well.

For a model ℓ, let $B_\ell = \{j \neq \ell : \theta_j^{\mathcal{B}} > \theta_\ell^{\mathcal{B}}\}$. Recall that $A_{\ell,r}$ contains the vertices with a directed path to ℓ, and $D_{\ell,r}$ contains the vertices reached by a directed path from ℓ, in both cases excluding ℓ itself. Equation (26) gives

$$A_{\ell,r} \subseteq B_\ell, \qquad D_{\ell,r} \subseteq \{j \neq \ell : \theta_j^{\mathcal{B}} < \theta_\ell^{\mathcal{B}}\}.$$

The second set is disjoint from B_ℓ, so $|A_{\ell,r}| \leq |B_\ell|$ and $|B_\ell| + |D_{\ell,r}| \leq L - 1$. With the true rank defined by $\mathrm{rank}_B(\ell) = 1 + |B_\ell|$, these inequalities give

$$1 + |A_{\ell,r}| \leq \mathrm{rank}_B(\ell) \leq L - |D_{\ell,r}|, \qquad \mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r}, \tag{27}$$

where $\mathcal{I}_{\ell,r} = [1 + |A_{\ell,r}|, L - |D_{\ell,r}|]$ is the interval in (13). This is the rank definition used in Corollary 1, which counts strictly better means and permits ties.

Finally, if a nonempty proper set $\mathcal{T}$ is certified, every $a \in \mathcal{T}$ reaches every $b \notin \mathcal{T}$. Applying (26) to every such pair and using $\tau \geq 0$ gives

$$\min_{a\in\mathcal{T}} \theta_a^{\mathcal{B}} > \max_{b\notin\mathcal{T}} \theta_b^{\mathcal{B}}.$$

Thus $\mathcal{T}$ is the top set of its stated size. Ties within either set are allowed. A tie across their boundary cannot satisfy this certificate on Ω_R. All these implications hold pointwise on Ω_R for every $1 \leq r \leq R$, model, and reported set, so no additional multiplicity correction is needed for these graph consequences. □

D.4 Proof of Theorem 2

Proof. In Section A, the pilot data are $\mathcal{D}_0 = (Z_1^{(0)}, \ldots, Z_{R_0}^{(0)})$, and the confirmatory data are the independent subsequent runs $Z_1, \ldots, Z_R$. The time-zero sigma-field $\mathcal{F}_0$ of the confirmatory filtration contains the pilot and the fixed design, so $\sigma(\mathcal{D}_0, \mathcal{B}, \pi) \subseteq \mathcal{F}_0$, and $\mathcal{F}_r = \sigma(\mathcal{F}_0, Z_1, \ldots, Z_r)$. Let $(i_1, \ldots, i_L)$ be the permutation obtained by sorting the pairs $(-\widehat{\theta}_\ell^{(0)}, \ell)$ in increasing lexicographic order, with pilot means defined in (11). Thus larger pilot means come first, and a smaller model

index comes first when two pilot means are equal. The rule gives a unique permutation by finitely many comparisons of pilot-measurable scores. Consequently, $\mathcal{T}_0 = \{i_1, \dots, i_k\}$ is measurable with respect to $\sigma(\mathcal{D}_0, \mathcal{B}, \pi)$ and has exactly k members. Define

$$\mathcal{U}_0 = \{(a \to b) : a \in \mathcal{T}_0,\ b \notin \mathcal{T}_0\}, \qquad \mathcal{U}_{00} = \{e \in \mathcal{U}_0 : \Delta_e^{\mathcal{B}} \le \tau\},$$

where $\Delta_e^{\mathcal{B}} = \Delta_{a,b}^{\mathcal{B}}$ for $e = (a \to b)$. The family $\mathcal{H}(\mathcal{T}_0)$ consists of the hypotheses indexed by $\mathcal{U}_0$, and $\mathcal{U}_{00}$ contains precisely its true nulls. Both sets are measurable with respect to $\mathcal{F}_0$, i.e., membership in $\mathcal{U}_0$ is determined by the pilot, and the benchmark means are fixed given $(\mathcal{B}, \pi)$. In particular, $|\mathcal{U}_0| = k(L-k) > 0$, while $h_0 = |\mathcal{U}_{00}|$ is allowed to be zero. We first work under the conditional law given $\mathcal{F}_0$, for almost every initial design.

For any bounded measurable function h of a complete panel, Assumption 1(1) and then (3) imply

$$\mathbb{E}[h(Z_r) \mid \mathcal{F}_{r-1}] = \mathbb{E}[h(Z_r) \mid \mathcal{F}_0] = \mathbb{E}[h(Z_r) \mid \mathcal{B}, \pi], \tag{28}$$

where the first equality uses conditional independence of the new run from the preceding confirmatory runs, and the second uses independence of the entire initial information, including the pilot, from the confirmatory panels given $(\mathcal{B}, \pi)$. Conditional identical distribution across runs makes the last expectation independent of r. Thus, for each fixed direction $e = (a \to b)$ and its block means $\nu_{e,m,r} = \mathbb{E}[Y_{e,m,r} \mid \mathcal{F}_{r-1}]$,

$$\begin{aligned}\sum_{m=1}^{M} w_m \nu_{e,m,r} &= \mathbb{E}\left[\sum_{m=1}^{M} w_m Y_{e,m,r} \,\middle|\, \mathcal{F}_{r-1}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_{i,r}^{e} \,\middle|\, \mathcal{F}_{r-1}\right] \\ &= \frac{1}{2} + \frac{1}{2N}\sum_{i=1}^{N} \mathbb{E}[S_{i,r}^{(a)} - S_{i,r}^{(b)} \mid \mathcal{B}, \pi] = \frac{1 + \Delta_e^{\mathcal{B}}}{2},\end{aligned} \tag{29}$$

where the first equality uses predictable weights, the second the partition identity, and the third (28). These equalities hold simultaneously for all directions in the finite model universe, hence also for the pilot-selected directions. Under $e \in \mathcal{U}_{00}$, their common value is at most μ_0. No blockwise inequality $\nu_{e,m,r} \le \mu_0$ has been imposed.

We verify the evidence property for this conditional null family. For a fixed grid component set $u_m = \lambda_g w_m / w_\star \in [0, 1)$ and $p_{m,r} = \widehat{Y}_{e,m,r}$. Define

$$Q_{m,r} = \exp\{u_m(Y_{e,m,r} - \nu_{e,m,r}) - \psi_E(u_m)(Y_{e,m,r} - p_{m,r})^2\}.$$

Assumption 1(4) makes $p_{m,r}$ predictable and bounded, so Lemma 1 gives $\mathbb{E}[Q_{m,r} \mid \mathcal{F}_{r-1}] \le 1$. Conditional on this history, the $Q_{m,r}$ are functions of the separate block vectors. Their mutual conditional independence follows from Assumption 1(2). Consequently, for $e \in \mathcal{U}_{00}$,

$$\begin{aligned}\mathbb{E}[F_{e,r}^{\mathcal{C}}(\lambda_g) \mid \mathcal{F}_{r-1}] &= \exp\left\{\frac{\lambda_g}{w_\star}\left(\sum_m w_m \nu_{e,m,r} - \mu_0\right)\right\} \prod_m \mathbb{E}[Q_{m,r} \mid \mathcal{F}_{r-1}] \\ &\le 1.\end{aligned} \tag{30}$$

For clarity, $F_{e,r}^{\mathcal{C}}(\lambda_g)$ is the exponential factor in (6), evaluated on confirmatory run r. The bounds preceding (23) apply here and give

$$-\sum_{m=1}^{M}\{u_m + \psi_E(u_m)\} \le \log F_{e,r}^{\mathcal{C}}(\lambda_g) \le N.$$

Since $u_m < 1$, the lower bound is finite, and hence $0 < F_{e,r}^{\mathcal{C}}(\lambda_g) \le \exp(N)$. For a single grid value λ_g, the corresponding process and the mixture over the grid are, respectively,

$$E_{e,r}^{\mathcal{C}}(\lambda_g) = \prod_{s=1}^{r} F_{e,s}^{\mathcal{C}}(\lambda_g), \qquad E_{e,r}^{\mathcal{C}} = \sum_{g=1}^{G} \rho_g E_{e,r}^{\mathcal{C}}(\lambda_g).$$

The mixture coefficients satisfy $\rho_g \ge 0$ and $\sum_g \rho_g = 1$. Thus both process values lie in $(0, \exp(rN)]$, which gives

$$\mathbb{E}[|E_{e,r}^{\mathcal{C}}(\lambda_g)| \mid \mathcal{F}_0] \le \exp(rN) < \infty, \qquad \mathbb{E}[|E_{e,r}^{\mathcal{C}}| \mid \mathcal{F}_0] \le \exp(rN) < \infty.$$

This is the finite-time integrability required for the conditional expectations below. Each ρ_g is $\mathcal{F}_0$-measurable, and $E^{\mathcal{C}}_{e,r-1}(\lambda_g)$ is $\mathcal{F}_{r-1}$-measurable. For $e \in \mathcal{U}_{00}$, the product recursion and (30) therefore give

$$\begin{aligned}\mathbb{E}[E^{\mathcal{C}}_{e,r} \mid \mathcal{F}_{r-1}] &= \sum_g \rho_g E^{\mathcal{C}}_{e,r-1}(\lambda_g)\mathbb{E}[F^{\mathcal{C}}_{e,r}(\lambda_g) \mid \mathcal{F}_{r-1}] \\ &\le \sum_g \rho_g E^{\mathcal{C}}_{e,r-1}(\lambda_g) = E^{\mathcal{C}}_{e,r-1}.\end{aligned}$$

At $r = 0$, the empty products equal one, so $E^{\mathcal{C}}_{e,0} = 1$. For $s \ge 1$, the observed block mean $Y_{e,m,s}$ in each factor is computed from Z_s, not from any pilot run $Z^{(0)}_j$. In particular, no factor computed from a pilot run is multiplied into $E^{\mathcal{C}}_{e,r}(\lambda_g)$. The pilot-selected family and any allowed initial design information are fixed when conditioning on $\mathcal{F}_0$, and they are not additional confirmatory observations.

If $h_0 = 0$, no true cross-set null can be rejected. Suppose henceforth that $h_0 \ge 1$, and define

$$A_r = \frac{1}{h_0} \sum_{e \in \mathcal{U}_{00}} E^{\mathcal{C}}_{e,r}.$$

Because h_0 and $\mathcal{U}_{00}$ are $\mathcal{F}_0$-measurable, linearity of conditional expectation gives

$$\mathbb{E}[A_r \mid \mathcal{F}_{r-1}] = \frac{1}{h_0} \sum_{e \in \mathcal{U}_{00}} \mathbb{E}[E^{\mathcal{C}}_{e,r} \mid \mathcal{F}_{r-1}] \le \frac{1}{h_0} \sum_{e \in \mathcal{U}_{00}} E^{\mathcal{C}}_{e,r-1} = A_{r-1}.$$

Also $A_0 = h_0^{-1} \sum_{e \in \mathcal{U}_{00}} 1 = 1$. The calculation concerns a sum, not a product, so dependence among the directional processes does not change the equality. For $t = 1/\alpha$, the restricted-family cutoff is

$$J_r = \{e \in \mathcal{U}_0 : E^{\mathcal{C}}_{e,r} < t\}, \qquad c_r = t + \sum_{e \in J_r} (t - E^{\mathcal{C}}_{e,r}).$$

Suppose a true-null direction $e_* \in \mathcal{U}_{00}$ is rejected at r. Then $E^{\mathcal{C}}_{e_*,r} \ge c_r \ge t$, so $e_* \notin J_r$. Set $J_{0,r} = J_r \cap \mathcal{U}_{00}$ and $j_0 = |J_{0,r}|$. For every $e \in J_r$, $E^{\mathcal{C}}_{e,r} < t$, so $t - E^{\mathcal{C}}_{e,r} > 0$. Hence

$$\begin{aligned}E^{\mathcal{C}}_{e_*,r} \ge c_r &= t + \sum_{e \in J_{0,r}} (t - E^{\mathcal{C}}_{e,r}) + \sum_{e \in J_r \setminus J_{0,r}} (t - E^{\mathcal{C}}_{e,r}) \\ &\ge t + \sum_{e \in J_{0,r}} (t - E^{\mathcal{C}}_{e,r}),\end{aligned}$$

where the second sum is nonnegative, including when it is empty. There are $h_0 - j_0 - 1$ directions in $\mathcal{U}_{00} \setminus (J_r \cup \{e_*\})$. Each satisfies $E^{\mathcal{C}}_{e,r} \ge t$ because it is outside J_r. Therefore

$$\begin{aligned}\sum_{e \in \mathcal{U}_{00}} E^{\mathcal{C}}_{e,r} &\ge E^{\mathcal{C}}_{e_*,r} + \sum_{e \in J_{0,r}} E^{\mathcal{C}}_{e,r} + (h_0 - j_0 - 1)t \\ &\ge t + j_0 t + (h_0 - j_0 - 1)t = h_0 t. \end{aligned} \tag{31}$$

Thus any false rejection requires $A_r \ge t$. This derives the needed direct e-Holm implication rather than assuming independent tests (Hartog & Lei, 2025, Theorem 4.2).

