Efficient Evaluation of LLM Performance with Statistical Guarantees

Skyler Wu
Department of Statistics
Stanford University
skylerw@stanford.edu

Yash Nair
Department of Statistics
Stanford University
yashnair@stanford.edu

Emmanuel J. Candès
Department of Statistics
Stanford University
candes@stanford.edu

Abstract

Exhaustively evaluating many large language models (LLMs) on a large suite of benchmarks is expensive. We cast benchmarking as finite-population inference and, under a fixed query budget, seek tight confidence intervals (CIs) for model accuracy with asymptotic frequentist coverage. We propose *Factorized Active Querying* (FAQ), which (a) leverages historical information through a Bayesian factor model; (b) adaptively selects questions using a hybrid variance-reduction/active-learning sampling policy; and (c) maintains validity through *Pro-Active Inference*—a finite-population extension of active inference [Zrnic and Candès, 2024] that enables direct question selection while preserving asymptotic frequentist coverage. In our experiments, FAQ delivers up to $5\times$ effective sample size gains over strong baselines on two benchmark suites, across varying historical-data missingness levels, with negligible computational overhead: this means that it matches the CI width of uniform sampling while using up to $5\times$ fewer queries. We release our source code and our curated datasets to support reproducibility and future research.

1 Introduction

Motivation Large language models (LLMs) are rapidly moving from research demos to deployed systems in domains such as healthcare, law, and education [Chen et al., 2024, Li et al., 2023, Dehghani et al., 2025, Kasneci et al., 2023, Artsi et al., 2025, Achiam et al., 2023]. As adoption grows, so does the number of candidate models organizations must evaluate and monitor. This scale is visible in public ecosystems: HuggingFace hosts at least 1.86M models, including 300K+ tagged as text-generation—many of them fine-tuned variants of certain base models [Laufer et al., 2025]. Indeed, private institutions may maintain hundreds to thousands of internal variants, each exhibiting their own query-dependent strengths and weaknesses. As post-training becomes cheaper, the number of viable candidate models per deployment decision will only grow.

A common practice is to evaluate candidate models' performances on finite, curated benchmark suites. Such suites increasingly include questions that (a) require expensive domain-expert adjudication [Szymanski et al., 2025, Li et al., 2025, Enguehard et al., 2025, Croxford et al., 2025]; (b) involve multi-turn interactions (e.g., helpfulness and safety) [He et al., 2024, Kwan et al., 2024]; and/or (c) demand long-form reasoning with verification [Bai et al., 2024, Ding et al., 2025]. With tens of thousands of questions (e.g., 12K on MMLU-Pro [Wang et al., 2024]), and hundreds to thousands of candidate models, exhaustively evaluating every model on every question is expensive in API costs,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

runtime, and expert annotation [Zhou et al., 2025].[1] In practice, organizations may instead resort to "vibes"-based judgments on ad-hoc subsets [Mazzoni and Sharma, 2025].

Meanwhile, many organizations have access to *historical performance data*—from prior benchmark runs (e.g., the Open LLM Leaderboard [Fourrier et al., 2024]) and/or production evaluation pipelines. As we show, this data can be leveraged to make model evaluation substantially faster and cheaper.

Problem description Consider a benchmark suite of N_q evaluation units—questions, tasks, or trajectories—viewed as a fixed finite population. For a new model, let $z_j \in \{0, 1\}$ indicate whether it succeeds on unit j. We wish to estimate the *model's accuracy* over the entire finite evaluation bank: $\theta := \frac{1}{N_q} \sum_{j=1}^{N_q} z_j \in [0, 1]$.

We also assume access to historical data $H \in \{0, 1, \text{NA}\}^{N_{\text{old}} \times N_q}$ from N_{old} prior models (e.g., previously benchmarked and/or observed in deployment), potentially with many missing entries (i.e., NA) due to resource and deployment constraints. Here, H_{ij} records historical model i's outcome on question j: 1 for correct, 0 for incorrect, and NA for unobserved. Given a new model and budget n_b, the problem is: how should we adaptively select which n_b questions to query so as to construct a confidence interval (CI) for θ that is as narrow as possible, while maintaining asymptotically frequentist coverage?

Here, coverage (say, 95%) means that over repeated runs of the evaluation procedure (i.e., over sampling randomness), the reported 95% CI contains the true finite-bank accuracy θ at least 95% of the time. This guarantee is crucial for model selection and monitoring: miscalibrated uncertainty can lead to systematic overconfidence and poor deployment decisions, and the risk is amplified under adaptive evaluation where i.i.d. assumptions and heuristic error bars may fail.

Our contributions In this paper, we introduce *Factorized Active Querying* (FAQ), an efficient LLM evaluation method that, in our experiments, produces CIs of comparable width using up to $5\times$ fewer queries than competitive baselines (including uniform sampling), with little to no additional computational cost, while maintaining asymptotic frequentist coverage. FAQ has three components (see Figure 1 for a full schematic):

(a) A Bayesian factor model that extracts information from partially-observed historical data.
(b) A hybrid sampling policy combining variance reduction and active learning.
(c) An asymptotic-coverage-preserving finite-population extension of active inference [Zrnic and Candès, 2024], which we call Pro-Active Inference (PAI). Unlike active inference, PAI directly selects the next question to query in a data-adaptive manner, rather than rigidly proceeding according to a fixed ordering and only making label/skip decisions.

FAQ is designed for repeated evaluation settings, where organizations must compare many candidate or deployed model variants on massive fixed benchmark suites. One group of intended users would be frontier labs monitoring checkpoints (e.g, fine-tunes, distillations, and/or quantized variants). However, a much larger set of organizations whom FAQ is designed for are those performing evaluations who might not be the same ones conducting large-scale pre/post-training. Examples include domain-specific teams (e.g., legal, medical, or scientific organizations), benchmark creators and leaderboard services, and safety/governance teams evaluating many candidate models via paid APIs. In such settings, repeated model evaluation itself can be a meaningful bottleneck in cost (e.g., one evaluation of HELM can cost $9,337 in API credits, see [Li et al., 2024]), latency, and throughput, especially when tasks require multi-turn interactions, simulation/tool use, long-context reasoning, and/or expensive expert verification (e.g., doctors or lawyers).

We evaluate FAQ on two benchmark suites under varying historical-data missingness and quantify the additional gains attributable to PAI. We release our compiled and curated datasets (with 4.4K+ models and 21.5K+ questions); datasets and code are available at `https://github.com/skbwu/efficiently-evaluating-llms`.

[1] Although we use "questions" throughout for readability, the same finite-bank formulation applies to any suites of interactive tasks or trajectories that report task-level success indicators.

2 Related Work

Probabilistic factor modeling FAQ is inspired by probabilistic factor modeling [Mnih and Salakhutdinov, 2007, Collins et al., 2001, Chen et al., 2023], which assumes a (Bayesian) generative factor model for a partially-observed binary matrix—here, the matrix of model-question correctness outcomes—and learns low-dimensional latent factors from the data. Joint-estimation approaches include [Chen et al., 2019] and [Cui and Xu, 2026], which impose additional constraints to support theoretical analysis of latent factor models. Relatedly, Cai et al. [2025] develop a family-aware latent-variable model for aggregate LLM benchmark performance, incorporating model metadata and benchmark-specific loadings. Importantly, any uncertainty guarantees for these approaches are tied to correct specification of the factor model. In contrast, FAQ uses the Bayesian factor model only as a *mental model* to guide adaptive query selection, while delivering asymptotically valid frequentist coverage regardless of factor-model fidelity. If the data *is* well-described by the factor model, FAQ automatically adapts and delivers tighter CIs.

Active statistical inference For model-free asymptotic frequentist coverage, FAQ builds on active statistical inference [Zrnic and Candès, 2024, Gligorić et al., 2025], which uses machine learning predictions over unlabeled covariates to adaptively select labels under a fixed budget and construct CIs. Our benchmark setting, however, provides no explicit covariates, so we use the factor model to construct latent features to guide adaptive query selection. Moreover, unlike sequential active inference—which, in a pre-specified "nature's order" stream, can only make label/skip decisions for the next item and enforces the budget constraints only in expectation—PAI/FAQ directly selects the next question from the full benchmark bank using all available information and enforces the query budget exactly.[2]

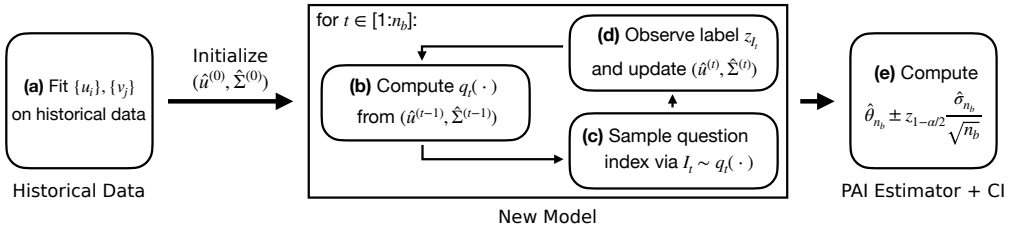

Figure 1: **High-level overview of Factorized Active Querying (FAQ). (a)** Using partially-observed historical outcomes H, we fit historical model and question factors (Eq. 6). For a new model, we initialize its latent factor from the historical fit. Then, at each sampling round t, we **(b)** compute the hybrid sampling policy $q_t(\cdot)$ from the current factor posterior (Eq. 4), **(c)** sample a question $I_t \sim q_t(\cdot)$, and **(d)** update the model factor using the observed outcome z_{I_t} (Eqs. 7-8). Finally, we **(e)** output the PAI estimate and confidence interval (Eqs. 2, 9).

Active learning FAQ draws on Bayesian active learning ideas [c.f. e.g., Settles, 2009, Di Fiore et al., 2023, MacKay, 1992, Fedorov, 2013, Schein and Ungar, 2007] to guide question selection. One component of our hybrid policy upweights questions predicted (under the factor model) to yield larger posterior variance reductions for θ. Because our goal is tight *frequentist* uncertainty quantification of θ, we combine this active-learning component with an oracle-inspired variance-reduction component.

LLM evaluations FAQ contributes to a growing literature on evaluating LLMs using only a fraction of benchmark questions [Hardt, 2025, Zhang et al., 2025, Vivek et al., 2024, Polo et al., 2024, Zhong et al., 2025, Li et al., 2024]. However, most of this line of work targets point estimation of accuracy: with the exception of a simple AIPW estimator considered in Zhang et al. [2025], *none of these prior works provide any model-free guarantees—on point estimation or coverage.* For example, Li et al. [2024] assume a probabilistic model of response correctness given questions, imposing much stronger assumptions on the benchmark/outcome-generating process than our fixed-bank, model-free setup. Relatedly, Zouhar et al. [2025], Zhou et al. [2025], Perlitz et al. [2024] focus on model ranking/selection under limited queries, which is orthogonal to our goal of estimating model accuracy

[2]While preparing this camera-ready manuscript, we became aware of Hamilton et al. [2026]'s closely related Active Measurement (AM) framework, which also uses online-updated ML predictions to guide adaptive sampling for finite-population inference. Unlike AM, however, FAQ additionally incorporates its online-updated factor-model predictions into PAI's regression adjustment, providing another mechanism for variance reduction beyond adaptive query selection.

with rigorous uncertainty quantification. Perlitz et al. [2024] additionally study benchmark design, which is very different from our goal. Closest in spirit to our work is Angelopoulos et al. [2025], which uses active inference to reduce the cost of rating LLM outputs; by contrast, we reduce the number of inference calls per model. More importantly, Angelopoulos et al. [2025]'s approach posits an i.i.d. model for the benchmark data and targets the *expected* accuracy as judged by a strong autorater, whereas we estimate the *true, finite-bank* accuracy of each LLM without i.i.d. assumptions. This distinction is practically important: new models arrive over time and distribution shift is the norm, potentially breaking the i.i.d. assumption. By targeting finite-bank accuracy, FAQ provides uncertainty quantification that remains meaningful over such drifts. Relatedly, Cowen-Breen et al. [2026] develop a statistical framework for LLM evaluation leveraging outputs from multiple cheaper models, but do not address adaptive question selection under a fixed evaluation budget. Separately, Blackwell et al. [2024] document run-to-run variability in LLM evaluation even at zero temperature and with a fixed random seed, highlighting another source of uncertainty in benchmark evaluation.