Let $T = \inf\{1 \le r \le R : A_r \ge t\}$, with $\inf \varnothing = \infty$. For $1 \le n \le R$, the stopped process satisfies the pathwise identity

$$A_{T \wedge n} = 1 + \sum_{s=1}^{n} \mathbb{1}\{T \ge s\}(A_s - A_{s-1}).$$

Since $\{T \ge s\} = \bigcap_{j=1}^{s-1}\{A_j < t\} \in \mathcal{F}_{s-1}$, the tower property and the supermartingale inequalities give

$$\mathbb{E}[A_{T \wedge n} \mid \mathcal{F}_0] = 1 + \sum_{s=1}^{n} \mathbb{E}[\mathbb{1}\{T \ge s\}\mathbb{E}[A_s - A_{s-1} \mid \mathcal{F}_{s-1}]|\mathcal{F}_0] \le 1.$$

Nonnegativity gives $A_{T\wedge n} \geq t\mathbb{1}\{T \leq n\}$, and hence $\mathbb{P}(T \leq n \mid \mathcal{F}_0) \leq 1/t = \alpha$. Let F_R denote the false-rejection event in (12). Taking $n = R$ and using (31) gives $\mathbb{P}(F_R \mid \mathcal{F}_0) \leq \alpha$. This finite stopped-sum argument is the form of Ville's inequality needed here (Howard et al., 2020). Since $\sigma(\mathcal{D}_0, \mathcal{B}, \pi) \subseteq \mathcal{F}_0$, the tower property gives

$$\mathbb{P}(F_R \mid \mathcal{D}_0, \mathcal{B}, \pi) = \mathbb{E}[\mathbb{P}(F_R \mid \mathcal{F}_0) \mid \mathcal{D}_0, \mathcal{B}, \pi] \leq \alpha,$$

which proves (12) for almost every pilot realization. Averaging once more over the pilot data yields

$$\mathbb{P}(F_R \mid \mathcal{B}, \pi) = \mathbb{E}[\mathbb{P}(F_R \mid \mathcal{D}_0, \mathcal{B}, \pi) \mid \mathcal{B}, \pi] \leq \alpha,$$

including pilot outcomes for which $h_0 = 0$.

For the Top-k certificate, let $C_r = \{\mathcal{U}_0 \subseteq \mathcal{E}_r\}$. On $F_R^c \cap C_r$, every cross-set gap exceeds τ. The two sets are finite and nonempty, so choose $a_* \in \arg\min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}}$ and $b_* \in \arg\max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}}$. Then

$$\min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}} - \max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}} = \Delta_{a_*,b_*}^{\mathcal{B}} > \tau \geq 0. \tag{32}$$

The same argument applies to every $1 \leq r \leq R$ on F_R^c. Conversely, if $\min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}} \leq \max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}} + \tau$, then $\Delta_{a_*,b_*}^{\mathcal{B}} \leq \tau$. If C_r also occurs, $(a_* \to b_*) \in \mathcal{U}_{00} \cap \mathcal{E}_r$, so F_R occurs. This proves

$$\left\{\exists r \leq R : C_r \text{ and } \min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}} \leq \max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}} + \tau\right\} \subseteq F_R. \tag{33}$$

Taking complements and using the conditional error bound proves the precise certificate assertion

$$\mathbb{P}\left(\forall\, 1 \leq r \leq R :\ C_r \implies \min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}} > \max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}} + \tau \,\middle|\, \mathcal{D}_0, \mathcal{B}, \pi\right) \geq 1 - \alpha.$$

Since $\tau \geq 0$, every member of a certified set has a larger benchmark mean than every model outside it on F_R^c.

For $1 \leq r \leq R$, write $C_{\leq r} = \bigcup_{s=1}^{r} C_s$, where $C_s = \{\mathcal{U}_0 \subseteq \mathcal{E}_s\}$. Thus $C_{\leq R}$ is the event that a certificate is issued by run R. Define the candidate's margin-failure event by

$$B_0 = \left\{\min_{a\in\mathcal{T}_0} \theta_a^{\mathcal{B}} \leq \max_{b\notin\mathcal{T}_0} \theta_b^{\mathcal{B}} + \tau\right\}.$$

The inclusion (33) gives

$$\mathbb{P}(C_{\leq R} \cap B_0 \mid \mathcal{D}_0, \mathcal{B}, \pi) \leq \alpha. \tag{34}$$

Given $(\mathcal{D}_0, \mathcal{B}, \pi)$, the candidate and its true benchmark means are fixed, so B_0 is already determined. For any such pilot outcome in B_0, (34) reduces to $\mathbb{P}(C_{\leq R} \mid \mathcal{D}_0, \mathcal{B}, \pi) \leq \alpha$. An incorrect pilot selection is therefore allowed without assuming that the selected set satisfies the required margin.

For this fixed-candidate procedure, the event $A_{\leq r}$ in (55) is exactly $C_{\leq r} \cap B_0$. Since $C_{\leq r} \subseteq C_{\leq R}$, monotonicity of conditional expectation and (34) give, almost surely,

$$\mathbb{P}(C_{\leq r} \cap B_0 \mid \mathcal{D}_0, \mathcal{B}, \pi) \leq \mathbb{P}(C_{\leq R} \cap B_0 \mid \mathcal{D}_0, \mathcal{B}, \pi) \leq \alpha.$$

Averaging over the pilot data with the tower property gives the error metric in (56),

$$\begin{aligned}\mathrm{FCP}_r &= \mathbb{P}(C_{\leq r} \cap B_0 \mid \mathcal{B}, \pi) \\ &= \mathbb{E}[\mathbb{P}(C_{\leq r} \cap B_0 \mid \mathcal{D}_0, \mathcal{B}, \pi) \,|\, \mathcal{B}, \pi] \leq \alpha.\end{aligned}$$

If $\mathrm{CertProb}_r = \mathbb{P}(C_{\leq r} \mid \mathcal{B}, \pi) > 0$, then

$$\mathbb{P}(B_0 \mid C_{\leq r}, \mathcal{B}, \pi) = \frac{\mathbb{P}(C_{\leq r} \cap B_0 \mid \mathcal{B}, \pi)}{\mathbb{P}(C_{\leq r} \mid \mathcal{B}, \pi)} = \frac{\mathrm{FCP}_r}{\mathrm{CertProb}_r}.$$

Therefore, the guarantee $\mathrm{FCP}_r \leq \alpha$ controls the probability of issuing an incorrect certificate, rather than the conditional error probability given that certification occurs. This distinction is particularly clear after conditioning on the pilot. That is, given $(\mathcal{D}_0, \mathcal{B}, \pi)$, the event B_0 is fixed. Hence, if B_0 holds for a particular pilot realization and certification occurs with positive conditional probability, then

$$\mathbb{P}(B_0 \mid C_{\leq R}, \mathcal{D}_0, \mathcal{B}, \pi) = 1.$$

Finally, if some $a \in \mathcal{T}_0$ and $b \notin \mathcal{T}_0$ have $\theta_a^{\mathcal{B}} = \theta_b^{\mathcal{B}}$, then $\Delta_{a,b}^{\mathcal{B}} = 0 \leq \tau$. On F_R^c, this direction is never rejected, so the requirement $\mathcal{U}_0 \subseteq \mathcal{E}_r$ cannot be met at any $r \leq R$. This concerns a tie in true benchmark means across the two groups, not a tie in the pilot scores. □

D.5 Proof of Corollary 1

Proof. Fix an integer $R \geq 1$ and consider the first R completed runs. For the true-null set $\mathcal{N}_0$ in Theorem 1, define

$$\begin{aligned}\Omega_R &= \bigcap_{r=1}^{R} \{\mathcal{E}_r \cap \mathcal{N}_0 = \varnothing\} \\ &= \left\{\forall\, 1 \leq r \leq R,\ \forall (a \to b) \in \mathcal{E}_r : \theta_a^{\mathcal{B}} - \theta_b^{\mathcal{B}} = \Delta_{a,b}^{\mathcal{B}} > \tau\right\}.\end{aligned} \tag{35}$$

By (10), $\mathbb{P}(\Omega_R \mid \mathcal{B}, \pi) \geq 1 - \alpha$. For the interval in (13), write

$$\mathcal{I}_{\ell,r} = \{x \in \mathbb{R} : 1 + |\mathcal{A}_{\ell,r}| \leq x \leq L - |\mathcal{D}_{\ell,r}|\}.$$

Define the joint coverage event

$$C_R = \bigcap_{r=1}^{R} \bigcap_{\ell \in \mathcal{V}} \{\mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r}\}.$$

We will prove $\Omega_R \subseteq C_R$.

Fix $\omega \in \Omega_R$, $1 \leq r \leq R$, and $\ell \in \mathcal{V}$. All graph quantities in the following calculation are evaluated at ω. A directed cycle $v_0 \to v_1 \to \cdots \to v_d = v_0$ with $d \geq 1$ would imply

$$0 = \theta_{v_0}^{\mathcal{B}} - \theta_{v_d}^{\mathcal{B}} = \sum_{j=0}^{d-1} (\theta_{v_j}^{\mathcal{B}} - \theta_{v_{j+1}}^{\mathcal{B}}) > d\tau \geq 0,$$

This gives $0 > 0$, a contradiction. Thus $(\mathcal{V}, \mathcal{E}_r)$ is acyclic.

Define the true sets of strictly better, strictly worse, and tied models by

$$\begin{aligned}\mathcal{B}_\ell &= \{j \neq \ell : \theta_j^{\mathcal{B}} > \theta_\ell^{\mathcal{B}}\}, \\ \mathcal{W}_\ell &= \{j \neq \ell : \theta_j^{\mathcal{B}} < \theta_\ell^{\mathcal{B}}\}, \\ \mathcal{O}_\ell &= \{j \neq \ell : \theta_j^{\mathcal{B}} = \theta_\ell^{\mathcal{B}}\}.\end{aligned}$$

They are pairwise disjoint and satisfy

$$\mathcal{B}_\ell \cup \mathcal{W}_\ell \cup \mathcal{O}_\ell = \mathcal{V} \setminus \{\ell\}, \qquad |\mathcal{B}_\ell| + |\mathcal{W}_\ell| + |\mathcal{O}_\ell| = L - 1.$$

If $j \in \mathcal{A}_{\ell,r}$, there exists a path $j = v_0 \to \cdots \to v_d = \ell$ of length $d \geq 1$. Summing its true edge inequalities gives

$$\theta_j^{\mathcal{B}} - \theta_\ell^{\mathcal{B}} = \sum_{q=0}^{d-1} (\theta_{v_q}^{\mathcal{B}} - \theta_{v_{q+1}}^{\mathcal{B}}) > d\tau \geq 0.$$

Thus $j \in \mathcal{B}_\ell$. If instead $j \in \mathcal{D}_{\ell,r}$, choose a path $\ell = u_0 \to \cdots \to u_q = j$, where $q \geq 1$. Then

$$\theta_\ell^{\mathcal{B}} - \theta_j^{\mathcal{B}} = \sum_{s=0}^{q-1} (\theta_{u_s}^{\mathcal{B}} - \theta_{u_{s+1}}^{\mathcal{B}}) > q\tau \geq 0,$$

so $j \in \mathcal{W}_\ell$. We have proved the two inclusions

$$\mathcal{A}_{\ell,r} \subseteq \mathcal{B}_\ell, \qquad \mathcal{D}_{\ell,r} \subseteq \mathcal{W}_\ell \subseteq (\mathcal{V} \setminus \{\ell\}) \setminus \mathcal{B}_\ell. \tag{36}$$

In particular,

$$\mathcal{A}_{\ell,r} \cap \mathcal{D}_{\ell,r} = \varnothing, \qquad (\mathcal{A}_{\ell,r} \cup \mathcal{D}_{\ell,r}) \cap \mathcal{O}_\ell = \varnothing.$$

The rank in the corollary is $\mathrm{rank}_B(\ell) = 1 + |\mathcal{B}_\ell|$. Using (36) and the partition above, the differences between this rank and the two interval endpoints are

$$\begin{aligned}\mathrm{rank}_B(\ell) - (1 + |\mathcal{A}_{\ell,r}|) &= |\mathcal{B}_\ell| - |\mathcal{A}_{\ell,r}| = |\mathcal{B}_\ell \setminus \mathcal{A}_{\ell,r}| \geq 0, \\ (L - |\mathcal{D}_{\ell,r}|) - \mathrm{rank}_B(\ell) &= L - 1 - |\mathcal{B}_\ell| - |\mathcal{D}_{\ell,r}| \\ &= |\mathcal{W}_\ell \setminus \mathcal{D}_{\ell,r}| + |\mathcal{O}_\ell| \geq 0.\end{aligned} \tag{37}$$

Consequently,

$$\begin{aligned} 1 + |\mathcal{A}_{\ell,r}| \le 1 + |\mathcal{B}_\ell| = \mathrm{rank}_B(\ell) \le L - |\mathcal{D}_{\ell,r}|, \\ \mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r} = [1 + |\mathcal{A}_{\ell,r}|,\ L - |\mathcal{D}_{\ell,r}|] \neq \varnothing. \end{aligned} \tag{38}$$

The calculation holds for every $r \le R$ and every ℓ at the fixed outcome ω. Since $\omega \in \Omega_R$ was arbitrary,

$$\Omega_R \subseteq C_R, \qquad \mathbb{1}\{\Omega_R\} \le \mathbb{1}\{C_R\}. \tag{39}$$

To justify taking conditional expectations in (39), note first that $E^{\mathcal{C}}_{e,r}$ and the cutoff c_r are $\mathcal{F}_r$-measurable. Thus $\{e \in \mathcal{E}_r\} = \{E^{\mathcal{C}}_{e,r} \ge c_r\} \in \mathcal{F}_r$ for every tested direction; an untested direction has an empty edge event. For distinct vertices a, b, reachability satisfies

$$\{a \rightsquigarrow_r b\} = \bigcup_{d=1}^{L-1} \bigcup_{\substack{v_0=a,\, v_d=b \\ v_0,\dots,v_d \text{ distinct}}} \bigcap_{j=0}^{d-1} \{(v_j \to v_{j+1}) \in \mathcal{E}_r\} \in \mathcal{F}_r. \tag{40}$$

Indeed, deleting a segment between repeated vertices shortens a walk without changing its endpoints; successive deletions give a path with at most $L-1$ edges. The unions in (40) are therefore finite. It follows that

$$|\mathcal{A}_{\ell,r}| = \sum_{j \neq \ell} \mathbb{1}\{j \rightsquigarrow_r \ell\}, \qquad |\mathcal{D}_{\ell,r}| = \sum_{j \neq \ell} \mathbb{1}\{\ell \rightsquigarrow_r j\}$$

are $\mathcal{F}_r$-measurable. The true rank is $\mathcal{F}_0$-measurable, so

$$\begin{aligned} \{\mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r}\} = \{1 + |\mathcal{A}_{\ell,r}| \le \mathrm{rank}_B(\ell)\} \\ \cap \{\mathrm{rank}_B(\ell) \le L - |\mathcal{D}_{\ell,r}|\} \in \mathcal{F}_r. \end{aligned}$$

Hence $C_R, \Omega_R \in \mathcal{F}_R$. Taking conditional expectations of the indicator inequality in (39) gives

$$\begin{aligned} \mathbb{P}(C_R \mid \mathcal{B}, \pi) &= \mathbb{E}[\mathbb{1}\{C_R\} \mid \mathcal{B}, \pi] \\ &\ge \mathbb{E}[\mathbb{1}\{\Omega_R\} \mid \mathcal{B}, \pi] = \mathbb{P}(\Omega_R \mid \mathcal{B}, \pi) \ge 1 - \alpha. \end{aligned} \tag{41}$$

This proves the corollary for the first R completed runs.

For the additional statement following the corollary, suppose Assumption 1 holds for one sequence $(Z_r)_{r \ge 1}$. For every integer $n \ge 1$, (41) applies to its first n runs. The corresponding events satisfy

$$C_{n+1} \subseteq C_n, \qquad C_\infty := \bigcap_{n=1}^{\infty} C_n = \bigcap_{r=1}^{\infty} \bigcap_{\ell \in \mathcal{V}} \{\mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,r}\}.$$

Since $\mathbb{1}\{C_n^c\} \uparrow \mathbb{1}\{C_\infty^c\}$, conditional monotone convergence gives

$$\begin{aligned} \mathbb{P}(C_\infty \mid \mathcal{B}, \pi) &= 1 - \mathbb{E}[\mathbb{1}\{C_\infty^c\} \mid \mathcal{B}, \pi] \\ &= 1 - \lim_{n \to \infty} \mathbb{E}[\mathbb{1}\{C_n^c\} \mid \mathcal{B}, \pi] = \lim_{n \to \infty} \mathbb{P}(C_n \mid \mathcal{B}, \pi) \ge 1 - \alpha \quad \text{a.s.} \end{aligned}$$

The bounds $\mathbb{P}(C_n \mid \mathcal{B}, \pi) \ge 1 - \alpha$ in (41) hold simultaneously for all n outside a countable union of probability-zero sets. □

D.6 Proof of Theorem 3

Proof. For an epoch v that is entered, $\mathcal{K}_{v-1}$ is the sigma-field containing all information available before its fresh runs, as defined in Section C. Its within-epoch filtration is

$$\mathcal{F}_0^{(v)} = \mathcal{K}_{v-1}, \qquad \mathcal{F}_r^{(v)} = \sigma(\mathcal{K}_{v-1}, Z_1^{(v)}, \dots, Z_r^{(v)}).$$

Here $Z_r^{(v)}$ contains all scores on the frozen benchmark $B^{(v)}$. In this proof, $S_{i,r}^{(\ell)}$ denotes its component for item i and model ℓ in epoch v. We first condition on $\mathcal{K}_{v-1}$. By the design in Section C,

$$\sigma\Big(B^{(v)}, \mathcal{H}^{(v)}, \mathcal{C}^{(v)}, \Lambda^{(v)}, (\rho_g^{(v)})_g, \alpha_v\Big) \subseteq \mathcal{K}_{v-1},$$

where $\mathcal{C}^{(v)}$ is the partition and $\Lambda^{(v)}$ is the stake grid in epoch v. Define

$$\mathcal{U}^{(v)} = \{e : H_e^{(v,\tau)} \in \mathcal{H}^{(v)}\}, \qquad \mathcal{N}_0^{(v)} = \{e \in \mathcal{U}^{(v)} : \Delta_e^{B^{(v)}} \le \tau\},$$

$$\Delta_e^{B^{(v)}} = \Delta_{a,b}^{B^{(v)}} \quad (e = (a \to b)), \qquad h_v = |\mathcal{N}_0^{(v)}|.$$

Since the benchmark means are functions of $(B^{(v)}, \pi)$,

$$\mathbb{1}\{e \in \mathcal{N}_0^{(v)}\} = \mathbb{1}\{e \in \mathcal{U}^{(v)}\}\mathbb{1}\{\Delta_e^{B^{(v)}} \le \tau\}$$

is $\mathcal{K}_{v-1}$-measurable for each direction e, as is h_v.

Write $P_{b,\pi}$ for the fixed execution law of a fresh complete panel on benchmark b. The execution-law condition in the theorem means that, for every bounded measurable panel function h,

$$\mathbb{E}[h(Z_r^{(v)}) \mid \mathcal{F}_{r-1}^{(v)}] = \mathbb{E}[h(Z_r^{(v)}) \mid \mathcal{K}_{v-1}] = \int h(z)\, P_{B^{(v)},\pi}(dz). \tag{42}$$

The first equality uses conditional run independence in Assumption 1(1); the second uses the fixed execution law in Theorem 3. Because $\sigma(B^{(v)}, \pi) \subseteq \mathcal{K}_{v-1}$, the tower property also gives

$$\begin{aligned}\mathbb{E}[h(Z_r^{(v)}) \mid B^{(v)}, \pi] &= \mathbb{E}\Big[\mathbb{E}[h(Z_r^{(v)}) \mid \mathcal{K}_{v-1}]\Big| B^{(v)}, \pi\Big] \\ &= \mathbb{E}\left[\int h(z) P_{B^{(v)},\pi}(dz)\middle| B^{(v)}, \pi\right] = \int h(z) P_{B^{(v)},\pi}(dz).\end{aligned}$$

Thus the conditional expectations in (42) use the same law as the benchmark means $\theta_\ell^{B^{(v)}}$.

Let the epoch partition have M_v nonempty blocks with $w_m^{(v)} = |C_m^{(v)}|/N_v$ and $w_\star^{(v)} = \max_m w_m^{(v)} > 0$. For $e = (a \to b)$, define

$$Y_{e,m,r}^{(v)} = \frac{1}{|C_m^{(v)}|} \sum_{i \in C_m^{(v)}} \frac{S_{i,r}^{(a)} - S_{i,r}^{(b)} + 1}{2}, \qquad \nu_{e,m,r}^{(v)} = \mathbb{E}[Y_{e,m,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}].$$

The partition identity and (42) give

$$\begin{aligned}\sum_{m=1}^{M_v} w_m^{(v)} \nu_{e,m,r}^{(v)} &= \mathbb{E}\left[\frac{1}{N_v} \sum_{i \in B^{(v)}} \frac{S_{i,r}^{(a)} - S_{i,r}^{(b)} + 1}{2}\middle| \mathcal{F}_{r-1}^{(v)}\right] \\ &= \frac{1 + \Delta_e^{B^{(v)}}}{2} \le \mu_0, \qquad e \in \mathcal{N}_0^{(v)}.\end{aligned} \tag{43}$$

Here $e \in \mathcal{N}_0^{(v)}$ implies $(1 + \Delta_e^{B^{(v)}})/2 \le (1+\tau)/2 = \mu_0$.

For each grid value $\lambda_g^{(v)}$, define $u_{g,m}^{(v)} = \lambda_g^{(v)} w_m^{(v)} / w_\star^{(v)} \in [0, 1)$. The prediction $\widehat{Y}_{e,m,r}^{(v)}$ is $\mathcal{F}_{r-1}^{(v)}$-measurable and belongs to $[0, 1]$ by Assumption 1(4). Write

$$Q_{g,m,r}^{(v)} = \exp\Big\{u_{g,m}^{(v)}(Y_{e,m,r}^{(v)} - \nu_{e,m,r}^{(v)}) - \psi_E(u_{g,m}^{(v)})(Y_{e,m,r}^{(v)} - \widehat{Y}_{e,m,r}^{(v)})^2\Big\}.$$

Lemma 1 gives $\mathbb{E}[Q_{g,m,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] \le 1$. Given $\mathcal{F}_{r-1}^{(v)}$, each $Q_{g,m,r}^{(v)}$ is a function of the block score vector $V_{m,r}^{(v)} = (S_{i,r}^{(\ell)} : i \in C_m^{(v)},\ 1 \le \ell \le L)$. Assumption 1(2) therefore yields

$$\mathbb{E}\left[\prod_{m=1}^{M_v} Q_{g,m,r}^{(v)}\middle| \mathcal{F}_{r-1}^{(v)}\right] = \prod_{m=1}^{M_v} \mathbb{E}[Q_{g,m,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] \le 1.$$

For $e \in \mathcal{N}_0^{(v)}$, (43) gives

$$\sum_m u_{g,m}^{(v)}(\nu_{e,m,r}^{(v)} - \mu_0) = \frac{\lambda_g^{(v)}}{w_\star^{(v)}}\left(\sum_m w_m^{(v)} \nu_{e,m,r}^{(v)} - \mu_0\right) \le 0. \tag{44}$$

Decomposing the factor in (6) accordingly gives

$$\mathbb{E}[F_{e,r}^{(v)}(\lambda_g^{(v)}) \mid \mathcal{F}_{r-1}^{(v)}] = \exp\left\{\sum_m u_{g,m}^{(v)}(\nu_{e,m,r}^{(v)} - \mu_0)\right\} \prod_m \mathbb{E}[Q_{g,m,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] \le 1. \tag{45}$$

For grid value $\lambda_g^{(v)}$, the process is $E_{e,r}^{(v)}(\lambda_g^{(v)}) = \prod_{s=1}^{r} F_{e,s}^{(v)}(\lambda_g^{(v)})$, and its mixture coefficient $\rho_g^{(v)} \geq 0$ is fixed at the start of epoch v, with $\sum_g \rho_g^{(v)} = 1$. The bounds in (23), applied to the epoch partition, give

$$0 < E_{e,r}^{(v)}(\lambda_g^{(v)}) \leq \exp(rN_v), \qquad 0 < E_{e,r}^{(v)} = \sum_g \rho_g^{(v)} E_{e,r}^{(v)}(\lambda_g^{(v)}) \leq \exp(rN_v).$$

Since N_v is finite and $\mathcal{K}_{v-1}$-measurable,

$$\mathbb{E}[|E_{e,r}^{(v)}(\lambda_g^{(v)})| \mid \mathcal{K}_{v-1}] \leq \exp(rN_v) < \infty, \qquad \mathbb{E}[|E_{e,r}^{(v)}| \mid \mathcal{K}_{v-1}] \leq \exp(rN_v) < \infty.$$

The product $E_{e,r-1}^{(v)}(\lambda_g^{(v)})$ and the coefficient $\rho_g^{(v)}$ are $\mathcal{F}_{r-1}^{(v)}$-measurable. The product recursion and (45) give, for $e \in \mathcal{N}_0^{(v)}$,

$$\begin{aligned}\mathbb{E}[E_{e,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] &= \sum_g \rho_g^{(v)} E_{e,r-1}^{(v)}(\lambda_g^{(v)}) \mathbb{E}[F_{e,r}^{(v)}(\lambda_g^{(v)}) \mid \mathcal{F}_{r-1}^{(v)}] \\ &\leq \sum_g \rho_g^{(v)} E_{e,r-1}^{(v)}(\lambda_g^{(v)}) = E_{e,r-1}^{(v)}, \qquad E_{e,0}^{(v)} = \sum_g \rho_g^{(v)} = 1. \end{aligned} \tag{46}$$

Define the epoch error event by

$$F_v = \bigcup_{r=1}^{\infty} \{\mathcal{E}_r^{(v)} \cap \mathcal{N}_0^{(v)} \neq \varnothing\}.$$

If $\alpha_v = 0$, then $\mathcal{E}_r^{(v)} = \varnothing$ for all r, so $F_v = \varnothing$. If $h_v = 0$, then $\mathcal{N}_0^{(v)} = \varnothing$, again giving $F_v = \varnothing$. It remains to consider $\alpha_v > 0$ and $h_v \geq 1$. Define

$$A_r^{(v)} = \frac{1}{h_v} \sum_{e \in \mathcal{N}_0^{(v)}} E_{e,r}^{(v)}, \qquad t_v = 1/\alpha_v.$$

Since h_v and $\mathcal{N}_0^{(v)}$ are $\mathcal{K}_{v-1}$-measurable, (46) gives

$$\mathbb{E}[A_r^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] = \frac{1}{h_v} \sum_{e \in \mathcal{N}_0^{(v)}} \mathbb{E}[E_{e,r}^{(v)} \mid \mathcal{F}_{r-1}^{(v)}] \leq A_{r-1}^{(v)}, \qquad A_0^{(v)} = 1.$$

Set

$$J_r^{(v)} = \{e \in \mathcal{U}^{(v)} : E_{e,r}^{(v)} < t_v\}, \qquad c_r^{(v)} = t_v + \sum_{e \in J_r^{(v)}} (t_v - E_{e,r}^{(v)}).$$

If some $e_* \in \mathcal{N}_0^{(v)}$ is rejected, it lies outside $J_r^{(v)}$ and satisfies $E_{e_*,r}^{(v)} \geq c_r^{(v)}$. With $J_{0,r}^{(v)} = J_r^{(v)} \cap \mathcal{N}_0^{(v)}$, we have

$$\begin{aligned} E_{e_*,r}^{(v)} \geq c_r^{(v)} &= t_v + \sum_{e \in J_{0,r}^{(v)}} (t_v - E_{e,r}^{(v)}) + \sum_{e \in J_r^{(v)} \setminus J_{0,r}^{(v)}} (t_v - E_{e,r}^{(v)}) \\ &\geq t_v + \sum_{e \in J_{0,r}^{(v)}} (t_v - E_{e,r}^{(v)}), \end{aligned}$$

because $E_{e,r}^{(v)} < t_v$ for every $e \in J_r^{(v)}$. Each of the $h_v - |J_{0,r}^{(v)}| - 1$ directions in $\mathcal{N}_0^{(v)} \setminus (J_r^{(v)} \cup \{e_*\})$ satisfies $E_{e,r}^{(v)} \geq t_v$. Therefore

$$\begin{aligned} \sum_{e \in \mathcal{N}_0^{(v)}} E_{e,r}^{(v)} &\geq E_{e_*,r}^{(v)} + \sum_{e \in J_{0,r}^{(v)}} E_{e,r}^{(v)} + (h_v - |J_{0,r}^{(v)}| - 1) t_v \\ &\geq t_v + \sum_{e \in J_{0,r}^{(v)}} (t_v - E_{e,r}^{(v)}) + \sum_{e \in J_{0,r}^{(v)}} E_{e,r}^{(v)} + (h_v - |J_{0,r}^{(v)}| - 1) t_v = h_v t_v. \end{aligned}$$