Because the above approaches generally target different estimands/tasks—point estimation, ranking/selection, benchmark design, or superpopulation accuracy—and generally lack model-free coverage guarantees for each model's finite-bank accuracy, they are not comparable inferential baselines: as such, our empirical baselines (c.f. Section 4) comprise procedures that do have such guarantees.

3 Method

3.1 Factor Model: Extracting Information from History

To run FAQ, we require features for models and questions, but the partially-observed historical outcome matrix H provides no explicit covariates. Thus, we introduce latent factors—each historical model i has $u_i \in \mathbb{R}^k$ and each question j has $v_j \in \mathbb{R}^k$. We model the likelihood that model i answers question j correctly via a logistic link, where $\sigma(\cdot)$ is the sigmoid function:

$$P\left(H_{ij} = 1 \mid \{u_{i'}\}_{i'=1}^{N_{\text{old}}}, \{v_{j'}\}_{j'=1}^{N_q}\right) = \sigma\left(u_i^\top v_j\right). \tag{1}$$

This model is natural for heterogeneous benchmarks, where performance is driven by a small number of shared capabilities (e.g., coding, reasoning, and domain knowledge): u_i encodes model i's latent proficiencies, v_j the capabilities required by question j, and $u_i^\top v_j$ their match. The resulting probabilistic predictions provide uncertainty over unqueried model-question pairs, which we exploit in our hybrid sampling policy. We emphasize that the factor model only serves to improve efficiency: our FAQ coverage guarantees are model-free and do not depend on it.

Fitting on partially-observed historical data We jointly fit the historical model and question factors on the observed entries of H by minimizing the masked negative log-likelihood via AdamW with weight decay λ: (λ, k) are selected via cross-validation (see Section 4). The full objective is given in Appendix A.2.

Deploying to evaluate a new model At deployment, we fix the historically-learned question factors $\{v_j\}_{j=1}^{N_q}$ as features, since questions do not change across models. For a new model, we initialize $u \sim \mathcal{N}(\hat{u}^{(0)}, \hat{\Sigma}^{(0)})$ from the empirical mean and covariance of historical model factors $\{u_i\}_{i=1}^{N_{\text{old}}}$. The factor model is only a working approximation for querying and prediction, *not* for Bayesian credible intervals: we make no fidelity assumptions. Asymptotic frequentist validity is ensured by PAI (Section 3.2).

Online Laplace updates At round t, after observing z_{I_t}, we update $u \mid \mathcal{F}_t \approx \mathcal{N}(\hat{u}^{(t)}, \hat{\Sigma}^{(t)})$ via $\hat{\Sigma}^{(t)}, \hat{u}^{(t)} \leftarrow \texttt{ModelFactorUpdate}(z_{I_t}, v_{I_t}, \hat{u}^{(t-1)}, \hat{\Sigma}^{(t-1)})$, a computable Laplace update derived and specified in Appendices A.4 and A.3.1. This refines predictions $\hat{p}_j^{(t)} = \sigma(\hat{u}^{(t)\top} v_j)$. We adopt this Gaussian approximation for computational tractability, enabling fast, closed-form updates of the new model's latent-factor mean and covariance, which guide our active-learning sampling policy.

3.2 Pro-Active Inference

Motivation and desiderata The factor model is only a working model and offers no coverage guarantee. We also make no i.i.d. assumptions on benchmark questions: labels z_j are fixed-but-unknown, since curated suites often include adversarial/edge cases and need not represent a superpopulation

of user queries. Thus, we need a *model-free* inference layer that wraps around the factor model and yields valid coverage for the finite-bank estimand θ.

Sequential active inference [Zrnic and Candès, 2024] provides coverage, but follows *nature's order*: at time t, one observes only the next item in a fixed stream and can only make a label/skip decision via an adaptive Bernoulli policy. If an informative item is skipped, it cannot be revisited. In our benchmark-bank setting, by contrast, we have access to the *full* pool of question factors at each step and wish to *proactively select* the next question, labeling each selected question deterministically.

These desiderata are important for FAQ because the new model factor u is initialized from historical models and may be inaccurate as models evolve. Proactively selecting informative questions early accelerates learning of u (Section 3.4), potentially improving subsequent query choices and inference. Practically, our desiderata also better match typical evaluation workflows: selecting a budgeted subset of questions, rather than streaming through the entire bank with probabilistic labeling decisions.

Pro-Active Inference estimator Fix a query budget n_b. At each round $t = 1, \ldots, n_b$, we sample a question $I_t \in \{1, \ldots, N_q\}$ from an adaptive distribution $q_t(\cdot)$,[3] query the new model, and observe label $z_{I_t} \in \{0, 1\}$. We define the *Pro-Active Inference* (PAI) estimator $\hat{\theta}_{n_b}$:

$$\hat{\theta}_{n_b} := \frac{1}{n_b} \sum_{t=1}^{n_b} \phi_t,\ \phi_t := \frac{1}{N_q} \sum_{j=1}^{N_q} \hat{p}_j^{(t-1)} + \frac{1}{N_q} \frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t\left(I_t\right)}. \tag{2}$$

Here, $\hat{p}_j^{(t-1)} := \sigma\left(\hat{u}^{(t-1)\top} v_j\right)$ and $\hat{u}^{(t-1)}$ is the current posterior-mean estimate of the new model factor. The first term in ϕ_t is the plug-in factor-model estimate of θ, while the second serves as a bias correction. Let $\mathcal{F}_{t-1} := \sigma(\{I_s, z_{I_s}\}_{s=1}^{t-1})$. We require $\hat{p}_j^{(t-1)}$ and $q_t(\cdot)$ to be $\mathcal{F}_{t-1}$-measurable: depending only on past queries/labels. In Theorem 3.1 (proven in Appendix A.5), we show that $\hat{\theta}_{n_b}$ is unbiased for θ and derive its limiting distribution via a martingale central limit theorem (CLT). To formalize asymptotics, we work under a triangular-array setup indexed by m. Specifically, we consider a sequence of problems $\{\mathcal{P}^{(m)}\}_{m \geq 1}$, in which $N_q = N_q^{(m)}$, $n_b = n_b^{(m)}$, and all involved quantities—$\{z_1, \ldots, z_{N_q}\}$, θ, $\hat{\theta}_{n_b}$, $\{\hat{p}^{(t)}\}_{0 \leq t < n_b}$, and $\{q_t(\cdot)\}_{1 \leq t \leq n_b}$—may depend on m. For readability, we suppress the superscript (m) in the statement and proof.

Theorem 3.1. *Consider a sequence of problems (indexed by m) such that $n_b \uparrow \infty$ as $m \to \infty$ and $n_b \leq N_q$ for every m. Then (a) $\hat{\theta}_{n_b}$ is unbiased for θ for every m; and (b) under mild regularity conditions (variance stabilization/control and Lindeberg; Assumptions A.1–A.3), $\hat{\theta}_{n_b}$ obeys $\sqrt{n_b}(\hat{\theta}_{n_b} - \theta)/\hat{\sigma}_{n_b} \xrightarrow{d} \mathcal{N}(0, 1)$, where $\hat{\sigma}_{n_b}^2$'s computable closed-form expression is in Appendix A.3.2.*

By Theorem 3.1, it follows that an asymptotic $(1 - \alpha)$-level CI for θ is $[\hat{\theta}_{n_b} \pm z_{1-\alpha/2}(\hat{\sigma}_{n_b}/\sqrt{n_b})]$. For a fixed bank of outcomes $(z_1, \ldots, z_{N_q})$, the coverage probability of this interval is

$$P_{\text{sampling}}\left(\theta \in \left[\hat{\theta}_{n_b} \pm z_{1-\alpha/2} \frac{\hat{\sigma}_{n_b}}{\sqrt{n_b}}\right] \mid z_1, \ldots, z_{N_q}\right),$$

where randomness arises solely from adaptive question selection. Although the theorem is stated for a sequence of problem instances, we observe a single one in practice. Consequently, Theorem 3.1 implies that the above CI attains coverage at approximately the nominal level $1 - \alpha$ for sufficiently large problem sizes.[4]

3.3 An Oracle-Optimal Sequential Sampling Policy

In practice, constructing a sequential policy $q_t(\cdot)$ that yields narrow CIs is difficult, partly because $\hat{p}_j^{(t)}$ evolves over rounds t. We therefore derive guidance from an oracle idealization. For this subsection

[3] To support Assumptions A.1–A.2 for Theorem 3.1 we require sampling with replacement. Appendix B discusses how ad-hoc without-replacement variants can fail to yield valid coverage.

[4] Nonasymptotic confidence intervals can also be constructed using martingale concentration inequalities, but straightforward approaches can be prohibitively conservative. For example, applying Azuma-Hoeffding under our hybrid sampling policy (c.f. Section 3.5) with the default uniform-mixing weight of $\tau = 0.05$ and a budget of $n_b = 1,200$ labels, yields a 95% confidence interval with half-width approximately 1.57, exceeding the entire parameter range $[0, 1]$. Sharper, potentially variance-adaptive nonasymptotic bounds may be possible, but their development is beyond the scope of this work. We also empirically assess FAQ's finite-budget coverage in Section 5.

only, assume an oracle superpopulation model: (a) $z_j \sim \text{Bern}(p_j)$ independently across questions j; (b) the oracle has access to the true, time-invariant $\{p_j\}$; and (c) the target is $\theta^* := N_q^{-1} \sum_{j=1}^{N_q} p_j$. As a corollary of Theorem 3.1(a), $\hat{\theta}_{n_b}$ is unbiased for θ^*. A natural direction is to minimize its variance in the oracle world (proof in Appendix A.6):[5]

Theorem 3.2. *In the aforementioned oracle setup, the variance-minimizing policy $q_t^*(j)$ is time-independent and uniquely defined by $q_t(j) \propto \sqrt{p_j(1-p_j)}$.*

3.4 An Active-Learning-Driven Sampling Policy

In practice, we use the factor model to estimate question difficulty as $\hat{p}_j^{(t-1)} := \sigma(\hat{u}^{(t-1)\top} v_j)$. Since question factors $\{v_j\}$ are fixed, prediction quality depends on how well we have learned $\hat{u}^{(t-1)}$—motivating an active-learning component that quickly learns u, but only insofar as it improves inference for θ^*. For this subsection only, assume the factor model is correctly specified: (a) $z_j \sim \text{Bern}\,(p_j)$ with $p_j := \sigma\left(u^\top v_j\right)$; (b) u has a normal prior; and (c) θ^* is random only through u:[6]

$$\theta^* = \frac{1}{N_q} \sum_{j=1}^{N_q} p_j = \frac{1}{N_q} \sum_{j=1}^{N_q} \sigma\left(u^\top v_j\right) =: g(u).$$

Given $u \mid \mathcal{F}_{t-1} \sim \mathcal{N}\left(\hat{u}^{(t-1)}, \hat{\Sigma}^{(t-1)}\right)$, we seek the query that most reduces the posterior variance of $g(u)$ (c.f. MacKay [1992]). To obtain a closed-form expression, we approximate the updated posterior $u \mid \mathcal{F}_{t-1}, z_j$ as normal and linearize g around $\hat{u}^{(t-1)}$. Under these approximations, the index I_t that maximally reduces θ^*'s posterior variance is given by $I_t = \text{argmax}_j\, d^{(t)}(j)$, whose computable closed-form expression is presented in Appendix A.3.3 and derived in Appendix A.7.