Dividing by h_v proves

$$\{\mathcal{E}_r^{(v)} \cap \mathcal{N}_0^{(v)} \neq \varnothing\} \subseteq \{A_r^{(v)} \geq 1/\alpha_v\}. \tag{47}$$

Fix an integer $n \geq 1$ through which the epoch assumptions hold, and let

$$T_{v,n} = \inf\{1 \leq r \leq n : A_r^{(v)} \geq t_v\}, \qquad \inf \varnothing = \infty.$$

For $1 \leq s \leq n$,

$$\{T_{v,n} \geq s\} = \bigcap_{j=1}^{s-1} \{A_j^{(v)} < t_v\} \in \mathcal{F}_{s-1}^{(v)}, \qquad A_{T_{v,n}\wedge n}^{(v)} = 1 + \sum_{s=1}^{n} \mathbb{1}\{T_{v,n} \geq s\}(A_s^{(v)} - A_{s-1}^{(v)}).$$

Taking conditional expectations of this identity gives

$$\mathbb{E}[A_{T_{v,n}\wedge n}^{(v)} \mid \mathcal{K}_{v-1}] = 1 + \sum_{s=1}^{n} \mathbb{E}\Big[\mathbb{1}\{T_{v,n} \geq s\}\mathbb{E}[A_s^{(v)} - A_{s-1}^{(v)} \mid \mathcal{F}_{s-1}^{(v)}]\Big|\mathcal{K}_{v-1}\Big] \leq 1.$$

Nonnegativity and $t_v = 1/\alpha_v$ imply

$$\mathbb{1}\{T_{v,n} \leq n\} \leq \alpha_v A_{T_{v,n}\wedge n}^{(v)}.$$

Since α_v is $\mathcal{K}_{v-1}$-measurable,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\leq r\leq n} A_r^{(v)} \geq 1/\alpha_v \Big| \mathcal{K}_{v-1}\Big) &= \mathbb{P}(T_{v,n} \leq n \mid \mathcal{K}_{v-1}) \\ &\leq \alpha_v \mathbb{E}[A_{T_{v,n}\wedge n}^{(v)} \mid \mathcal{K}_{v-1}] \leq \alpha_v.\end{aligned} \tag{48}$$

For a finite run budget $n_v \geq 1$ fixed at the start of epoch v, the reporting convention gives $\mathcal{E}_r^{(v)} = \varnothing$ for $r > n_v$. Hence

$$F_v = \bigcup_{r=1}^{n_v} \{\mathcal{E}_r^{(v)} \cap \mathcal{N}_0^{(v)} \neq \varnothing\} \subseteq \left\{\max_{1\leq r\leq n_v} A_r^{(v)} \geq t_v\right\},$$

by (47). Applying (48) with $n = n_v$ bounds $\mathbb{P}(F_v \mid \mathcal{K}_{v-1})$ by α_v.

If the assumptions instead hold for the entire potential sequence $(Z_r^{(v)})_{r\geq 1}$, define

$$D_{v,n} = \bigcup_{r=1}^{n} \{A_r^{(v)} \geq t_v\}, \qquad D_{v,n} \uparrow D_{v,\infty} := \bigcup_{r=1}^{\infty} \{A_r^{(v)} \geq t_v\}.$$

Again (47) gives $F_v \subseteq D_{v,\infty}$, including when evaluation stops after a data-dependent number of runs, because no reports are issued after that stop. Conditional monotone convergence and (48) yield

$$\begin{aligned}\mathbb{P}(F_v \mid \mathcal{K}_{v-1}) &\leq \mathbb{E}[\mathbb{1}\{D_{v,\infty}\} \mid \mathcal{K}_{v-1}] \\ &= \lim_{n\to\infty} \mathbb{E}[\mathbb{1}\{D_{v,n}\} \mid \mathcal{K}_{v-1}] \leq \alpha_v.\end{aligned}$$

If $h_v = 0$, $F_v = \varnothing$. If $\alpha_v = 0$ or epoch v is not entered, $\mathcal{E}_r^{(v)} = \varnothing$ for all r, so $F_v = \varnothing$ as well. Combining these cases proves

$$\mathbb{E}[\mathbb{1}\{F_v\} \mid \mathcal{K}_{v-1}] = \mathbb{P}(F_v \mid \mathcal{K}_{v-1}) \leq \alpha_v \quad \text{almost surely.} \tag{49}$$

Now remove the conditioning separately for each epoch. The tower property, countable subadditivity, and Tonelli's theorem for the nonnegative budgets give

$$\begin{aligned}\mathbb{P}\left(\bigcup_{v\geq 1} F_v\right) &\leq \sum_{v\geq 1} \mathbb{P}(F_v) = \sum_{v\geq 1} \mathbb{E}[\mathbb{E}[\mathbb{1}\{F_v\} \mid \mathcal{K}_{v-1}]] \\ &\leq \sum_{v\geq 1} \mathbb{E}[\alpha_v] = \mathbb{E}\left[\sum_{v\geq 1} \alpha_v\right] \leq \alpha.\end{aligned} \tag{50}$$

The final inequality follows from $\sum_{v\geq 1} \alpha_v \leq \alpha$ almost surely. The calculation uses no factorization of probabilities across epochs.

By the definition of $\mathcal{N}_0^{(v)}$,

$$\bigcup_{v\geq 1} F_v = \left\{\exists v \geq 1,\ \exists r \geq 1, \quad \exists (a \to b) \in \mathcal{E}_r^{(v)} : \Delta_{a,b}^{B^{(v)}} \leq \tau\right\}.$$

Substituting this identity into (50) proves Theorem 3. □

D.7 Proof of Corollary 2

Proof. For the initial stage $v = 0$, the initial panels satisfy Assumption 1. Therefore, Theorem 1 gives $\mathbb{P}(F_0 \mid \mathcal{B}, \pi) \le \alpha_0$, where $F_0 = \left\{\exists r \le R^{(0)}, \exists e \in \mathcal{E}^{(0)}_{r,\mathrm{new}} \text{ such that } \Delta^{\mathcal{B}}_e \le \tau\right\}$.

Fix a stage $v \ge 1$. Let $\mathcal{Q}_v$ collect the model-inclusion rule, pairing rule, directional family, partition, stake grid, mixture weights, and all other stage-v design choices. Define the replay filtration

$$\mathcal{F}^{(v)}_r = \sigma\Big(\mathcal{B}, \pi, \mathcal{Q}_v, \widetilde{Z}^{(v)}_1, \dots, \widetilde{Z}^{(v)}_r\Big), \qquad r = 0, \dots, R^{(v)},$$

with $\mathcal{F}^{(v)}_0 = \sigma(\mathcal{B}, \pi, \mathcal{Q}_v)$.

By assumption, $\mathcal{Q}_v$ is independent of the augmented inference panels conditional on $(\mathcal{B}, \pi)$, and the augmented panels satisfy Assumption 1 with respect to $(\mathcal{F}^{(v)}_r)^{R^{(v)}}_{r=0}$.

Although the stored outcomes are available when stage v begins, the new e-processes replay the augmented panels in replicate order. Hence every prediction and stake used at replicate r is $\mathcal{F}^{(v)}_{r-1}$-measurable and does not depend on stored outcomes from replicate r or later. The supermartingale argument in Theorem 1 therefore applies with respect to the replay filtration.

The proof of Theorem 1 applies to any directional family fixed before inference and therefore applies to $\mathcal{H}^{(v)}_{\mathrm{new}}$. Defining

$$F_v = \left\{\exists r \le R^{(v)},\ \exists e \in \mathcal{E}^{(v)}_{r,\mathrm{new}} \text{ such that } \Delta^{\mathcal{B}}_e \le \tau\right\},$$

we obtain $\mathbb{P}(F_v \mid \mathcal{B}, \pi) \le \alpha_v$.

The events $(F_v)_{v \ge 0}$ may be arbitrarily dependent because stored outcomes can be reused across stages. Countable subadditivity nevertheless gives

$$\mathbb{P}\left(\bigcup_{v \ge 0} F_v \,\middle|\, \mathcal{B}, \pi\right) \le \sum_{v \ge 0} \mathbb{P}(F_v \mid \mathcal{B}, \pi) \le \sum_{v \ge 0} \alpha_v \le \alpha.$$

On the complementary event, every retained or newly certified edge is a true margin claim. Therefore, all path implications, simultaneous rank intervals, and Top-k certificates derived from the combined graph are valid. □

D.8 The grid-dependent power statement

This subsection supplies the condition and proof for the power statement already used in the main text, which concerns the same fixed grid and predictions, not a different evidence construction. For a tested alternative $e = (a \to b)$, set $\delta_e = (\Delta^{\mathcal{B}}_{a,b} - \tau)/2 > 0$ and define

$$\begin{aligned} q_{e,m,r} &= \mathbb{E}[(Y_{e,m,r} - \widehat{Y}_{e,m,r})^2 \mid \mathcal{F}_{r-1}], \\ a_{e,g,r} &= \frac{\lambda_g}{w_\star}\delta_e - \sum_m \psi_E(\lambda_{g,m}) q_{e,m,r}. \end{aligned} \tag{51}$$

The quantity $a_{e,g,r}$ is the conditional expected log increment, as verified in the proof below.

Theorem 4. *Suppose Assumption 1 holds for the entire infinite sequence, with a finite family and a fixed finite grid. Fix a tested direction e with $\Delta^{\mathcal{B}}_{a,b} > \tau$. If for some grid component g with $\rho_g > 0$,*

$$\liminf_{r \to \infty} \frac{1}{r} \sum_{s=1}^{r} a_{e,g,s} > 0 \quad \textit{almost surely}, \tag{52}$$

then $E^{\mathcal{C}}_{e,r} \to \infty$ almost surely, and direct e-Holm rejects $H^{(\tau)}_e$ at every sufficiently large completed replicate almost surely. In particular, its probability of rejection tends to one.

Proof. Let $\mathcal{F}_0 = \sigma(\mathcal{B}, \pi, \mathcal{Q})$, and fix the grid index g in the theorem, for which $\rho_g > 0$. Define

$$G_g = \left\{\liminf_{r \to \infty} \frac{1}{r} \sum_{s=1}^{r} a_{e,g,s} > 0\right\}.$$

Assumption (52) gives $\mathbb{P}(G_g^c) = 0$. By the tower property,

$$0 = \mathbb{P}(G_g^c) = \mathbb{E}[\mathbb{P}(G_g^c \mid \mathcal{F}_0)], \qquad \mathbb{P}(G_g^c \mid \mathcal{F}_0) \geq 0,$$

so $\mathbb{P}(G_g \mid \mathcal{F}_0) = 1$ almost surely. We work under the conditional law given $\mathcal{F}_0$ for such initial designs. Define $\ell_{g,r} = \log F_{e,r}^{\mathcal{C}}(\lambda_g)$. Taking conditional expectations in (6) gives

$$\begin{aligned}\mathbb{E}[\ell_{g,r} \mid \mathcal{F}_{r-1}] &= \sum_m \lambda_{g,m}(\nu_{e,m,r} - \mu_0) - \sum_m \psi_E(\lambda_{g,m}) q_{e,m,r} \\ &= \frac{\lambda_g}{w_\star}\delta_e - \sum_m \psi_E(\lambda_{g,m}) q_{e,m,r} = a_{e,g,r}, \end{aligned} \tag{53}$$

where the second equality is (20). Both $\nu_{e,m,r}$ and $\widehat{Y}_{e,m,r}$ are $\mathcal{F}_{r-1}$-measurable. Since $\mathbb{E}[Y_{e,m,r} - \nu_{e,m,r} \mid \mathcal{F}_{r-1}] = 0$,

$$\begin{aligned} q_{e,m,r} &= \mathbb{E}\Big[\{(Y_{e,m,r} - \nu_{e,m,r}) + (\nu_{e,m,r} - \widehat{Y}_{e,m,r})\}^2 \Big| \mathcal{F}_{r-1}\Big] \\ &= \mathbb{E}[(Y_{e,m,r} - \nu_{e,m,r})^2 \mid \mathcal{F}_{r-1}] \\ &\quad + 2(\nu_{e,m,r} - \widehat{Y}_{e,m,r})\mathbb{E}[Y_{e,m,r} - \nu_{e,m,r} \mid \mathcal{F}_{r-1}] + (\nu_{e,m,r} - \widehat{Y}_{e,m,r})^2 \\ &= \mathrm{Var}(Y_{e,m,r} \mid \mathcal{F}_{r-1}) + (\nu_{e,m,r} - \widehat{Y}_{e,m,r})^2. \end{aligned}$$

To pass from conditional expected growth to realized growth, write

$$\xi_r = \ell_{g,r} - a_{e,g,r}, \qquad \mathsf{M}_r = \sum_{s=1}^{r} \xi_s, \qquad C_g = \sum_m \{\lambda_{g,m} + \psi_E(\lambda_{g,m})\}.$$

Since $Y_{e,m,r}, \widehat{Y}_{e,m,r}, \mu_0 \in [0,1]$, we have

$$|\ell_{g,r}| \leq C_g, \qquad |a_{e,g,r}| = |\mathbb{E}[\ell_{g,r} \mid \mathcal{F}_{r-1}]| \leq C_g, \qquad |\xi_r| \leq 2C_g.$$

Here $C_g < \infty$ is $\mathcal{F}_0$-measurable because the grid is finite and $\lambda_{g,m} < 1$ for all m. By (53),

$$\mathbb{E}[\xi_r \mid \mathcal{F}_{r-1}] = \mathbb{E}[\ell_{g,r} \mid \mathcal{F}_{r-1}] - a_{e,g,r} = 0.$$

On $\{C_g = 0\}$, $a_{e,g,r} = 0$ for every r, so $\{C_g = 0\} \subseteq G_g^c$. Thus $C_g > 0$ almost surely under the designs considered here.