3.5 A Practical Hybrid Sampling Policy

In practice, we trade off exploitation (variance-reduction using the current $\hat{p}^{(t-1)}$) and exploration (querying informative questions to quickly refine $\hat{u}^{(t)}$ and hence $\hat{p}^{(t)}$). We implement this via a time-varying hybrid sampling policy. Inspired by Sections 3.3–3.4, at time t, define unnormalized oracle (o) and active-learning (a) scores as follows: $s_{\text{o}}^{(t)}(j) := \sqrt{\hat{p}_j^{(t-1)}(1-\hat{p}_j^{(t-1)})},\ s_{\text{a}}^{(t)}(j) := d^{(t)}(j)$. Define $\text{Norm}(x)(j) = x(j)/\sum_{j'} x(j')$. We convert each score to a probability vector via $h_{\text{o}}^{(t)} = \text{Norm}(s_{\text{o}}^{(t)})$ and $h_{\text{a}}^{(t)} = \text{Norm}(s_{\text{a}}^{(t)})$. Given hyperparameters $\rho, \gamma, \beta_0, \tau \in (0,1)$, set mixing weight $\alpha_t = \max\left(0, 1 - \frac{t}{\rho n_b}\right)$ and tempering exponent $\beta_t = \beta_0 \min\left(1, \frac{t}{\gamma n_b}\right)$. We mix the normalized scores and apply tempering to form our combined-and-tempered score $h_{\text{cat}}^{(t)}$:

$$h_{\text{cat}}^{(t)}(j) = \left((1-\alpha_t)\, h_{\text{o}}^{(t)}(j) + \alpha_t h_{\text{a}}^{(t)}(j)\right)^{\beta_t}. \tag{3}$$

Finally, we normalize $h_{\text{cat}}^{(t)}(\cdot)$ to be a probability vector and apply τ-uniform mixing to lower-bound all sampling probabilities. Our final hybrid policy $q_t(\cdot)$ is given by:

$$q_t(j) = \frac{\tau}{N_q} + (1-\tau)\,\text{Norm}(h_{\text{cat}}^{(t)})(j). \tag{4}$$

Here, α_t sets the exploration-to-exploitation schedule (decaying to 0 after ρn_b queries), while β_t tempers early scores towards near-uniform exploration (i.e., scores closer to each other) and reaches β_0 after γn_b queries. The tempering and τ-mixing ensure that all questions retain nontrivial selection probability—supporting Theorem 3.1's regularity conditions.

We initialize a new model's latent factor as $u \sim \mathcal{N}(\hat{u}^{(0)}, \hat{\Sigma}^{(0)})$. For rounds $t = 1, \ldots, n_b$, FAQ:

(a) Updates $q_t(\cdot)$ under the current factor-model and samples $I_t \sim q_t(\cdot)$ (Eq. 4).
(b) Queries the model to observe z_{I_t}.

[5]While access to $\{p_j\}$ makes θ known, we wish to find the variance-minimizing policy among estimators of form $\hat{\theta}_{n_b}$.
[6]We emphasize that the factor model is only a *mental model* that helps us in deriving a reasonable adaptive sampling policy.

(c) Updates the PAI estimate (Eq. 2)
(d) Updates $(\hat{u}^{(t)}, \hat{\Sigma}^{(t)})$ (Eqs. 7–8).

After n_b rounds, we estimate the variance and form a CI (Eq. 9). Algorithm A.1 gives pseudocode.

4 Experimental Setup

Datasets Using the HuggingFace `datasets` API, we collect per-question 0/1 outcomes for 4.4K models from the Open LLM Leaderboard [Fourrier et al., 2024, Myrzakhan et al., 2024]. Each model is evaluated on 21.6K questions across six datasets, which we aggregate into two suites: (a) MMLU-Pro with 12K questions; and (b) BBH+GPQA+IFEval+MATH+MuSR with 9.5K questions. Full dataset citations are in Appendix C. We release our compiled datasets to support future research. We focus on these suites because they provide large, reusable model-by-unit outcome matrices; interactive agent benchmarks with task-level success outcomes are natural targets for FAQ/PAI, but running them at comparable scale would require substantially greater evaluation infrastructure.

Historical splits and missingness For each suite, we sort models by release date, take the first 2.2K as historical H and the next 2.2K as new test models. We induce missingness in H by selecting $n_{\text{full-obs}}$ fully-observed rows uniformly at random, and independently masking each remaining entry with probability $1 - p_{\text{obs}}$ (i.e., MCAR: missing completely at random). Details are in Appendix C.

Factor model selection For each non-cold-start missingness setting, we split H by rows (models) into H_{train} (first 80%) and H_{val} (last 20%). We choose (λ, k) via 5-fold cross-validation on H_{train}: (a) fit on observed entries (Eq. 6) and evaluate binary prediction accuracy on an MCAR-masked holdout set; and (b) apply the 1-SE rule [Friedman et al., 2010].

Budgets and coverage We consider ten evenly-spaced budgets $n_b \in [2.5\%, 25\%]$ of the question bank and target 95% nominal coverage for all experiments.

FAQ hyperparameters For each non-cold-start budget/missingness setting, we tune $(\rho, \gamma, \beta_0, \tau)$ to minimize mean CI width on H_{val} (averaged over 5 seeds), using a factor model refit on all of H_{train}.[7] For tuning, we run FAQ for n_b rounds per validation model, querying selected questions when labels are missing. This mirrors deployment, where tuning expends additional queries on held-out models, beyond the reported query budget for evaluating each new model.

Baselines Let $\overline{p}_j$ be the mean of the *observed* outcomes in H for question j[8], used as a proxy for question difficulty. For each budget/missingness setting, we compare against model-assisted, design-based AIPW estimators for finite-population inference, using established sequential sampling schemes from survey sampling and active inference. Unlike FAQ/PAI, these baselines follow *nature's order*: questions arrive in a fixed stream and question t is labeled with probability $\pi_t(t)$. Let $\xi_t \sim \text{Bern}(\pi_t(t))$ indicate labeling. Each baseline uses

$$\hat{\theta}_{\text{base}} = \frac{1}{N_q} \sum_{t=1}^{N_q} \left(p_{\text{base}}(t) + (z_t - p_{\text{base}}(t)) \frac{\xi_t}{\pi_t(t)} \right), \tag{5}$$

with baseline predictor $p_{\text{base}}(t) \in \{0, \overline{p}_t\}$. We consider: (a) $\pi_t(t) = n_b/N_q$ (Bernoulli); (b) $\pi_t(t) \propto \sqrt{\overline{p}_t(1-\overline{p}_t)}$ (Neyman); and (c) $\pi_t(t) \propto \min(\overline{p}_t, 1-\overline{p}_t)$ (c.f. Zrnic and Candès [2024]). For (b)-(c), we normalize, apply τ-mixing, and use the budget-stabilization wrapper in Equation 8 of Zrnic and Candès [2024]. Note that $p_{\text{base}}(t) = 0$ and $\pi_t(t) = n_b/N_q$ reduces to *uniform sampling*.

Traditional active inference ablation Since the baselines above omit the factor model—a core component of FAQ—we additionally ablate FAQ to isolate the contribution of PAI and our factor-model-guided hybrid sampling policy. We replace PAI and hybrid sampling with Zrnic and Candès [2024]'s sequential active inference—keeping the same factor-model predictions $\hat{p}_j^{(t)}$, using their suggested labeling policies (Eqs. 3 and 8 therein), and tuning their τ-mixing parameter (Appendix C).

Metrics We evaluate (a) coverage and (b) effective sample size (ESS). For each setting, we run 100 random seeds for each of 2.2K test models. In each seed/model, methods output Wald intervals of the form $[\hat{\theta}_{\text{method}} \pm 1.96\hat{\sigma}_{\text{method}}]$ (clipped to be in $[0, 1]$), where $\hat{\sigma}_{\text{method}}$ is computed via the method's

[7] On H_{val}, $(\beta_0, \tau) = (1.0, 0.05)$ is a robust default across budget/missingness settings. (ρ, γ) are more setting-dependent.
[8] If question j is never observed in H (e.g., when 0.1% entries in H are observed), we impute $\overline{p}_j$ with H's global mean.

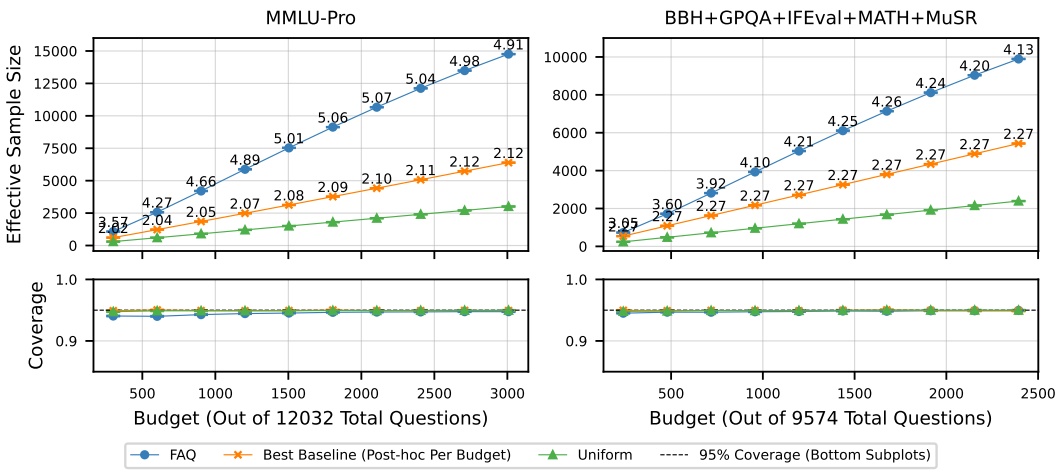

Figure 2: **Evaluating 2.2K LLMs on MMLU-Pro and BBH+GPQA+IFEval+MATH+MuSR with fully-observed historical data. (Top Row)** Effective sample size (ESS) vs. budget for FAQ, the strongest baseline per budget (selected post-hoc, favoring baselines), and uniform sampling. Numbers annotate the ESS multiplier relative to uniform sampling. For example, on MMLU-Pro at budget 1500, an ESS multiplier of 5.01 means FAQ matches uniform sampling's CI width at budget 7515 (a $5.01\times$ gain). **(Bottom Row)** Empirical coverage of the corresponding 95% normal-approximation CIs, averaged over 2.2K test models and 100 seeds. Standard errors are shown, but negligible on the plot scale. **Higher ESS and ESS-multiplier values indicate better performance.**

plug-in variance formula at budget n_b. Coverage is computed as the fraction of intervals containing θ, averaged over models and seeds. ESS at budget n_b is defined as $n_{\text{eff}} := (\hat{\sigma}^2_{\text{uniform}}/\hat{\sigma}^2_{\text{method}})n_b$, where $\hat{\sigma}^2_{\text{uniform}}$ is the plug-in variance of uniform sampling at the same budget n_b. Intuitively, n_{eff} is the uniform-sampling budget that would yield the same variance as the method at budget n_b: thus, larger ESS implies stronger performance. We emphasize ESS (over raw interval width) since it directly corresponds to query-cost savings: an ESS multiplier of m means matching uniform sampling's CI width with roughly a $1/m$ fraction of the queries.[9]

Good-faith evaluation FAQ is tuned using only historical train/validation data (without access to test models). In contrast, to avoid understating baseline performance, we report baselines *after post-hoc selection*: for each budget/dataset/missingness setting, we sweep baseline/ablation variants/hyperparameters on the test models (100 seeds) and report the best configuration (averaged across seeds), yielding an oracle-style upper bound that favors the baselines.