The conditional form of Hoeffding's bounded-variable lemma (Hoeffding, 1963), applied to the centered variable $\xi_r \in [-2C_g, 2C_g]$, gives, for each real t,

$$\mathbb{E}[\exp(t\xi_r) \mid \mathcal{F}_{r-1}] \leq \exp(2t^2 C_g^2).$$

Since M_{r-1} is $\mathcal{F}_{r-1}$-measurable, the tower property gives

$$\begin{aligned}\mathbb{E}[\exp(t\mathsf{M}_r) \mid \mathcal{F}_0] &= \mathbb{E}[\exp(t\mathsf{M}_{r-1})\mathbb{E}[\exp(t\xi_r) \mid \mathcal{F}_{r-1}]|\mathcal{F}_0] \\ &\leq \exp(2t^2C_g^2)\mathbb{E}[\exp(t\mathsf{M}_{r-1}) \mid \mathcal{F}_0] \leq \cdots \leq \exp(2rt^2C_g^2),\end{aligned}$$

where $M_0 = 0$. For $\varepsilon > 0$, exponential Markov's inequality and minimization over $t > 0$ give

$$\mathbb{P}(\mathsf{M}_r \geq r\varepsilon \mid \mathcal{F}_0) \leq \inf_{t>0} \exp\{-tr\varepsilon + 2rt^2C_g^2\} = \exp\left\{-\frac{r\varepsilon^2}{8C_g^2}\right\}.$$

The minimizing value is $t = \varepsilon/(4C_g^2)$, with C_g fixed under the conditional law. Applying the same bound to $-\mathsf{M}_r$ and adding the two tail probabilities yields

$$\mathbb{P}(|\mathsf{M}_r| \geq r\varepsilon \mid \mathcal{F}_0) \leq 2\exp\left\{-\frac{r\varepsilon^2}{8C_g^2}\right\}. \tag{54}$$

For every integer $j \geq 1$, (54) implies

$$\begin{aligned}\mathbb{P}\left(\bigcup_{r=n}^{\infty}\{|\mathsf{M}_r| \geq r/j\}\Big|\mathcal{F}_0\right) &\leq 2\sum_{r=n}^{\infty}\exp\left\{-\frac{r}{8j^2C_g^2}\right\} \\ &= \frac{2\exp\{-n/(8j^2C_g^2)\}}{1-\exp\{-1/(8j^2C_g^2)\}} \longrightarrow 0.\end{aligned}$$

The events on the left decrease to $\bigcap_{n\geq 1}\bigcup_{r\geq n}\{|\mathsf{M}_r|\geq r/j\}$, which therefore has conditional probability zero. Taking the countable union over j gives

$$\mathbb{P}\left(\limsup_{r\to\infty}\frac{|\mathsf{M}_r|}{r}>0\middle|\mathcal{F}_0\right)=0,\qquad \frac{\mathsf{M}_r}{r}\longrightarrow 0\quad\text{a.s.}$$

The product definition of the component process now gives

$$\frac{1}{r}\log E^{\mathcal{C}}_{e,r}(\lambda_g)=\frac{1}{r}\sum_{s=1}^{r}a_{e,g,s}+\frac{\mathsf{M}_r}{r}.$$

On $G_g\cap\{\mathsf{M}_r/r\to 0\}$, whose conditional probability is one,

$$d_g:=\liminf_{r\to\infty}\frac{1}{r}\log E^{\mathcal{C}}_{e,r}(\lambda_g)=\liminf_{r\to\infty}\frac{1}{r}\sum_{s=1}^{r}a_{e,g,s}>0.$$

The coefficient $\rho_g>0$ is fixed given $\mathcal{F}_0$, so

$$\frac{1}{r}\log E^{\mathcal{C}}_{e,r}\geq\frac{\log\rho_g}{r}+\frac{1}{r}\log E^{\mathcal{C}}_{e,r}(\lambda_g).$$

Since $(\log\rho_g)/r\to 0$, the last inequality yields

$$\liminf_{r\to\infty}\frac{1}{r}\log E^{\mathcal{C}}_{e,r}\geq d_g>0.$$

For each such outcome, set $\gamma=d_g/2>0$. There exists an integer r_0 such that $E^{\mathcal{C}}_{e,r}\geq\exp(\gamma r)$ for every $r\geq r_0$. Thus $\mathbb{P}(E^{\mathcal{C}}_{e,r}\to\infty\mid\mathcal{F}_0)=1$.

Let $h=|\mathcal{H}|\geq 1$. Whenever $E^{\mathcal{C}}_{e,r}\geq h/\alpha$, the direction e is not in $\mathcal{J}_r$. Thus $|\mathcal{J}_r|\leq h-1$ and

$$c_r=\frac{1}{\alpha}+\sum_{j\in\mathcal{J}_r}\left(\frac{1}{\alpha}-E^{\mathcal{C}}_{j,r}\right)\leq\frac{1}{\alpha}+\frac{h-1}{\alpha}=\frac{h}{\alpha}\leq E^{\mathcal{C}}_{e,r}.$$

Consequently, for every r,

$$\{E^{\mathcal{C}}_{e,r}\geq h/\alpha\}\subseteq\{e\in\mathcal{E}_r\}.$$

Because $h/\alpha<\infty$ is fixed given $\mathcal{F}_0$, divergence gives

$$\mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{r=m}^{\infty}\{e\in\mathcal{E}_r\}\middle|\mathcal{F}_0\right)=1.$$

In particular, $\mathbb{1}\{e\in\mathcal{E}_r\}\to 1$ almost surely. Since these indicators lie in $[0,1]$, conditional bounded convergence gives

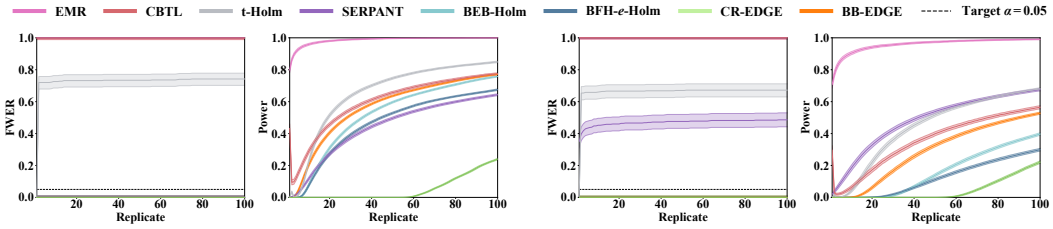
$$\mathbb{P}(e\in\mathcal{E}_r\mid\mathcal{F}_0)=\mathbb{E}[\mathbb{1}\{e\in\mathcal{E}_r\}\mid\mathcal{F}_0]\longrightarrow 1\quad\text{a.s.}$$

Applying the tower property and bounded convergence once more yields

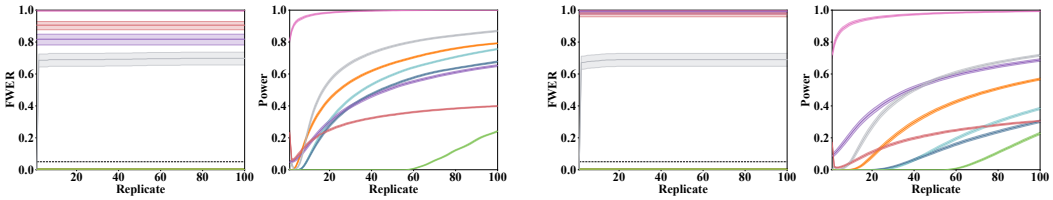
$$\mathbb{P}(e\in\mathcal{E}_r\mid\mathcal{B},\pi)=\mathbb{E}[\mathbb{P}(e\in\mathcal{E}_r\mid\mathcal{F}_0)\mid\mathcal{B},\pi]\longrightarrow 1\quad\text{a.s.},$$
$$\mathbb{P}(e\in\mathcal{E}_r)=\mathbb{E}[\mathbb{P}(e\in\mathcal{E}_r\mid\mathcal{F}_0)]\longrightarrow 1.$$

□

E Ablation Study

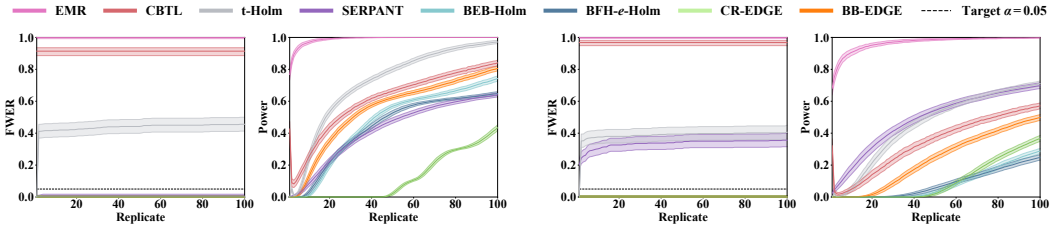


(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$

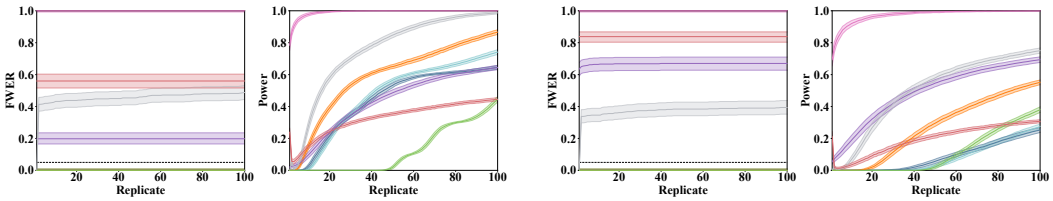


(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

Figure 6: Ablation study on the number of models ($L = 20$).

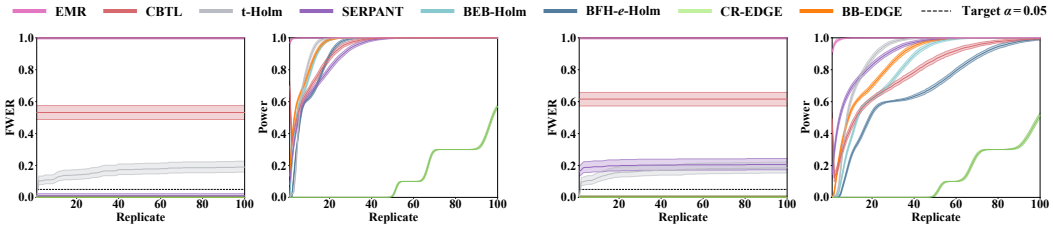


(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$



(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

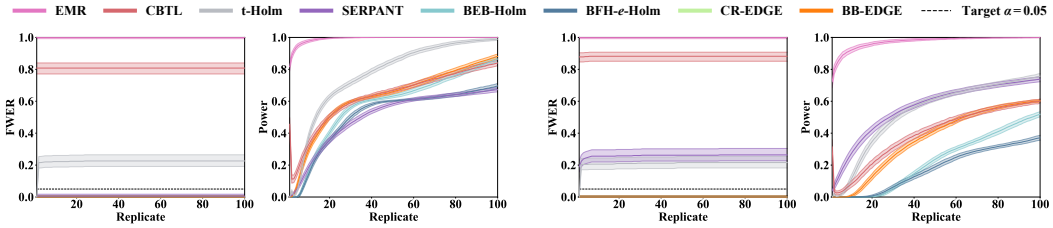
Figure 7: Ablation study on the effect of item size ($N = 50$).



(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$

(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

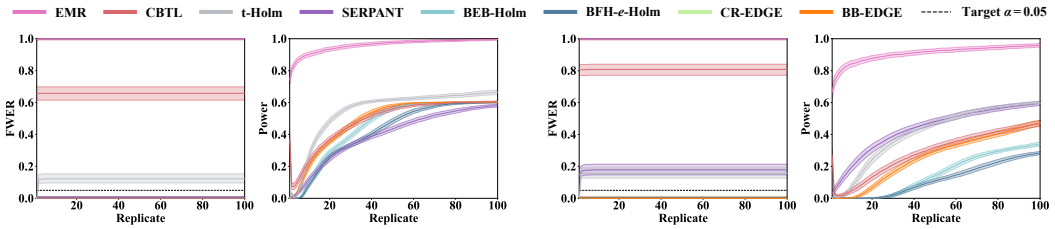
Figure 8: Ablation study on the effect of item size ($N = 500$).



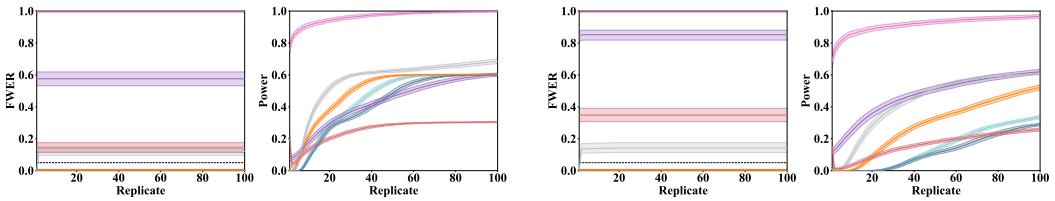
(a) $|C_m| = 1$, $h_{i,\ell} = 0$ (b) $|C_m| = 5$, $h_{i,\ell} = 0$

(c) $|C_m| = 1$, $h_{i,\ell} = 0.2$ (d) $|C_m| = 5$, $h_{i,\ell} = 0.2$

Figure 9: Ablation study on prespecified margins ($\tau = 0.01$).

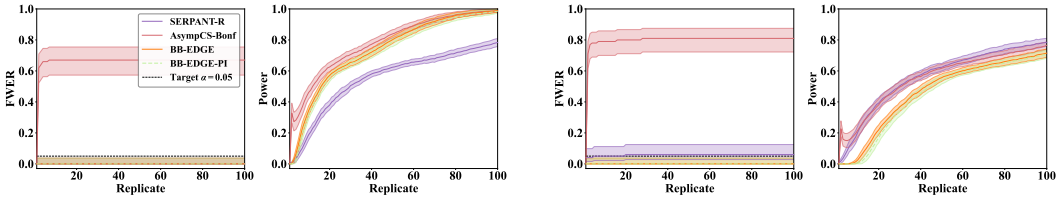


(a) $|C_m| = 1, h_{i,\ell} = 0$ (b) $|C_m| = 5, h_{i,\ell} = 0$

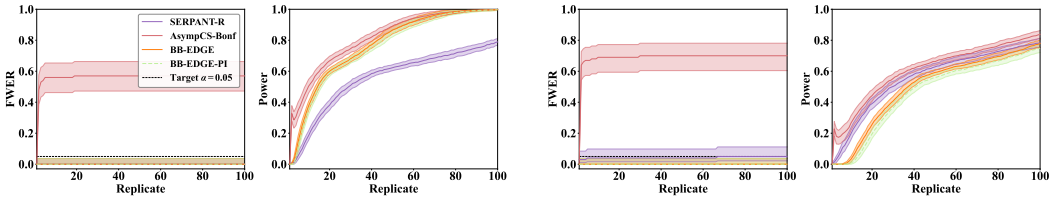


(c) $|C_m| = 1, h_{i,\ell} = 0.2$ (d) $|C_m| = 5, h_{i,\ell} = 0.2$

Figure 10: Ablation study on prespecified margins ($\tau = 0.03$).

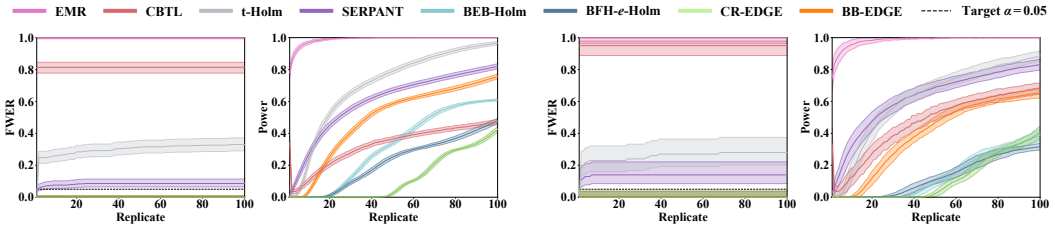


(a) $|C_m| = 1, h_{i,\ell} = 0$ (b) $|C_m| = 5, h_{i,\ell} = 0$



(c) $|C_m| = 1, h_{i,\ell} = 0.2$ (d) $|C_m| = 5, h_{i,\ell} = 0.2$

Figure 11: Ablation study on more baselines: BB-EDGE versus BB-EDGE-PI, AsympCS-Bonf and SERPANT-R.



(a) Heterogeneous Block Correlations (b) Heterogeneous Block Sizes

Figure 12: Ablation study under heterogeneous block structures. (a) The within-block correlation varies uniformly from 0.1 to 0.7 across blocks. (b) The prespecified block sizes vary from 1 to 10.

F Additional Real Data Results

F.1 Leaderboard

Tables 2–5 report the mean benchmark score and its standard deviation for ten LLMs on four real datasets across the 20 replicates. Leaderboards offer an intuitive ranking of models, but a ranking based on point estimates alone cannot separate models whose mean scores are nearly identical. For example, Gemma-4-E4B-It and Qwen3.5-9B score 0.7751 and 0.7790 on ACEBench, a gap smaller than either of their standard deviations across replicates (0.0104 and 0.0085). Consistently, BB-EDGE leaves this pair unresolved at $r = 20$ (Figure 2 (b)). Additionally, Figure 14 also shows that item-level advantages are positive for some items and negative for others despite near-zero benchmark-average differences: $\widehat{\Delta}_e^{\mathcal{B}} = -0.008$ for Gemma-4-E4B-it versus MiMo-7B-RL on GSM8K.

Model	Accuracy
Qwen3.5-9B	0.6436 ± 0.0031
Phi-4	0.6267 ± 0.0043
Granite-4.1-8B	0.6124 ± 0.0028
Falcon-H1R-7B	0.5492 ± 0.0035
Intern-S1-Mini	0.5427 ± 0.0030
MiMo-7B-RL	0.5291 ± 0.0030
Gemma-4-E4B-It	0.5015 ± 0.0029
DeepSeek-R1-Distill-Qwen-7B	0.4728 ± 0.0028
Llama-3.1-8B-Instruct	0.4223 ± 0.0032
Mistral-7B-Instruct-v0.3	0.3527 ± 0.0030

Table 2: Accuracy (mean $\pm$ std) on MMLU-Pro for each model, sorted in descending order of accuracy.

Model	Exact Match
Granite-4.1-8B	0.9138 ± 0.0043
Phi-4	0.9067 ± 0.0065
Qwen3.5-9B	0.8488 ± 0.0079
Intern-S1-Mini	0.7971 ± 0.0076
MiMo-7B-RL	0.7444 ± 0.0074
Gemma-4-E4B-It	0.7362 ± 0.0076
Llama-3.1-8B-Instruct	0.7235 ± 0.0090
DeepSeek-R1-Distill-Qwen-7B	0.7102 ± 0.0091
Mistral-7B-Instruct-v0.3	0.5054 ± 0.0097
Falcon-H1R-7B	0.4765 ± 0.0099

Table 3: Exact Match (mean $\pm$ std) on GSM8K for each model, sorted in descending order of accuracy.

F.2 Full-Ranking Confidence Graph

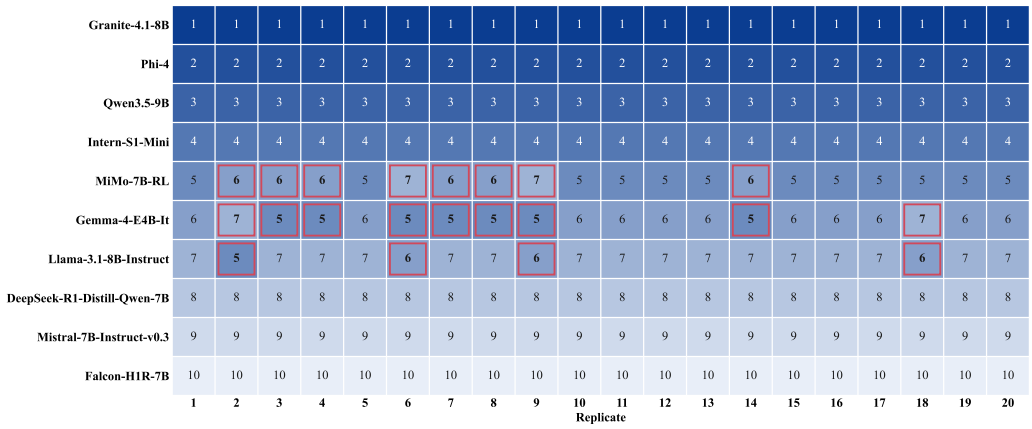
Figure 15 presents the directed acyclic confidence graphs at replicate $r = 20$ for MMLU-Pro and MT-Bench. The MMLU-Pro graph forms a fully resolved linear chain headed by Qwen3.5-9B, followed by Phi-4 and Granite-4.1-8B, with each displayed edge representing a statistically certified directional comparison. The MT-Bench graph is nearly a total order, with only three pairs remaining unresolved after multiplicity correction: Gemma-4-E4B-It versus Qwen3.5-9B, Intern-S1-Mini versus Llama-3.1-8B-Instruct, and Falcon-H1R-7B versus DeepSeek-R1-Distill-Qwen-7B. For the binary-score benchmarks, SERPANT's preference target is equivalent to the mean-score target after fair tie breaking. On MT-Bench, which uses graded judge scores, SERPANT instead targets the induced pairwise-preference ordering and is included only as a reference. Compared with $\tau = 0$, Figure 16 shows that setting $\tau = 0.01$ yields fewer certified edges because each comparison must

Model	LLM Judge
Gemma-4-E4B-It	0.7422 ± 0.0076
Qwen3.5-9B	0.7418 ± 0.0139
Granite-4.1-8B	0.6960 ± 0.0098
Phi-4	0.6614 ± 0.0117
Intern-S1-Mini	0.5945 ± 0.0150
Llama-3.1-8B-Instruct	0.5714 ± 0.0146
Mistral-7B-Instruct-v0.3	0.4961 ± 0.0113
Falcon-H1R-7B	0.4497 ± 0.0141
DeepSeek-R1-Distill-Qwen-7B	0.4233 ± 0.0102
MiMo-7B-RL	0.3824 ± 0.0149

Table 4: LLM judge scores (mean $\pm$ standard deviation) on MT-Bench for each model, sorted in descending order of mean score.

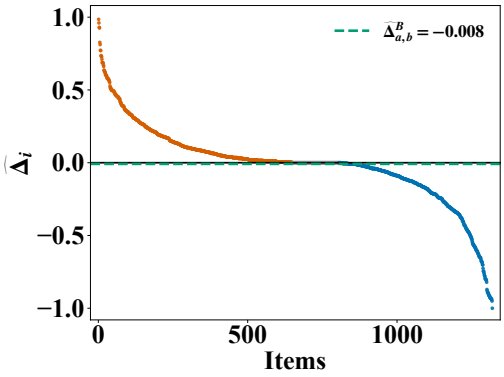
Model	AST-based Accuracy
Qwen3.5-9B	0.7790 ± 0.0104
Gemma-4-E4B-It	0.7751 ± 0.0085
Phi-4	0.7308 ± 0.0093
Intern-S1-Mini	0.7295 ± 0.0140
Granite-4.1-8B	0.6619 ± 0.0083
Mistral-7B-Instruct-v0.3	0.5465 ± 0.0114
Llama-3.1-8B-Instruct	0.4996 ± 0.0167
DeepSeek-R1-Distill-Qwen-7B	0.3769 ± 0.0213
MiMo-7B-RL	0.3700 ± 0.0218
Falcon-H1R-7B	0.2620 ± 0.0388

Table 5: AST-based accuracy (mean $\pm$ standard deviation) on ACEBench for each model, sorted in descending order of mean score.



(a) GSM8K

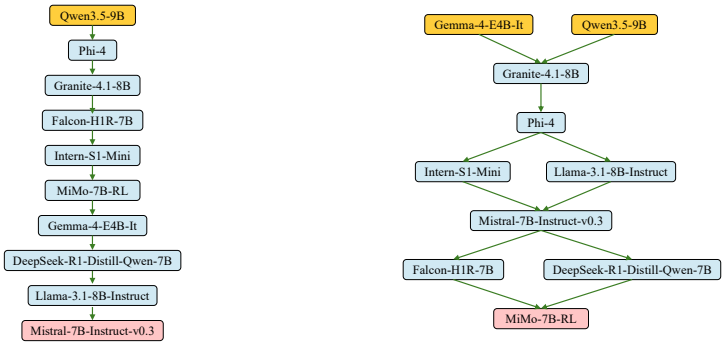
Figure 13: Single-run rankings of ten LLMs on GSM8K over 20 replicates. Rows are ordered by the 20-replicate mean score; outlined cells mark ranks that differ from the 20-replicate ranking.



(a) GSM8K

Figure 14: Item-level heterogeneity on GSM8K. Different items favor different models despite near-zero benchmark-average differences: Gemma-4-E4B-it versus MiMo-7B-RL on GSM8K ($\widehat{\Delta}^{\mathcal{B}}_{\text{Gemma,MiMo}} = -0.008$).

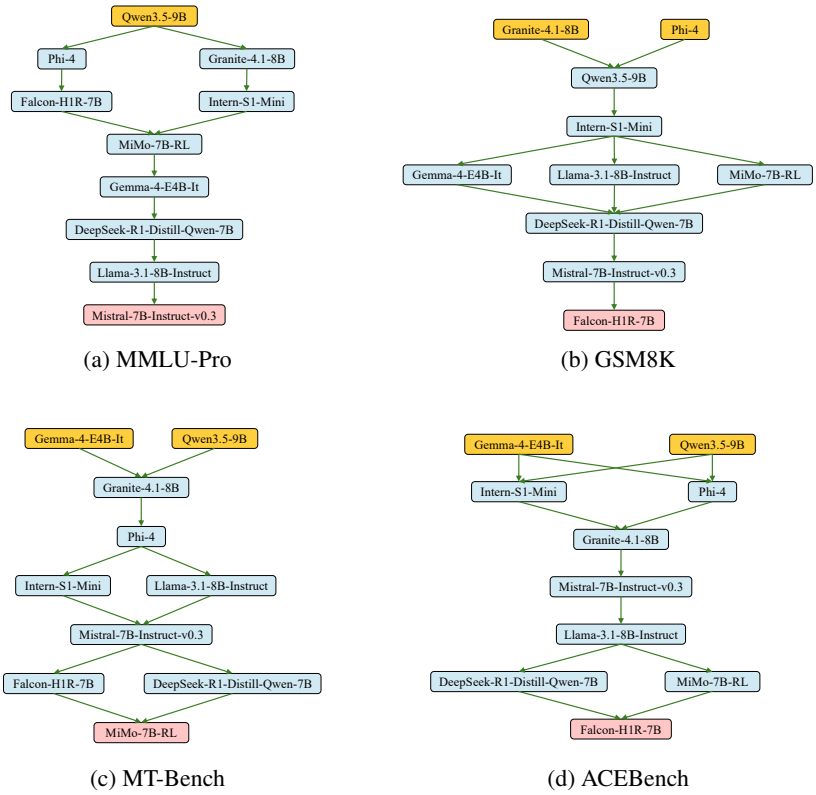
establish an advantage greater than 0.01. For example, increasing τ from 0 to 0.01 changes the fully resolved MMLU-Pro ranking into a partial order, leaving Phi-4 versus Granite-4.1-8B and Falcon-H1R-7B versus Intern-S1-Mini unresolved. Nevertheless, BB-EDGE still identifies clear partial ranking structures across all four benchmarks.