Cold-start and cross-suite transfer While FAQ as presented above leverages target-suite historical data to fit the factor model, practitioners may ask how it performs on a new benchmark before

[9] As all tested methods are unbiased and asymptotically normal, ESS gains translate into proportional reductions in MSE.

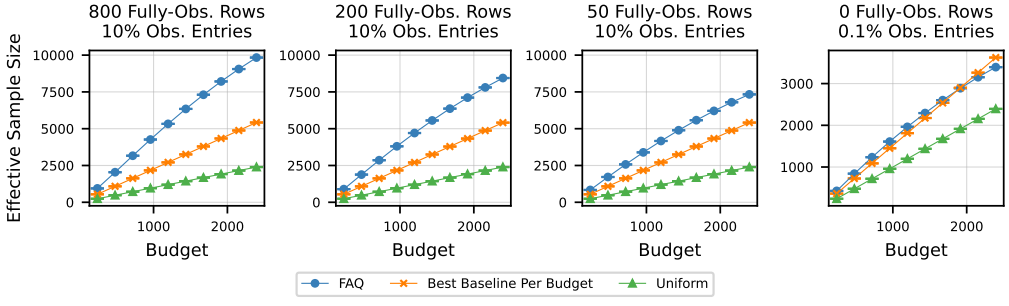

Figure 3: **ESS under missing historical data on BBH+GPQA+IFEval+MATH+MuSR.** ESS vs. budget across missingness settings (Section 4). Standard errors are shown, but negligible on the plot scale. MMLU-Pro and coverage results are in Appendix E.

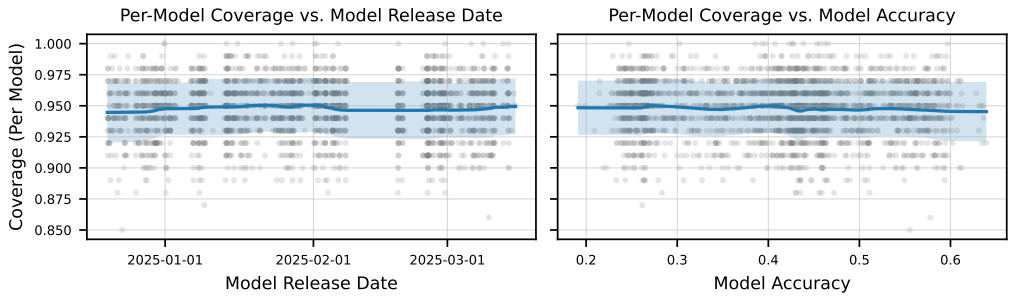

Figure 4: **Per-model coverages on BBH+GPQA+IFEval+MATH+MuSR with fully-observed historical data at budget** 7.5%**.** Per-model coverages (averaged over 100 seeds) vs. **(Left)** model release date and **(Right)** true accuracy. Gray dots: individual models; solid-blue curve: locally-weighted mean with a $K = 501$ nearest-neighbor Gaussian kernel; light-blue band: local ± 1 SD.

such data have accumulated. We therefore explore a cold-start factor-model experiment on GPQA (All), initializing factors via MMLU-Pro transfer: question factors use weighted k-NN over LLM-embedding cosine similarities of question text, while model factors come from the MMLU-Pro source fit. Because no target-suite historical matrix is available for factor-model selection, we use a fixed a priori factor-model configuration. FAQ sampling hyperparameters are tuned analogously on held-out validation models and minimizing CI width only. Full details are in Appendix C.

Hyperparameter search ranges are in Appendix C. Full results (including widths) are in Appendix E.

5 Empirical Results

Our primary experiments compare FAQ's ESS and coverage against uniform sampling and post-hoc-selected baselines across two benchmark suites and varying historical-data missingness levels (Figures 2–3). We also audit FAQ's per-model coverage under realistic deployment shifts (Figure 4) and explore a cold-start transfer without target-suite historical data for factor-model fitting (Figure 5). A factor-model-matched active-inference ablation is summarized below, with full results in Appendix D. Per setting, one parallelized FAQ run over 2.2K models and 12K questions takes under one minute on a single H100 GPU: negligible compared to model inference/evaluation time.[10]

Fully-observed historical data With fully-observed H, FAQ achieves: (a) 4–$5\times$ higher ESS than uniform sampling; (b) 1.8–$2.4\times$ higher ESS than the strongest post-hoc baseline; and (c) empirical coverage near the 95% target (Figure 2). These gains are stable across budgets and translate to roughly proportional query-cost savings. At the smallest budgets, FAQ shows mild ESS degradation and miscoverage under 1%—likely from noisy early factor estimates and less accurate CLT normal approximations, respectively.[11] Even there, FAQ still considerably outperforms all baselines.

Partially-observed historical data Since fully-observed history is rarely available at scale, we evaluate FAQ under progressively sparser H. Figure 3 shows that even with 12–42% of H observed (e.g., $n_{\text{full-obs}} \in \{50, 800\}$ with $p_{\text{obs}} = 0.1$ on remaining rows), FAQ substantially outperforms the strongest post-hoc baselines. Even with only 0.1% of historical entries observed, FAQ retains a modest but consistent ESS edge (Figure 3, last column) except at the two largest budgets, where noisy factors learned under such extreme sparsity can be over-trusted as tempering weakens late-budget.

Coverage audits under deployment shifts Figure 2 reports coverage aggregated over 2.2K models and 100 seeds. For deployment, it is also important that coverage does not vary systematically across models—e.g., that it is not driven by "easy" extreme-accuracy models while under-covering moderate, high-uncertainty ones. This is especially relevant here because we learn question factors from historical data and then freeze them, initializing new model factors from the historical mean/covariance. As newer models drift from the historical cohort (e.g., via new architectures or training recipes), these factors may become less representative, potentially impacting finite-sample coverage. Consistent

[10] Additional exploratory experiments examine structured historical-data missingness, historical-label corruption, and sensitivity to factor-model predictive accuracy at selected query budgets (Appendix F). These results further support FAQ's empirical efficiency gains and near-nominal coverage under a range of perturbations.

[11] Future work could explore t- or bootstrap-calibrated (over the martingale increments) quantiles.

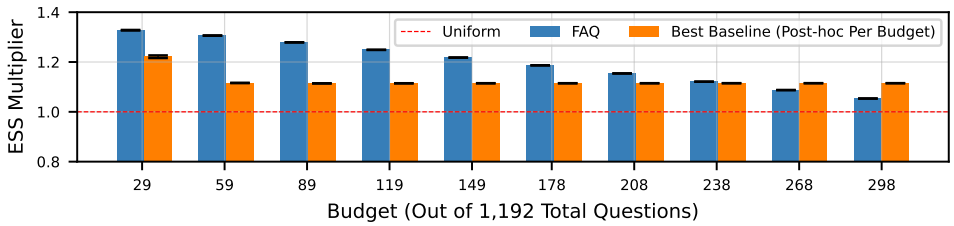

Figure 5: **Cold-start FAQ on GPQA (All).** ESS multipliers relative to uniform sampling across budgets for cold-start FAQ *with zero target-suite historical data for factor-model fitting* vs. strongest post-hoc baseline per budget *with access to* 0.1% *target-suite historical data.* Standard errors are shown, but negligible on the plot scale. **Higher ESS-multiplier values indicate better performance.**

with PAI's guarantee of valid coverage for every evaluated model, Figure 4 shows stable coverage across model release date and model accuracy, with no systematic deviations.

Cold-start and cross-suite transfer FAQ, *with zero target-suite historical data for factor-model fitting*, achieves ESS multipliers up to $1.32\times$ relative to uniform sampling and outperforms the strongest post-hoc selected baselines *with access to* 0.1% *target-suite historical data* on 8 of 10 budgets (Figure 5; coverages in Figure 12). This suggests that cross-benchmark structure transferred through LLM embeddings can make FAQ useful even before target-suite history accumulates, providing a natural onboarding path for continuous monitoring.

Active-inference ablation Appendix D compares FAQ to a factor-model-matched sequential active-inference ablation. FAQ yields narrower CIs in the low-budget regime most relevant for query-cost savings, while the sequential ablation can match or outperform FAQ at larger budgets.

6 Discussion and Conclusion

We introduced FAQ, a finite-bank framework for LLM evaluation that, in our experiments, achieves up to $5\times$ higher ESS than uniform sampling and $2.4\times$ higher ESS than the strongest baselines, with negligible computational overhead and model-free asymptotic frequentist coverage. FAQ also supports *continuous monitoring*: each evaluation leverages accumulated historical data while contributing fresh outcomes, making uncertainty-quantified evaluation progressively more cost-effective. Although hyperparameter tuning incurs additional query costs on held-out validation models (c.f. Section 4), this one-time cost per evaluation setting can be amortized across repeated model evaluations. Our empirically robust default of $(\beta_0, \tau) = (1.0, 0.05)$ further reduces the need for extensive tuning. This is particularly valuable when repeatedly evaluating model variants (e.g., checkpoints, fine-tunes, distillations, quantizations, or deployed systems) under limited query, latency, or adjudication budgets, where even modest per-run savings can compound.

Our limitations suggest future work. FAQ/PAI offers a framework for statistically rigorous, budget-constrained benchmarking: flexible predictions guide query selection, while PAI preserves asymptotically valid finite-bank inference. Our simple factor-model instantiation illustrates the core idea, but is not fundamental: FAQ's coverage is model-free, and any probability-outputting predictive model and adaptive sampling policy satisfying PAI's regularity conditions would retain valid coverage. Updating question factors after deployment waves, or using drift-aware priors informed by metadata, embeddings, or other side information, could improve efficiency without changing validity. For example, hierarchical latent-variable models could capture model-family and question-domain effects, while more flexible predictors based on Gaussian processes, neural networks, or LLMs could leverage additional metadata, question text, and embeddings. In the complete absence of historical data, transferable benchmark information, or other predictive side information, FAQ cannot be expected to improve upon uniform sampling. Nevertheless, PAI reduces exactly to classical uniform sampling by using constant predictions and uniform query probabilities, retaining its inferential guarantees.