(a) MMLU-Pro (b) MT-Bench

Figure 15: Certified comparison graphs at replicate $r = 20$ on MMLU-Pro and MT-Bench, respectively. An edge $A \to B$ indicates that model A is certified to outperform model B on the benchmark under the evaluation protocol. Pairs not connected by a directed path remain unresolved.

Figure 17 shows the GSM8K confidence graph at $r = 100$ with a single block ($M = 1$), which requires no within-replicate independence. It resolves 24 of 45 comparisons, certifying exactly the pairs with mean gaps above 0.14, whereas the item-factorized graph ($M = N$) resolves 41 after only two replicates. The one-block resolution requires roughly two orders of magnitude more replicates, so we recommend it only as a fallback when the protocol cannot justify conditional independence across blocks.



(a) MMLU-Pro (b) GSM8K

(c) MT-Bench (d) ACEBench

Figure 16: Certified comparison graphs at replicate $r = 20$ on four datasets, with a prespecified margin of $\tau = 0.01$. An edge $A \to B$ indicates that model A is certified to outperform model B on the benchmark under the evaluation protocol. Pairs not connected by a directed path remain unresolved.

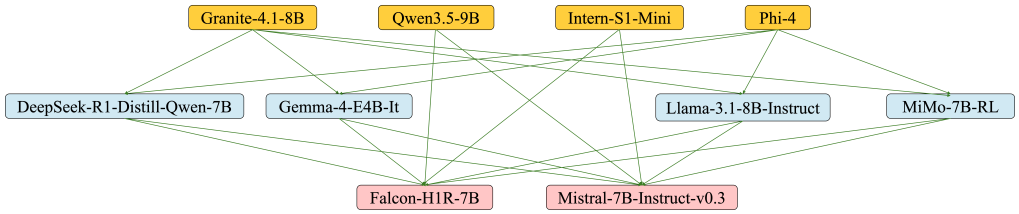


Figure 17: Certified comparison graphs at replicate $r = 100$ on GSM8K under the one-block partition ($M = 1$) with $\tau = 0$. An edge $A \to B$ indicates that model A is certified to outperform model B on the benchmark under the evaluation protocol. Pairs not connected by a directed path remain unresolved.

F.3 TOP-k CERTIFICATION

Table 6 reports the certified Top-4 sets across all four benchmarks. Qwen3.5-9B and Phi-4 appear in every benchmark's Top-4, indicating consistently strong performance across diverse task types, while Granite-4.1-8B is certified in three of the four (all but ACEBench). The remaining slot varies by benchmark: Falcon-H1R-7B on MMLU-Pro, Intern-S1-Mini on GSM8K and ACEBench, and Gemma-4-E4B-It on MT-Bench and ACEBench.

Table 7 reports the minimum number of complete benchmark replicates required for Top-4 certification ($\alpha = 0.05$, $\tau = 0$) on the four benchmarks. The results show that the two BB-EDGE variants always require the fewest replicates. Full-graph BB-EDGE is fastest on MMLU-Pro and GSM8K, certifying the Top-4 set after one and two replicates, respectively. On both benchmarks, the pilot-targeted variant requires one additional replicate for pilot selection, resulting in slightly later certification. The pilot-targeted variant performs better on MT-Bench and ACEBench, where more replicates are otherwise required. For example, on MT-Bench, full-graph BB-EDGE also certifies within 5 replicates, whereas the pilot-targeted variant only requires 3 replicates;

	MMLU-Pro	GSM8K	MT-Bench	ACEBench
Top-4	Falcon-H1R-7B Granite-4.1-8B Phi-4 Qwen3.5-9B	Intern-S1-Mini Granite-4.1-8B Phi-4 Qwen3.5-9B	Gemma-4-E4B-It Granite-4.1-8B Phi-4 Qwen3.5-9B	Gemma-4-E4B-It Intern-S1-Mini Phi-4 Qwen3.5-9B

Table 6: Certified Top-4 model sets on the four benchmarks at $\alpha = 0.05$ and $\tau = 0$.

Method	MMLU-Pro	GSM8K	MT-Bench	ACEBench
SERPANT	**1**	4	8	20+
BFH-e-Holm	2	7	17	17
CR-EDGE	20+	20+	20+	20+
BB-EDGE	**1**	**2**	5	4
+Pilot	2	3	**3**	**3**

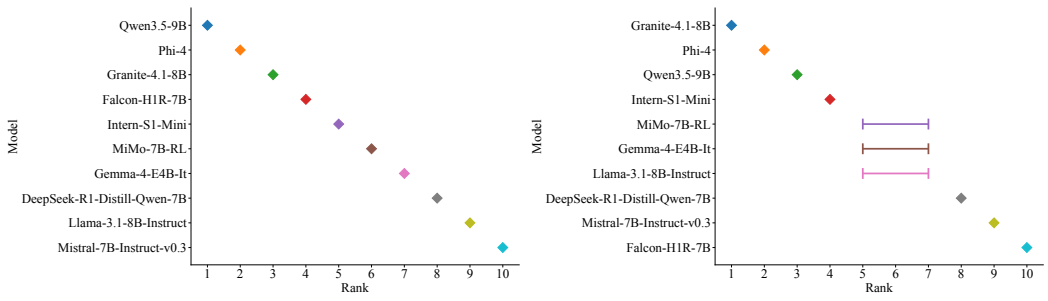
Table 7: Minimum number of complete benchmark replicates required for Top-4 certification at $\alpha = 0.05$ and $\tau = 0$. BB-EDGE and the baselines use full-graph certification; +pilot uses the pilot-targeted procedure with $R_0 = 1$ pilot replicates, which are included in its counts. "20+" indicates that no certificate was issued within the budget of 20 replicates. Bold marks the fewest replicates among methods with anytime FWER control.

F.4 RANK INTERVAL

Figure 18 shows that ranking uncertainty varies across four benchmarks. On MMLU-Pro, BB-EDGE uniquely determines every model's rank. On GSM8K, most ranks are resolved, while Llama-3.1-8B-Instruct, Gemma-4-E4B-it, and MiMo-7B-RL have unresolved ranks within 5–7. On MT-Bench, uncertainty remains at ranks 1–2, 5–6, and 8–9, corresponding to Gemma-4-E4B-it and Qwen3.5-9B, Intern-S1-Mini and Llama-3.1-8B-Instruct, and Falcon-H1R-7B and DeepSeek-R1-Distill-Qwen-7B, respectively. On ACEBench, ranks 1–2, 3–4, and 8–9 remain unresolved, while all other ranks are uniquely determined. These results illustrate how BB-EDGE certifies precise ranks when supported by the evidence while retaining uncertainty in unresolved comparisons.

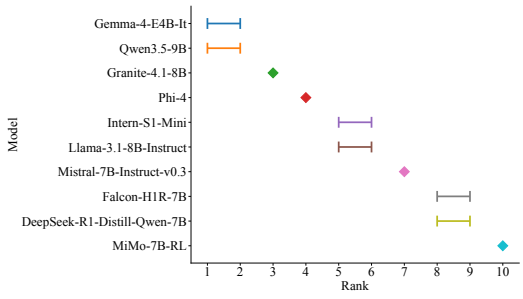
F.5 CLOSED-SOURCE APIS

We further evaluate BB-EDGE on five API models (DeepSeek-V4-Flash (Xu et al., 2026), GLM-5.3-Flash (Zeng et al., 2026), GPT-5.6-Luna (OpenAI, 2026), MiniMax-M3 (Lai et al., 2026), and Qwen3.8-Flash (Qwen Team, 2026b)). Due to limited resources, this evaluation is restricted to ACEBench Normal Multi-Turn. Table 8 reports the mean scores, and Figure 19 presents the directed acyclic graph representing the inferred ranking of closed-source APIs at replicate $r = 20$ on

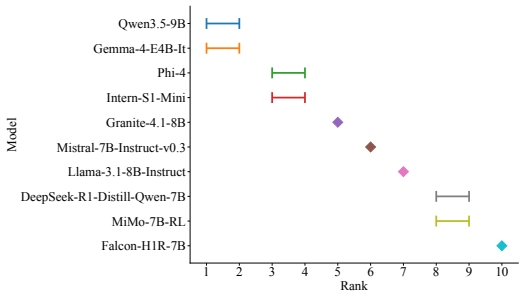


(a) MMLU-Pro

(b) GSM8K



(c) MT-Bench



(d) ACEBench

Figure 18: Rank intervals for LLMs at replicate $r = 20$ on four datasets. Each horizontal segment represents the inferred rank interval of a model. Overlapping intervals indicate unresolved ranking.

ACEBench. GLM-5.3-Flash and Qwen3.8-Flash jointly occupy the top tier, each certified above GPT-5.6-Luna and DeepSeek-V4-Flash, respectively, both of which are in turn certified above MiniMax-M3. The graph exhibits a symmetric two-branch structure that converges to a single lowest-ranked model.

Model	AST-based Accuracy
GLM-5.3-Flash	0.8041 ± 0.0089
Qwen3.8-Flash	0.7982 ± 0.0065
GPT-5.6-Luna	0.7797 ± 0.0087
DeepSeek-V4-Flash	0.7756 ± 0.0080
MiniMax-M3	0.7198 ± 0.0133

Table 8: ACEBench scores (mean $\pm$ standard deviation) for each model, sorted in descending order of mean score.

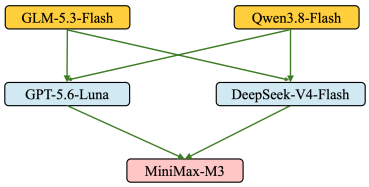


Figure 19: Certified comparison graph representing the inferred ranking of APIs on ACEBench at replicate $r = 20$.

G Experimental Setup Details

G.1 Dataset

We evaluate on four publicly available benchmarks: MMLU-Pro (Wang et al., 2024a), GSM8K (Cobbe et al., 2021), MT-Bench (Zheng et al., 2023), and ACEBench (Chen et al., 2025). Table 9 summarizes the statistics, splits, evaluation metrics, and block sizes $|C_m|$ (the number of scored items per block); the four datasets span a range of block sizes and pool sizes.

MMLU-Pro is an enhanced version of MMLU that expands each question's candidate options from 4 to up to 10, yielding a more challenging and discriminative benchmark. Each MMLU-Pro item is evaluated through a separate model invocation and has no shared conversational context with other items. We therefore use singleton blocks under the working assumption that separate stateless item-level executions are conditionally independent given the evaluation protocol.

GSM8K is a benchmark of grade-school-level math word problems that require multi-step arithmetic reasoning to arrive at a final numeric answer. Each GSM8K problem is evaluated through a separate model invocation without cross-item context. Accordingly, we use singleton blocks under the working assumption that separate item-level executions are conditionally independent given the benchmark and protocol.

MT-Bench is a benchmark designed to evaluate models' multi-turn conversational and instruction-following abilities. It consists of 80 two-turn questions. Following Zheng et al. (2023), the judge rates the two turns separately: the first-turn response is rated given the first question, and the second-turn response is rated given the full two-turn dialogue. Each question therefore contributes two scored items, giving $N = 160$. Ratings on the 1–10 scale are divided by 10. We group the two turns of each dialogue into one block, yielding $M = 80$ dialogue-level blocks of size $|C_m| = 2$. This grouping allows arbitrary dependence between the two turns within each dialogue, while assuming conditional independence across separately executed dialogues given the fixed benchmark and evaluation protocol.

ACEBench is a comprehensive benchmark for evaluating LLM tool usage. We focus on its Normal Multi-Turn subset, which contains $N = 223$ scored tool-call steps, each scored via AST-based function-call matching against the ground truth. Each dialogue contains two or three scored steps. We treat each self-contained dialogue as one block, so the blocks have sizes $|C_m| \in \{2, 3\}$ and weights $w_m = |C_m|/N$. We allow arbitrary dependence among steps within a dialogue and assume conditional independence across separately executed dialogues.

Permutation tests comparing the variance of replicate-level mean scores with that implied by independent blocks find no significant dependence across blocks for any model after Holm correction (smallest raw p: GSM8K 0.073, MMLU-Pro 0.097, MT-Bench 0.14, ACEBench 0.37), supporting the prespecified partitions. For each model, we compare the variance of replicate-level mean scores with the sum of block-level score variances divided by N^2, which is its value under independence across blocks, and obtain a p-value by independently permuting each block's scores across replicates, which preserves within-block dependence while removing any dependence across blocks.

Table 9: The statistics, split and evaluation metrics of each dataset.

Data	Train Set	Test Set	Evaluation	$\lvert C_m \rvert$
MMLU-Pro	70	12,032	Accuracy	1
GSM8K	7,470	1,319	Exact Match	1
MT-Bench	-	160 (80 questions)	LLM Judge*	2
ACEBench	-	223	AST-based Accuracy	2–3

*On MT-Bench, BB-EDGE certifies model comparisons based on the LLM-based scores generated by MiniMax-M3 under the given protocol; the reliability of the protocol itself is beyond the scope of this work.

G.2 BASELINES

EMR is the simplest baseline: it computes the empirical mean performance difference $\widehat{\Delta}_{a,b}$ over the r observed replicates and certifies (a, b) whenever $\widehat{\Delta}_{a,b} > \tau$. It accounts for neither sampling uncertainty nor multiplicity across the H directional comparisons.