FAQ also applies to any fixed benchmark bank with binary task-completion outcomes, including multi-turn or agentic benchmarks summarized by task success rates. Extending FAQ to non-binary feedback and batched querying would broaden applicability and parallelize evaluation across adjudicators/inference workers.

Acknowledgments and Disclosure of Funding

We thank Kevin Luo and Joon Lee for their many insightful comments and helpful discussions. S.W. and Y.N. were both supported by National Science Foundation Graduate Research Fellowships. E.J.C. was supported by the Office of Naval Research grant N000014-20-1-2157. We also gratefully acknowledge support from Arena Intelligence's Academic Partnerships Program. This work used Marlowe [Kapfer et al., 2025], Stanford University's GPU-based Computational Instrument, supported by Stanford HAI and Stanford Research Computing. This work also used Stanford University's Sherlock computing cluster. We thank Stanford University and the Stanford Research Computing Center for providing the computational resources and support that contributed to these research results. We also thank Scott Linderman for access to his H100 GPUs used in some experiments. The authors have no competing interests to declare.

References

Tijana Zrnic and Emmanuel Candès. Active Statistical Inference. In *International Conference on Machine Learning*, pages 62993–63010. PMLR, 2024.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *arXiv preprint arXiv:2405.01769*, 2024.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382, 2023.

Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. Large Language Models in Legal Systems: A Survey. *Humanities and Social Sciences Communications*, 12(1):1977, 2025.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and individual differences*, 103:102274, 2023.

Yaara Artsi, Vera Sorin, Benjamin S Glicksberg, Panagiotis Korfiatis, Girish N Nadkarni, and Eyal Klang. Large Language Models in Real-World Clinical Workflows: A Systematic Review of Applications and Implementation. *Frontiers in Digital Health*, 7:1659134, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Benjamin Laufer, Hamidah Oderinwale, and Jon Kleinberg. Anatomy of a Machine Learning Ecosystem: 2 Million Models on HuggingFace. *arXiv preprint arXiv:2508.06811*, 2025.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966, 2025.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference Leakage: A Contamination Problem in LLM-as-a-Judge. *arXiv preprint arXiv:2502.01534*, 2025.

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Cha, Arijit Ghosh Chowdhury, Prashanth Kallur Ramaswamy, Jeremy Roghair, Hannah R Marlowe, Carina Suzana Negreanu, Kitty Boxall, et al. LeMAJ (Legal LLM-as-a-Judge): Bridging Legal Reasoning and LLM Evaluation. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 318–337, 2025.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Current and Future State of Evaluation of Large Language Models for Medical Summarization Tasks. *NPJ Health Systems*, 2 (1):6, 2025.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-IF: Benchmarking LLMs on Multi-Turn and Multilingual Instructions Following. *arXiv preprint arXiv:2410.15553*, 2024.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

Jiayu Ding, Shuming Ma, Lei Cui, Nanning Zheng, and Furu Wei. LongReasonArena: A Long Reasoning Benchmark for Large Language Models. *arXiv preprint arXiv:2508.19363*, 2025.

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, 2024.

Jin Peng Zhou, Christian K Belardi, Ruihan Wu, Travis Zhang, Carla Gomes, Wen Sun, and Kilian Q Weinberger. On Speeding Up Language Model Evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Claudio Mazzoni and Anubhav Sharma. Beyond Vibes: How to Properly Select the Right LLM for the Right Task, 2025. URL `https://aws.amazon.com/blogs/machine-learning/beyond-vibes-how-to-properly-select-the-right-llm-for-the-right-task/`.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM Leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active Evaluation Acquisition for Efficient LLM Benchmarking. *arXiv preprint arXiv:2410.05952*, 2024.

Andriy Mnih and Russ R Salakhutdinov. Probabilistic Matrix Factorization. *Advances in Neural Information Processing Systems*, 20, 2007.

Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A Generalization of Principal Components Analysis to the Exponential Family. *Advances in Neural Information Processing Systems*, 14, 2001.

Yunxiao Chen, Chengcheng Li, Jing Ouyang, and Gongjun Xu. Statistical inference for noisy incomplete binary matrix. *Journal of Machine Learning Research*, 24(95):1–66, 2023.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis. *Psychometrika*, 84(1):124–146, 2019.

Chengyu Cui and Gongjun Xu. Identifiability and Inference for Generalized Latent Factor Models. *Journal of the American Statistical Association*, (just-accepted):1–29, 2026.

Peiyao Cai, Chengyu Cui, Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Mikhail Yurochkin, Yuekai Sun, Kean Ming Tan, and Gongjun Xu. A Latent Variable Framework for Scaling Laws in Large Language Models. *arXiv preprint arXiv:2512.06553*, 2025.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candès, and Dan Jurafsky. Can Unconfident LLM Annotations be used for Confident Conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533, 2025.

Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhransu Maji, and Daniel Sheldon. Active Measurement: Efficient Estimation at Scale. *Advances in Neural Information Processing Systems*, 38:145769–145807, 2026.

Burr Settles. Active Learning Literature Survey. 2009.

Francesco Di Fiore, Michela Nardelli, and Laura Mainini. Active Learning and Bayesian optimization: A Unified Perspective to Learn with a Goal. *arXiv preprint arXiv:2303.01560*, 2023.

David JC MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Valerii Vadimovich Fedorov. *Theory of Optimal Experiments*. Elsevier, 2013.

Andrew I Schein and Lyle H Ungar. Active Learning for Logistic Regression: An Evaluation. *Machine Learning*, 68(3):235–265, 2007.

Moritz Hardt. The Emerging Science of Machine Learning Benchmarks. *Manuscript. https://mlbenchmarks. org*, 2025.

Guanhua Zhang, Florian E Dorner, and Moritz Hardt. How Benchmark Prediction from Fewer Data Misses the Mark. *arXiv preprint arXiv:2506.07673*, 2025.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor Points: Benchmarking Models with Much Fewer Examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with Fewer Examples. *arXiv preprint arXiv:2402.14992*, 2024.

Xu-Xiang Zhong, Chao Yi, and Han-Jia Ye. Efficient Evaluation of Large Language Models via Collaborative Filtering. *arXiv preprint arXiv:2504.08781*, 2025.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. How to Select Datapoints for Efficient Human Evaluation of NLG Models? *Transactions of the Association for Computational Linguistics*, 13: 1789–1811, 2025.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient Benchmarking (of Language Models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536, 2024.

Anastasios N Angelopoulos, Jacob Eisenstein, Jonathan Berant, Alekh Agarwal, and Adam Fisch. Cost-Optimal Active AI Model Evaluation. *arXiv preprint arXiv:2506.07949*, 2025.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-Prediction-Powered Inference. *arXiv preprint arXiv:2603.27414*, 2026.

Robert E Blackwell, Jon Barry, and Anthony G Cohn. Towards Reproducible LLM Evaluation: Quantifying Uncertainty in LLM Benchmark Scores. *arXiv preprint arXiv:2410.03492*, 2024.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-LLM-Leaderboard: From Multi-Choice to Open-Style Questions for LLMs Evaluation, Benchmark, and Arena. *arXiv preprint arXiv:2406.07545*, 2024.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Craig Kapfer, Kurt Stine, Balasubramanian Narasimhan, Christopher Mentzel, and Emmanuel Candès. Marlowe: Stanford's GPU-based Computational Instrument, 2025. URL `https://doi.org/10.5281/zenodo.14751899`.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

Peter Hall and Christopher C Heyde. *Martingale Limit Theory and its Application*. Academic press, 2014.

Ward Whitt. Proofs of the Martingale FCLT. 2007.

Jean Jacod and Philip Protter. *Discretization of Processes*. Springer, 2012.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging Big-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *CoRR*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *Advances in Neural Information Processing Systems*, 2021.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *The Twelfth International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 2019.

A Proofs, Algorithms, and other Formulas

A.1 Full FAQ Procedure

Algorithm 1 Factorized Active Querying

1: **Input:** Historical data matrix H, confidence level α, and budget n_b
2: Fit factor model on H to initialize prior $\mathcal{N}(\hat{u}^{(0)}, \hat{\Sigma}^{(0)})$ for new model's latent factor.
3: **for** $t = 1$ **to** n_b **do**
4: Compute hybrid probability vector over queries at time t:

$$q_t(j) = \frac{\tau}{N_q} + (1-\tau)\mathrm{Norm}(h_{\mathrm{cat}}^{(t)})(j),$$

where $h_{\mathrm{cat}}^{(t)}$ is the tempered convex combination of oracle variance-reduction and active-learning-driven policies given in Equation (3).
5: Sample new index $I_t \sim q_t(\cdot)$ and query LLM with question I_t to obtain z_{I_t}.
6: Compute

$$\phi_t := \frac{1}{N_q}\sum_{j=1}^{N_q} \hat{p}_j^{(t-1)} + \frac{1}{N_q} \cdot \frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}.$$

7: Update factor model prior via

$$(\hat{u}^{(t)}, \hat{\Sigma}^{(t)}) \leftarrow (\texttt{ModelFactorUpdate}(z_{I_t}, v_{I_t}, \hat{u}^{(t-1)}, \hat{\Sigma}^{(t-1)}),$$

given in Equations (7)–(8).
8: **end for**
9: Compute the PAI variance estimate given in Equation (9):

$$\hat{\sigma}_{n_b}^2 = \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b} \frac{\left(z_{I_t} - \hat{p}_{I_t}^{(t-1)}\right)^2}{q_t(I_t)^2} - \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b}\left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{z_{I_s}}{q_s(I_s)} - \sum_{j=1}^{N_q}\hat{p}_j^{(t-1)}\right)^2,$$

using samples $\{z_{I_t}\}_{t=1}^{n_b}$, computed probability vectors $\{q_t(\cdot)\}_{t=1}^{n_b}$, and factor model estimates $\{\hat{p}_j^{(t)}\}_{j\in[N_q], t\in[0:n_b-1]}$.
10: **Return** Estimate $\hat{\theta}_{n_b} := \frac{1}{n_b}\sum_{t=1}^{n_b}\phi_t$ and $(1-\alpha)$-CI: $\left[\hat{\theta}_{n_b} \pm z_{1-\alpha/2} \cdot \frac{\hat{\sigma}_{n_b}}{\sqrt{n_b}}\right]$.

A.2 Full Historical Factor Model Negative Log-Likelihood Objective

Let O be a 0/1-valued $N_{\text{old}} \times N_q$ matrix such that $O_{ij} = 1$ if H_{ij} is observed, else $O_{ij} = 0$. Our historical minimization objective is given by the following:

$$L\left(\{u_{i'}\}_{i'=1}^{N_{\text{old}}}, \{v_{j'}\}_{j'=1}^{N_q}\right) = -\sum_{i=1}^{N_{\text{old}}}\sum_{j=1}^{N_q} O_{ij}\left(H_{ij}\log\sigma\left(u_i^\top v_j\right) + (1-H_{ij})\log\left(1-\sigma\left(u_i^\top v_j\right)\right)\right), \tag{6}$$

We implement using PyTorch's binary cross-entropy function for numerical stability. In practice, $\{u_{i'}\}_{i'=1}^{N_{\text{old}}}$ and $\{v_{j'}\}_{j'=1}^{N_q}$ are fit using AdamW with weight decay λ for L2 regularization, with (λ, k) chosen via cross-validation (Section 4).

A.3 Formulas for Factor Model Online Laplace Updates, Variance Estimators, and Hybrid Sampling Policy Components

A.3.1 Factor Model Online Laplace Updates

At round t, after querying question I_t and observing z_{I_t}, we update the Gaussian approximation $u \mid \mathcal{F}_t \approx \mathcal{N}(\hat{u}^{(t)}, \hat{\Sigma}^{(t)})$ via the following online Laplace update (derivation in Appendix A.4),

denoting these updates as $\hat{\Sigma}^{(t)}, \hat{u}^{(t)} \leftarrow \texttt{ModelFactorUpdate}(z_{I_t}, v_{I_t}, \hat{u}^{(t-1)}, \hat{\Sigma}^{(t-1)})$:

$$\hat{p}_{I_t}^{(t-1)} := \sigma\left(\hat{u}^{(t-1)\top} v_{I_t}\right), \hat{w} := \hat{p}_{I_t}^{(t-1)}(1 - \hat{p}_{I_t}^{(t-1)}), \tag{7}$$

$$\hat{\Sigma}^{(t)} := \hat{\Sigma}^{(t-1)} - \frac{\hat{w}\hat{\Sigma}^{(t-1)} v_{I_t} v_{I_t}^\top \hat{\Sigma}^{(t-1)}}{1 + \hat{w} v_{I_t}^\top \hat{\Sigma}^{(t-1)} v_{I_t}}, \hat{u}^{(t)} := \hat{u}^{(t-1)} + \hat{\Sigma}^{(t)}\left(z_{I_t} - \hat{p}_{I_t}^{(t-1)}\right) v_{I_t}. \tag{8}$$

A.3.2 Computable Closed-Form Variance Estimator for Pro-Active Inference

The estimator for the limiting variance of $\hat{\theta}_{n_b}$ in Theorem 3.1 is as follows (derived in Appendix A.5):

$$\hat{\sigma}_{n_b}^2 := \frac{1}{n_b N_q^2} \sum_{t=1}^{n_b} \frac{\left(z_{I_t} - \hat{p}_{I_t}^{(t-1)}\right)^2}{q_t(I_t)^2} - \frac{1}{n_b N_q^2} \sum_{t=1}^{n_b} \left(\frac{1}{n_b} \sum_{s=1}^{n_b} \frac{z_{I_s}}{q_s(I_s)} - \sum_{j=1}^{N_q} \hat{p}_j^{(t-1)} \right)^2. \tag{9}$$

A.3.3 Computable Closed-Form Expression for Maximal Posterior-Variance-Reducing Index under Assumptions and Approximations in Section 3.4.

Under the assumptions and approximations in Section 3.4, the index I_t that maximally reduces θ^*'s posterior variance is given by $I_t = \text{argmax}_j \, d^{(t)}(j)$ (derived in Appendix A.7), where

$$d^{(t)}(j) = \frac{\hat{w}_j^{(t-1)} \left(v_j^\top \hat{\Sigma}^{(t-1)} \frac{1}{N_q} \sum_{j'=1}^{N_q} \hat{w}_{j'}^{(t-1)} v_{j'} \right)^2}{1 + \hat{w}_j^{(t-1)} v_j^\top \hat{\Sigma}^{(t-1)} v_j},$$
$$\hat{w}_j^{(t-1)} = \sigma\left(\hat{u}^{(t-1)\top} v_j\right)\left(1 - \sigma\left(\hat{u}^{(t-1)\top} v_j\right)\right).$$

A.4 Derivation of Online Logistic Regression Laplace Approximation Update

This derivation was inspired by Ch. 4 of Bishop and Nasrabadi [2006]'s discussion on Laplace approximations and Bayesian logistic regression. The key idea is to Taylor expand the log-posterior around the current mean, which yields approximate closed-form update formulas while preserving approximate Gaussian conjugacy.

Suppose at time $t-1$ we have $u \sim \mathcal{N}(m, S)$. At time t, we queried z with corresponding question factor v.[12] Because under our factor model, $z \sim \text{Bern}\left(\sigma\left(u^\top v\right)\right)$, by Bayes' Rule we have the following, where "$\sim$" refers to equality up to an additive constant that does not depend on u, and $\sigma(x) := \frac{1}{1+\exp(-x)} = \frac{\exp(x)}{1+\exp(x)}$:

$$p(u \mid z) \propto p(z \mid u) p(u) \propto \sigma\left(u^\top v\right)^z \left(1 - \sigma\left(u^\top v\right)\right)^{1-z} \exp\left(-\frac{1}{2}(u-m)^\top S^{-1}(u-m)\right)$$

$$\implies \log p(u \mid z) \sim z \log\left(\sigma\left(u^\top v\right)\right) + (1-z)\log\left(1 - \sigma\left(u^\top v\right)\right) - \frac{1}{2}(u-m)^\top S^{-1}(u-m)$$

$$= z u^\top v - z \log\left(1 + \exp\left(u^\top v\right)\right) - (1-z)\log\left(1 + \exp(u^\top v)\right) - \frac{1}{2}(u-m)^\top S^{-1}(u-m)$$

$$= z u^\top v - \log\left(1 + \exp(u^\top v)\right) - \frac{1}{2}(u-m)^\top S^{-1}(u-m).$$

To apply Laplace's approximation, we will require a second-order Taylor expansion of $\log p(u \mid z)$ about $u = m$. For now, we proceed by computing the gradient and Hessian of the exact log-posterior:

$$\nabla_u \log p(u \mid z) = zv - \frac{\exp(u^\top v)}{1 + \exp(u^\top v)} v - S^{-1}(u-m) = zv - \sigma\left(u^\top v\right) v - S^{-1}(u-m),$$

and, noting that $\nabla_x \sigma(x) = \sigma(x)(1 - \sigma(x))$, we have

$$\nabla_u^2 \log p(u \mid z) = -S^{-1} - \sigma\left(u^\top v\right)\left(1 - \sigma\left(u^\top v\right)\right) v v^\top.$$

[12] For notational brevity, we omit random indices I_t, superscripts, and subscripts. The derivations are left exactly unchanged.

Assembling our second-order Taylor expansion and defining $\hat{p} := \sigma\left(u^\top v\right)$, let

$$b = \nabla_u \log p(u \mid z) \mid_{u=m} = zv - \sigma\left(u^\top v\right) v := (z - \hat{p})\, v,$$

and

$$A = \nabla_u^2 \log p(u \mid z) \mid_{u=m} = -S^{-1} - \hat{p}\,(1 - \hat{p})\, vv^\top.$$

Then, we have

$$\begin{aligned}\log p(u \mid z) &\approx \log p(m \mid z) + b^\top (u - m) + \frac{1}{2}(u - m)^\top A\,(u - m) \\ &\sim b^\top u + \frac{1}{2} u^\top A u - u^\top A m = \frac{1}{2}\left(u^\top A u - 2\,(Am - b)^\top u\right).\end{aligned}$$

After completing the square, we obtain the following

$$\begin{aligned}\log p(u \mid z) &\sim \frac{1}{2}\left(u^\top A u - 2\,(Am - b)^\top u\right) \\ &\sim -\frac{1}{2}\left(\left(u - A^{-1}(Am - b)\right)^\top (-A)\left(u - A^{-1}(Am - b)\right)\right) \\ &= -\frac{1}{2}\left(\left(u - \left(m - A^{-1}b\right)\right)^\top (-A)\left(u - \left(m - A^{-1}b\right)\right)\right).\end{aligned}$$

Pattern-matching to the multivariate normal distribution, we deduce that

$$\begin{aligned}u \mid z &\dot{\sim} \mathcal{N}\left(m - A^{-1}b, -A^{-1}\right) \\ &\overset{d}{=} \mathcal{N}\left(m + \left(S^{-1} + \hat{p}\,(1 - \hat{p})\, vv^\top\right)^{-1}(z - \hat{p})\, v, \left(S^{-1} + \hat{p}\,(1 - \hat{p})\, vv^\top\right)^{-1}\right).\end{aligned}$$

For compute-efficient updates, the Sherman-Morrison formula tells us that

$$\left(S^{-1} + \hat{p}\,(1 - \hat{p})\, vv^\top\right)^{-1} = S - \frac{\hat{p}(1 - \hat{p}) S v v^\top S}{1 + \hat{p}(1 - \hat{p}) v^\top S v}.$$

Putting everything together,

$$u \mid z \dot{\sim} \mathcal{N}\left(m + \left(S - \frac{\hat{p}(1 - \hat{p}) S v v^\top S}{1 + \hat{p}(1 - \hat{p}) v^\top S v}\right)(z - \hat{p})\, v, S - \frac{\hat{p}(1 - \hat{p}) S v v^\top S}{1 + \hat{p}(1 - \hat{p}) v^\top S v}\right).$$

A.5 Proof of Theorem 3.1

To formalize asymptotics, we work under a triangular-array setup indexed by m. Specifically, we consider a sequence of problems $\{\mathcal{P}^{(m)}\}_{m\geq 1}$, in which the bank size $N_q = N_q^{(m)}$, the budget $n_b = n_b^{(m)}$, and all involved quantities–$\{z_1, \ldots, z_{N_q}\}$, θ, $\hat{\theta}_{n_b}$, $\{\hat{p}^{(t)}\}_{0\leq t<n_b}$, and $\{q_t(\cdot)\}_{1\leq t\leq n_b}$—may depend on m. All asymptotic statements are in the limit as $m \to \infty$ with $n_b^{(m)} \uparrow \infty$. Henceforth, we suppress the superscript (m) for readability.

A.5.1 Definitions

Recall that our estimand is $\theta = \frac{1}{N_q}\sum_{j=1}^{N_q} z_j$ and $\hat{\theta}_{n_b} = \frac{1}{n_b}\sum_{t=1}^{n_b}\phi_t$, where

$$\phi_t = \frac{1}{N_q}\sum_{j=1}^{N_q}\hat{p}_j^{(t-1)} + \frac{1}{N_q}\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}.$$

Let us define $\Delta_t := \phi_t - \frac{1}{N_q}\sum_{j=1}^{N_q} z_j = \phi_t - \theta$ and define the filtration $\mathcal{F}_t = \sigma(I_1, z_{I_1}, \ldots, I_t, z_{I_t})$.

A.5.2 Technical Assumptions/Conditions

We enumerate the technical assumptions/conditions required to prove the asymptotic normality stated in Theorem 3.1.