CBTL (Wang et al., 2024b) models pairwise comparison probabilities as smooth functions of the item context using kernel smoothing. It certifies (a, b) when the infimum of the fitted log-odds gap exceeds the transformed margin $\log\{(1+\tau)/(1-\tau)\}$, with the test calibrated by a Gaussian multiplier bootstrap over complete replicates.

t-Holm (Chandrahas, 2026) computes a paired one-sided t-test for each directional null $H_0 : \Delta^{\mathcal{B}}_{a,b} \leq \tau$ using the complete-replicate mean differences. It then applies Holm's step-down procedure to the resulting p-values. This procedure provides fixed-time FWER control under its working assumptions, but repeated monitoring is not covered.

SERPANT (Gu et al., 2026) is designed for online pairwise comparisons under an i.i.d. item stream. We convert paired item scores into binary preferences, breaking ties with complementary fair coins. Items are processed in replicate-by-item order; after every item, we update each direction's uniform-mixture likelihood-ratio e-process and apply direct e-Holm, with absorbing rejections. Thus, SERPANT has been monitored at all rN item-level times. Under a heterogeneous benchmark-average null, intermediate item prefixes need not satisfy the null even when the complete benchmark does, so this item-level procedure need not control the target anytime FWER. This transformation preserves the benchmark mean-difference hypothesis for binary scores, since the fair-tie preference probability equals $(1 + \Delta^{\mathcal{B}}_{a,b})/2$. For the graded MT-Bench scores, this equivalence need not hold, so SERPANT is included there only as a preference-based reference. We use this adaptation primarily to diagnose the effect of item-level monitoring on fixed-benchmark inference; its certification times are descriptive when its item-stream assumptions or estimand do not match the benchmark-average target.

BEB-Holm is a fixed-time empirical-Bernstein baseline for the same benchmark-average directional hypotheses. For the equal-sized blocks used in the synthetic experiments, it treats the $n = rM$ block differences $D^e_{m,s} = \frac{1}{|C_m|}\sum_{i\in C_m}\left(S^{(a)}_{i,s} - S^{(b)}_{i,s}\right)$ with $s \leq r$ and $m \in [M]$ as independent bounded observations. Let $\overline{D}^e_r$ and $\widehat{V}^e_r$ denote their sample mean and variance. We define $L^e_r(\delta) = \overline{D}^e_r - \sqrt{\frac{2\widehat{V}^e_r \log(4/\delta)}{n}} - \frac{16\log(4/\delta)}{3(n-1)}$ and obtain the directional p-value by inverting $L^e_r(\delta) > \tau$. Holm's procedure is then applied across all directions. BEB-Holm is recomputed at each r, but its guarantee is fixed-time and does not cover repeated monitoring.

BFH-e-Holm uses the same block partition, proportional stakes, stake grid, mixture weights, and direct e-Holm procedure as BB-EDGE, replacing only the empirical-Bernstein penalty with a Hoeffding penalty. This comparison isolates the benefit of variance adaptation.

CR-EDGE uses the same empirical-Bernstein construction, stake grid, mixture weights, and direct e-Holm procedure as BB-EDGE, but treats each complete benchmark replicate as a single block. This comparison isolates the efficiency gain from block factorization.

BB-EDGE-PI uses the same block observations, proportional stakes, predictions, and direct e-Holm procedure as BB-EDGE, but replaces the grid mixture with a single predictable plug-in stake $\lambda_r = \min\{0.95,\ 0.5\,\hat{g}_{r-1}/\hat{V}_{r-1}\}$, where $\hat{g}_{r-1} = \max\{0, \hat{\mu}_{r-1} - \mu_0\}$ and $\hat{V}_{r-1}$ are regularized running estimates of the mean gap and the residual variance. This isolates sensitivity to stake selection.

AsympCS-Bonf applies the one-sided Gaussian-mixture asymptotic confidence sequence to the replicate-level mean differences of each direction, using the sample standard deviation across replicates, $t^\star = 10$, and a Bonferroni-corrected level α/H. No certificate is issued before a burn-in of two replicates. Its time-uniform coverage holds only asymptotically, so it tests whether a variance-adaptive but non-factorized procedure suffices in the small-budget regime.

SERPANT-R converts paired item scores into binary preferences as in SERPANT, pools the preferences from all items of the first r completed replicates, and computes the same uniform-mixture likelihood-ratio e-values, which treat items as i.i.d. Bernoulli observations. Direct e-Holm is applied only at completed replicates. This tests whether restricting monitoring to replicate boundaries alone removes the false certifications of item-level monitoring.

G.3 Evaluation Metric

Top-k certificate events Fix $1 \le k < L$. Let $\mathcal{S}_s$ be the collection of sets certified by the procedure at completed run s. In full-graph certification,

$$\mathcal{S}_s = \{T \subseteq \mathcal{V} : |T| = k,\ a \rightsquigarrow_s b \text{ for every } a \in T,\ b \notin T\}.$$

In the pilot-targeted procedure, $\mathcal{S}_s = \{\mathcal{T}_0\}$ if all hypotheses in $\mathcal{H}(\mathcal{T}_0)$ are rejected, and $\mathcal{S}_s = \varnothing$ otherwise. These are different testing procedures and must be identified separately in an experiment. If evaluation stops at the first certificate, set $\mathcal{S}_s = \varnothing$ after that stopping time. Define

$$\begin{aligned} C_{\le r} &= \{\exists s \le r : \mathcal{S}_s \neq \varnothing\}, \\ A_{\le r} &= \left\{\exists s \le r,\ \exists T \in \mathcal{S}_s : \min_{a \in T} \theta_a^{\mathcal{B}} \le \max_{b \notin T} \theta_b^{\mathcal{B}} + \tau\right\}. \end{aligned} \tag{55}$$

Thus $C_{\le r}$ records any issuance and $A_{\le r} \subseteq C_{\le r}$ records any incorrect issuance by replicate R. The error definition includes a failure to attain the required margin, even when the selected set contains the k largest means. It also handles ties at the true Top-k boundary without choosing an arbitrary true set.

FCP_r The false Top-k certification probability is

$$\mathrm{FCP}_r = \mathbb{P}(A_{\le r} \mid \mathcal{B}, \pi). \tag{56}$$

On the no-false-edge event, every reported path from a to b satisfies $\Delta_{a,b}^{\mathcal{B}} > \tau$. Consequently,

$$\mathcal{A}_{\le r} \subseteq \left\{\exists s \le r,\ \exists (a \to b) \in \mathcal{E}_s : \Delta_{a,b}^{\mathcal{B}} \le \tau\right\}.$$

For full-graph BB-EDGE, Theorem 1 therefore gives $\mathrm{FCP}_r \le \mathrm{FWER}_r \le \alpha$. For pilot-targeted BB-EDGE, $A_{\le r} = C_{\le r} \cap B_0$, where B_0 is the candidate's margin-failure event in Appendix D.4. Theorem 2 gives the same bound, both conditional on the pilot and after averaging over it.

$\mathrm{CertProb}_r$ $\mathrm{CertProb}_r$ measures the probability of issuing any Top-k certificate at replicate r:

$$\mathrm{CertProb}_r = \mathbb{P}(C_{\le r} \mid \mathcal{B}, \pi).$$

Monte Carlo estimation. For B_{MC} independent simulation repetitions under the same setting, let $I_b(r) = \mathbb{1}\{A_{\le r} \text{ occurs in repetition } b\}$ and $J_b(r) = \mathbb{1}\{C_{\le r} \text{ occurs in repetition } b\}$. Compute

$$\widehat{\mathrm{FCP}}_r = \frac{\sum_{b=1}^{B_{\mathrm{MC}}} I_b(r)}{B_{\mathrm{MC}}}, \qquad \widehat{\mathrm{CertProb}}_r = \frac{\sum_{b=1}^{B_{\mathrm{MC}}} J_b(r)}{B_{\mathrm{MC}}}.$$

Both denominators include uncertified repetitions. The optional conditional estimate is $\sum_b I_b(r) / \sum_b J_b(r)$ when $\sum_b J_b(r) > 0$, and is reported as undefined otherwise. Report the event counts and pointwise 95% binomial Monte Carlo confidence intervals for FCP and certification probability. These intervals quantify simulation uncertainty, not a second anytime guarantee.

CR. Under the specified certification procedure, define

$$\mathrm{CR} = \inf\{s \in \{1, \dots, R\} : \mathcal{S}_s \neq \varnothing\}, \qquad \inf \varnothing = +\infty.$$

Then $C_{\le r} = \{\mathrm{CR} \le r\}$ for $r \le R$. For finite-budget experimental summaries, we censor CR at R: repetitions without a certificate are retained and recorded as $\mathrm{CR} = R$. For the pilot-targeted procedure, the reported value includes the R_0 pilot replicate, and an uncertified repetition is recorded as $\mathrm{CR} = R + R_0$. We report $\mathrm{CertProb}_R$ alongside the mean CR to distinguish successful certification from censoring. The full-graph procedure searches all size-k sets, whereas the pilot-targeted procedure checks only the fixed candidate $\mathcal{T}_0$.

$\mathrm{Coverage}_r$. Let $\mathcal{I}_{\ell,s} = [L_{\ell,s}, U_{\ell,s}]$ denote the rank interval of model ℓ after replicate s. The anytime coverage through replicate r is defined as

$$\mathrm{Coverage}_r = \mathbb{P}\left(\forall s \le r,\ \forall \ell \in \mathcal{V} : \mathrm{rank}_B(\ell) \in \mathcal{I}_{\ell,s} \mid B, \pi\right).$$

Width_r. The mean rank-interval width at replicate r is defined as

$$\mathrm{Width}_r = \mathbb{E}\left[\frac{1}{|\mathcal{V}|}\sum_{\ell\in\mathcal{V}}(U_{\ell,r} - L_{\ell,r}) \,\middle|\, B, \pi\right].$$

Smaller values of Width_r indicate that the rank interval contains fewer candidate models.

G.4 Experiment details

For simulated experiments, the five benchmark-level mean values are equally spaced over $[0.41, 0.59]$, each assigned to a tied pair of models; model labels are shuffled once. Item offsets are fixed and equally spaced over $[-0.08, 0.08]$. Within each tied pair, one model has item effects $+H$ on the first half of the benchmark and $-H$ on the second half, while the other has the opposite effects, with $H \in \{0, 0.2\}$. These effects remain fixed across replicates; Gaussian factors are independently regenerated for each replicate and Monte Carlo repetition. Probabilities outside $[0, 1]$ are rejected rather than clipped.

We ran locally hosted models on an NVIDIA L40 GPU with temperature 0.7, top-p 0.95, and a maximum generation length of 4000 tokens. Model-generation and judge calls were executed separately and sequentially using independent stateless requests and random seeds, with no randomness, state, context, batching, or judge output shared across items or models. For GSM8K and MMLU-Pro, we adopt the evaluation protocol from the lm-evaluation-harness (Gao et al., 2024). For MT-Bench, we follow Zheng et al. (2023), and for ACEBench, we follow Chen et al. (2025). Each evaluation block was processed through a separate stateless request. We fixed decoding configurations, randomized block order within each replicate, and used independent seeds. These design choices support, but do not by themselves establish, the conditional independence assumptions. For BB-EDGE, we used 41 equally spaced stakes over $[0, 0.95]$ with uniform mixture weights and initialized each block-specific running-mean prediction at $\mu_0 = (1 + \tau)/2$.

G.5 Query template

Table 10: Query template for the multi-task knowledge and reasoning benchmark (MMLU-Pro).

```
<initial_instruction for subject: $\{$subject$\}$>

... (4 more few-shot exemplars omitted for brevity) ...

Question:
<question>
Options:
A. <option_1>
B. <option_2>
...
Answer: Let's think step by step.
```

Table 11: Query template for the math reasoning task (GSM8K).

```
... (4 more few-shot exemplars omitted for brevity) ...

Question:
<question>
Answer: Let's think step by step.
```

Table 12: Query template for the multi-turn dialogues task (MT-Bench).

```
Turn 1:
User: <question_turn_1>
Assistant: <model response>

Turn 2:
User: <question_turn_2>
Assistant: <model response>
```

Table 13: Judge prompt structure for MT-Bench single-answer grading.

```
[System]: You are a helpful assistant.

[Instruction]
Act as an impartial judge and evaluate the quality of the AI assistant
    's response, considering helpfulness, relevance, accuracy, depth,
    creativity, and level of detail.
(For math/reasoning/coding categories: additionally compare against
    the provided reference answer and identify any mistakes.)
Provide a brief explanation, then output a rating from 1 to 10 in the
    strict format "[[rating]]".

--- Turn 1 grading (rate the first-turn response) ---
[Question]
<question_turn_1>

[The Start of Assistant's Answer]
<answer_turn_1>
[The End of Assistant's Answer]

--- Turn 2 grading (rate only the second-turn response, given the full
     dialogue) ---
[The Start of Assistant's Conversation with User]
### User: <question_turn_1>
### Assistant: <answer_turn_1>
### User: <question_turn_2>
### Assistant: <answer_turn_2>
[The End of Assistant's Conversation with User]
```

Table 14: Query template for the function-calling task (ACEBench).

```
You are an AI assistant with the role name "assistant." Based on the
    provided API specifications and conversation history from steps 1
    to t, generate the API requests that the assistant should call in
    step t+1. The API requests should be output in the format [ApiName
    (key1='value1', key2='value2', ...)], replacing ApiName with the
    actual API name, key1, key2, etc., with the actual parameter names
    , and value1, value2, etc., with the actual parameter values. The
    output should start with a square bracket "[" and end with a
    square bracket "]".
If there are multiple API requests, separate them with commas, for
    example: [ApiName(key1='value1', ...), ApiName(key1='value1', ...)
    , ...]. Do not include any other explanations, prompts, or API
    call results in the output.
... (formatting and parameter rules omitted for brevity) ...

Role Descriptions:
user: User
assistant: The AI assistant role that makes API requests
tool: Provides the results returned from tool calls

API Specifications:
<function>

Conversation history 1..t:
<question>
```