Assumption A.1 (Variance stabilization)**.** The average conditional variance converges in probability to a positive deterministic limit σ^2 (which does not depend on (m)):

$$\frac{1}{n_b}\sum_{t=1}^{n_b}\mathrm{Var}\left(\Delta_t \mid \mathcal{F}_{t-1}\right) \xrightarrow{p} \sigma^2.$$

Assumption A.2 (Lindeberg condition)**.** The triangular array of problem instances in combination with sampling policies obey the Lindeberg condition. Specifically, for every $\epsilon > 0$,

$$\frac{1}{n_b}\sum_{t=1}^{n_b}\mathbb{E}[\Delta_t^2\mathbf{1}(|\Delta_t| > \epsilon\sqrt{n_b}) \mid \mathcal{F}_{t-1}] \xrightarrow{p} 0.$$

Assumption A.3 (Conditional variance control)**.** The average conditional variance of the inverse-propensity-weighted predicted probabilities satisfies

$$\frac{1}{n_b}\sum_{t=1}^{n_b}\mathrm{Var}\left(\frac{1}{N_q}\cdot\frac{\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} \mid \mathcal{F}_{t-1}\right) = o_p(n_b).$$

Assumptions A.1–A.2 are the standard conditions under which a CLT for martingale difference arrays holds (c.f. Hall and Heyde [2014, Corollary 3.2]). Assumption A.3 is mild and is used to facilitate consistency of the variance estimator. Intuitively, in our setting of LLM evaluation with our proposed hybrid sampling policy (c.f. Eq. 4), the variance stabilization assumption is reasonable because as more outcome labels are observed, the model-factor estimate $(\hat{u}^{(t)}, \hat{\Sigma}^{(t)})$ stabilizes, which stabilizes $\hat{p}^{(t)}$, and hence $q_t(\cdot)$, too (because (α_t, β_t) also stabilize). In addition, the Lindeberg condition is satisfied, by construction, under our hybrid sampling policy because τ-mixing lower-bounds $q_t(j)$, avoiding extreme inverse-probability weights, and since $z_j, \hat{p}_j^{(t)} \in [0, 1]$, the increments Δ_t are uniformly bounded (by a constant depending only on τ). The τ-mixing also ensures the satisfaction of Assumption A.3.

A.5.3 Asymptotic Normality via Martingale CLT

By substitution, we begin with the following:

$$\sqrt{n_b}\left(\hat{\theta}_{n_b} - \theta\right) = \sqrt{n_b}\left(\frac{1}{n_b}\sum_{t=1}^{n_b}\phi_t - \frac{1}{N_q}\sum_{j=1}^{N_q} z_j\right).$$

Simplifying, we have

$$\sqrt{n_b}\left(\hat{\theta}_{n_b} - \theta\right) = \frac{1}{\sqrt{n_b}}\sum_{t=1}^{n_b}\phi_t - \frac{\sqrt{n_b}}{N_q}\sum_{j=1}^{N_q} z_j = \frac{1}{\sqrt{n_b}}\sum_{t=1}^{n_b}\phi_t - \frac{1}{\sqrt{n_b}N_q}\sum_{t=1}^{n_b}\sum_{j=1}^{N_q} z_j$$
$$= \sum_{t=1}^{n_b}\frac{1}{\sqrt{n_b}}\left(\phi_t - \frac{1}{N_q}\sum_{j=1}^{N_q} z_j\right) = \sum_{t=1}^{n_b}\frac{1}{\sqrt{n_b}}\Delta_t.$$

Next, we will show that the Δ_t are martingale differences—i.e., that their cumulative sums form a martingale.

Proposition A.4 (Martingale differences and unbiasedness)**.** *Define M_t as follows:*

$$M_t := \sum_{s=1}^{t}\left(\phi_s - \frac{1}{N_q}\sum_{j=1}^{N_q} z_j\right) := \sum_{s=1}^{t}\Delta_s. \tag{10}$$

Then, $(M_t, \mathcal{F}_t)$ is a martingale. In particular, $\hat{\theta}_{n_b}$ is unbiased for θ for every finite n_b.

Proof. Substituting, we have

$$\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[\phi_t - \frac{1}{N_q}\sum_{j=1}^{N_q} z_j \mid \mathcal{F}_{t-1}\right]$$
$$= \frac{1}{N_q}\mathbb{E}\left[\sum_{j=1}^{N_q}\hat{p}_j^{(t-1)} + \frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} - \sum_{j=1}^{N_q} z_j \mid \mathcal{F}_{t-1}\right]$$
$$= \frac{1}{N_q}\mathbb{E}\left[\sum_{j=1}^{N_q}\left(\hat{p}_j^{(t-1)} - z_j\right) + \frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} \mid \mathcal{F}_{t-1}\right]$$
$$= \frac{1}{N_q}\sum_{j=1}^{N_q}\left(\hat{p}_j^{(t-1)} - z_j\right) + \frac{1}{N_q}\mathbb{E}\left[\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} \mid \mathcal{F}_{t-1}\right]$$
$$= \frac{1}{N_q}\sum_{j=1}^{N_q}\left(\hat{p}_j^{(t-1)} - z_j\right) + \frac{1}{N_q}\sum_{j=1}^{N_q}\frac{z_j - \hat{p}_j^{(t-1)}}{q_t(j)}P(I_t = j \mid \mathcal{F}_{t-1})$$
$$= \frac{1}{N_q}\sum_{j=1}^{N_q}\left(\hat{p}_j^{(t-1)} - z_j\right) + \frac{1}{N_q}\sum_{j=1}^{N_q}\frac{z_j - \hat{p}_j^{(t-1)}}{q_t(j)}q_t(j) = 0,$$

thus confirming that we have a martingale. As a corollary, by summing and taking expectations, we find that $\mathbb{E}[\hat{\theta}_{n_b}] = \theta$: i.e., that $\hat{\theta}_{n_b}$ is unbiased for θ at any finite n_b. □

Given that we have a martingale, we now proceed to obtain the asymptotic variance of $\hat{\theta}_{n_b}$. To do, so, we need to first compute the conditional variances of our martingale increments. Because the z_j are treated as fixed, by substitution we have

$$\begin{aligned}
&\text{Var}\left(\Delta_t \mid \mathcal{F}_{t-1}\right) \\
&= \frac{1}{N_q^2}\text{Var}\left(\sum_{j=1}^{N_q}\hat{p}_j^{(t-1)} + \frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} - \sum_{j=1}^{N_q} z_j \mid \mathcal{F}_{t-1}\right) = \frac{1}{N_q^2}\text{Var}\left(\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} \mid \mathcal{F}_{t-1}\right) \\
&= \frac{1}{N_q^2}\mathbb{E}\left[\left(\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right)^2 \mid \mathcal{F}_{t-1}\right] - \frac{1}{N_q^2}\mathbb{E}\left[\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)} \mid \mathcal{F}_{t-1}\right]^2 \\
&= \frac{1}{N_q^2}\sum_{j=1}^{N_q}\left(\frac{z_j - \hat{p}_j^{(t-1)}}{q_t(j)}\right)^2 P(I_t = j \mid \mathcal{F}_{t-1}) - \frac{1}{N_q^2}\left(\sum_{j=1}^{N_q}\frac{z_j - \hat{p}_j^{(t-1)}}{q_t(j)} P(I_t = j)\right)^2 \\
&= \frac{1}{N_q^2}\left(\sum_{j=1}^{N_q}\frac{\left(z_j - \hat{p}_j^{(t-1)}\right)^2}{q_t(j)} - \left(\sum_{j=1}^{N_q}\left(z_j - \hat{p}_j^{(t-1)}\right)\right)^2\right).
\end{aligned}$$

Under Assumptions A.1–A.2, the martingale central limit theorem (c.f., Hall and Heyde [2014, Corollary 3.2]) implies

$$\sqrt{n_b}\left(\hat{\theta}_{n_b} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right). \tag{11}$$

Lemma A.5 below shows that $\hat{\sigma}^2_{n_b} \xrightarrow{p} \sigma^2$, and thus the result follows by Slutsky's theorem.

Lemma A.5. *Define*

$$\hat{\sigma}^2_{n_b} = \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b}\frac{\left(z_{I_t} - \hat{p}_{I_t}^{(t-1)}\right)^2}{q_t(I_t)^2} - \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b}\left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{z_{I_s}}{q_s(I_s)} - \sum_{j=1}^{N_q}\hat{p}_j^{(t-1)}\right)^2. \tag{12}$$

Then $\hat{\sigma}^2_{n_b} \xrightarrow{p} \sigma^2$.

Proof. Define

$$\hat{A}_{n_b} := \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b}\frac{\left(z_{I_t} - \hat{p}_{I_t}^{(t-1)}\right)^2}{q_t(I_t)^2}, \hat{B}_{n_b} := \frac{1}{n_b N_q^2}\sum_{t=1}^{n_b}\left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{z_{I_s}}{q_s(I_s)} - \sum_{j=1}^{N_q}\hat{p}_j^{(t-1)}\right)^2$$

as well as

$$A := \frac{1}{n_b}\sum_{t=1}^{n_b}\frac{1}{N_q^2}\sum_{j=1}^{N_q}\frac{\left(z_j - \hat{p}_j^{(t-1)}\right)^2}{q_t(j)},\ B := \frac{1}{n_b}\sum_{t=1}^{n_b}\frac{1}{N_q^2}\left(\sum_{j=1}^{N_q}\left(z_j - \hat{p}_j^{(t-1)}\right)\right)^2$$

To show that $\hat{\sigma}^2_{n_b} \xrightarrow{p} \sigma^2$, it suffices to show that (a) $\hat{A}_{n_b} = A + o_p(1)$ and (b) $\hat{B}_{n_b} = B + o_p(1)$. We show these below:

(a) Define $\tilde{\theta}^{(t-1)} := \frac{1}{N_q}\sum_{j=1}^{N_q}\hat{p}_j^{(t-1)}$. Under Assumptions A.1–A.2, Hall and Heyde [2014, Theorem 2.23] guarantees that

$$\begin{aligned}
o_p(1) &= \frac{1}{n_b}\sum_{t=1}^{n_b}\Bigg[\left(\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}+\tilde{\theta}^{(t-1)}-\theta\right)^2 \\
&\quad -\left(\frac{1}{N_q^2}\sum_{j=1}^{N_q}\frac{\left(z_j-\hat{p}_j^{(t-1)}\right)^2}{q_t(j)}-(\tilde{\theta}^{(t-1)}-\theta)^2\right)\Bigg] \\
&= \hat{A}_{n_b} - A + \frac{1}{n_b}\sum_{t=1}^{n_b}\left[2(\tilde{\theta}^{(t-1)}-\theta)^2 - 2(\theta-\tilde{\theta}^{(t-1)})\left(\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right)\right].
\end{aligned}$$

To show that $\hat{A}_{n_b} - A = o_p(1)$ it suffices to show that

$$\frac{1}{n_b}\sum_{t=1}^{n_b}\left(\theta-\tilde{\theta}^{(t-1)}\right)\left(\theta-\tilde{\theta}^{(t-1)}-\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right) = o_p(1). \tag{13}$$

To show (13), we apply the Lenglart inequality (c.f., e.g., Whitt [2007, Lemma 3.7] or Jacod and Protter [2012, pp. 45]) to obtain

$$\begin{aligned}
&P\left(\left|\frac{1}{n_b}\sum_{t=1}^{n_b}\left(\theta-\tilde{\theta}^{(t-1)}\right)\left(\theta-\tilde{\theta}^{(t-1)}-\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right)\right|\geq\epsilon\right)\leq \\
&\frac{\eta}{\epsilon}+P\left(\frac{1}{n_b^2}\sum_{t=1}^{n_b}\mathbb{E}\left[\left(\theta-\tilde{\theta}^{(t-1)}\right)^2\left(\theta-\tilde{\theta}^{(t-1)}-\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right)^2\mid\mathcal{F}_{t-1}\right]\geq\eta\right)
\end{aligned}$$

for all $\epsilon,\eta>0$. Hence, it suffices to show that

$$\frac{1}{n_b^2}\sum_{t=1}^{n_b}\mathbb{E}\left[\left(\theta-\tilde{\theta}^{(t-1)}\right)^2\left(\theta-\tilde{\theta}^{(t-1)}-\frac{1}{N_q}\frac{z_{I_t}-\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\right)^2\mid\mathcal{F}_{t-1}\right]=o_p(1),$$

which is true by virtue of the fact that $(\theta-\tilde{\theta}^{(t-1)})^2\leq 1$ in combination with Assumption A.1.

(b) Expanding both squares, we find that

$$\begin{aligned}
\hat{B}_{n_b}-B &= \left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}\right)^2-\theta^2-2\tilde{\theta}^{(t-1)}\left[\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}-\theta\right] \\
&= \left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}-\theta\right)^2+2\theta\left(\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}-\theta\right) \\
&\quad -2\tilde{\theta}^{(t-1)}\left[\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}-\theta\right].
\end{aligned}$$

As $\theta,\tilde{\theta}^{(t-1)}\in[0,1]$, it suffices to show that $\frac{1}{n_b}\sum_{s=1}^{n_b}\frac{1}{N_q}\cdot\frac{z_{I_s}}{q_s(I_s)}-\theta\xrightarrow{p}0$. Since $\hat{\theta}_{n_b}-\theta\xrightarrow{p}0$ it further suffices to show that $\frac{1}{n_b}\sum_{t=1}^{n_b}\frac{1}{N_q}\frac{\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}-\frac{1}{N_q}\sum_{j=1}^{N_q}\hat{p}_j^{(t-1)}=o_p(1)$. Following the same argument as in the previous item, this is implied by

$$\frac{1}{n_b^2}\sum_{t=1}^{n_b}\mathrm{Var}\left(\frac{1}{N_q}\cdot\frac{\hat{p}_{I_t}^{(t-1)}}{q_t(I_t)}\mid\mathcal{F}_{t-1}\right)\xrightarrow{p}0,$$

which is precisely Assumption A.3.

□

A.6 Proof of Theorem 3.2

Replacing $\hat{p}_j^{(t-1)}$ with our non-time-varying oracle p_j's, our estimator after n_b sampling rounds becomes

$$\hat{\theta}_{n_b,\text{oracle}} = \frac{1}{n_b}\sum_{t=1}^{n_b}\left(\theta^* + \frac{1}{N_q}\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right) = \theta^* + \frac{1}{n_b N_q}\sum_{t=1}^{n_b}\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}.$$

Because θ^* is deterministic, to minimize the variance of $\hat{\theta}_{n_b,\text{oracle}}$, it suffices to minimize the variance of $M_{n_b} = \sum_{t=1}^{n_b}\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}$. Define the filtration $\mathcal{F}_t = \sigma(I_1, z_{I_1}, \ldots, I_t, z_{I_t})$. Writing out the variance, we have the following, noting that because $\hat{\theta}_{n_b,\text{oracle}}$ is unbiased for θ, it must follow that $\mathbb{E}[M_{n_b}] = 0$ and thus

$$\begin{aligned}\text{Var}(M_{n_b}) = \mathbb{E}\left[M_{n_b}^2\right] - \mathbb{E}\left[M_{n_b}\right]^2 = \mathbb{E}[M_{n_b}^2] &= \mathbb{E}\left[\left(\sum_{s=1}^{n_b}\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\left(\sum_{t=1}^{n_b}\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)\right] \\ &= \mathbb{E}\left[\sum_{s=1}^{n_b}\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)^2 + 2\sum_{1\le s<t\le n_b}\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)\right].\end{aligned}$$

Now, for any $s < t$, we deduce the following, because $\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}$ and $q_t(\cdot)$ are $\mathcal{F}_{t-1}$-measurable:

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right) \mid \mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\mathbb{E}\left[\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right) \mid \mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\sum_{j=1}^{N_q}\mathbb{E}\left[\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right) \mid I_t = j, \mathcal{F}_{t-1}\right]P(I_t = j \mid \mathcal{F}_{t-1})\right] \\ &= \mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\sum_{j=1}^{N_q}\mathbb{E}\left[\left(\frac{z_j-p_j}{q_t(j)}\right) \mid I_t = j, \mathcal{F}_{t-1}\right]q_t(j)\right] \\ &= \mathbb{E}\left[\left(\frac{z_{I_s}-p_{I_s}}{q_s(I_s)}\right)\sum_{j=1}^{N_q}\mathbb{E}\left[z_j - p_j \mid I_t = j, \mathcal{F}_{t-1}\right]\right] = 0,\end{aligned}$$

because for all j, $\mathbb{E}\left[z_j - p_j \mid I_t = j, \mathcal{F}_{t-1}\right] = \mathbb{E}\left[z_j - p_j\right] = 0$ by our oracle setup assumptions. Thus, by linearity, conditional independences, and variance of the Bernoulli, we have the following, noting that $q_t(\cdot)$ is $\mathcal{F}_{t-1}$-measurable:

$$\begin{aligned}\text{Var}(M_{n_b}) &= \sum_{t=1}^{n_b}\mathbb{E}\left[\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)^2\right] = \sum_{t=1}^{n_b}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)^2 \mid \mathcal{F}_{t-1}\right]\right] \\ &= \sum_{t=1}^{n_b}\mathbb{E}\left[\sum_{j=1}^{N_q}\mathbb{E}\left[\left(\frac{z_{I_t}-p_{I_t}}{q_t(I_t)}\right)^2 \mid I_t = j, \mathcal{F}_{t-1}\right]P(I_t = j \mid \mathcal{F}_{t-1})\right] \\ &= \sum_{t=1}^{n_b}\mathbb{E}\left[\sum_{j=1}^{N_q}\mathbb{E}\left[\left(\frac{z_j-p_j}{q_s(j)}\right)^2 \mid I_t = j, \mathcal{F}_{t-1}\right]q_t(j)\right] \\ &= \sum_{t=1}^{n_b}\mathbb{E}\left[\sum_{j=1}^{N_q}\frac{\mathbb{E}\left[(z_j-p_j)^2 \mid I_t = j, \mathcal{F}_{t-1}\right]}{q_t(j)}\right] \\ &= \sum_{t=1}^{n_b}\mathbb{E}\left[\sum_{j=1}^{N_q}\frac{p_j(1-p_j)}{q_t(j)}\right].\end{aligned}$$

At this point, note that we are minimizing a sum over t of quantities that are nonnegative, as the p_j are nonnegative and the $q_t(\cdot)$ are strictly positive by definition. From our final expression, there is also no temporal dependency between $q_s(\cdot)$ and $q_t(\cdot)$ for $s < t$. Finally, once provided with our choice of candidate $q_t(\cdot)$, the quantities inside the expectations are now deterministic. Combined, these three points tell us that (a) minimizing $\text{Var}(M_{n_b})$ entails minimizing the instantaneous variance contributions $\sum_{j=1}^{N_q} p_j(1-p_j)/q_t(j)$ at each time step s; and (b) this instantaneous variance contribution minimization is the same across all timesteps. Specifically, at each time step t, we want to solve the following minimization problem, which is strictly convex because $p_j(1-p_j)$ are all nonnegative, and the $q_t(j)$ are all strictly positive:

$$\min_{q_t(\cdot)} \sum_{j=1}^{N_q} \frac{p_j(1-p_j)}{q_t(j)} \quad \text{s.t.} \quad q_t(j) > 0 \; \forall j, \; \sum_{j=1}^{N_q} q_t(j) = 1.$$

Treating $q_t(\cdot)$ as a vector and setting up the Lagrangian, we have

$$L(q_t(\cdot), \lambda) = \sum_{j=1}^{N_q} \frac{p_j(1-p_j)}{q_t(j)} + \lambda \left(\sum_{j=1}^{N_q} q_t(j) - 1 \right).$$

Partial-differentiating to obtain first-order conditions, we have

$$\frac{\partial L(q_t(\cdot), \lambda)}{\partial q_t(j)} = -\frac{p_j(1-p_j)}{q_t(j)^2} + \lambda = 0 \implies q_t^*(j)^2 = \frac{p_j(1-p_j)}{\lambda} \implies q_t^*(j) \propto \sqrt{p_j(1-p_j)},$$

exactly as desired, and emphasizing that the oracle-optimal $q_t^*(j)$ are not time-varying. Because of the strict convexity of our minimization problem, such a solution is the unique minimizer.

A.7 Derivation of (Approximate) Maximal Variance-Reducing Index

Under the idealized assumptions of Section 3.4, suppose at time t that

$$u \mid \mathcal{F}_{t-1} \sim \mathcal{N}\left(\hat{u}^{(t-1)}, \hat{\Sigma}^{(t-1)} \right),$$

and recall that

$$\theta^* := g(u) = \frac{1}{N_q} \sum_{j=1}^{N_q} \sigma\left(u^\top v_j \right).$$

Under our Laplace approximation update derivation setup in Appendix A.4, we know that with $\hat{p}_j^{(t-1)} := \sigma\left(\hat{u}^{(t-1)\top} v_j \right)$,

$$\begin{aligned} u \mid \mathcal{F}_{t-1}, z_j \\ \dot{\sim} \mathcal{N}\Bigg(\hat{u}^{(t-1)} + \left(\hat{\Sigma}^{(t-1)} - \frac{\hat{p}_j^{(t-1)}(1-\hat{p}_j^{(t-1)}) \hat{\Sigma}^{(t-1)} v_j v_j^\top \hat{\Sigma}^{(t-1)}}{1 + \hat{p}_j^{(t-1)}(1-\hat{p}_j^{(t-1)}) v_j^\top \hat{\Sigma}^{(t-1)} v_j} \right) \left(z_j - \hat{p}_j^{(t-1)} \right) v_j, \\ \hat{\Sigma}^{(t-1)} - \frac{\hat{p}_j^{(t-1)}(1-\hat{p}_j^{(t-1)}) \hat{\Sigma}^{(t-1)} v_j v_j^\top \hat{\Sigma}^{(t-1)}}{1 + \hat{p}_j^{(t-1)}(1-\hat{p}_j^{(t-1)}) v_j^\top \hat{\Sigma}^{(t-1)} v_j} \Bigg). \end{aligned} \tag{14}$$

Define $\hat{w}_j^{(t-1)} := \hat{p}_j^{(t-1)} \left(1 - \hat{p}_j^{(t-1)} \right) = \sigma\left(\hat{u}^{(t-1)\top} v_j \right) \left(1 - \sigma\left(\hat{u}^{(t-1)\top} v_j \right) \right)$. Then,

$$\text{Var}\left(u \mid \mathcal{F}_{t-1}, z_j \right) = \hat{\Sigma}^{(t-1)} - \frac{\hat{w}_j^{(t-1)} \hat{\Sigma}^{(t-1)} v_j v_j^\top \hat{\Sigma}^{(t-1)}}{1 + \hat{w}_j^{(t-1)} v_j^\top \hat{\Sigma}^{(t-1)} v_j}.$$

Now, differentiating and recalling that $\sigma'(z) = \sigma(z)\left(1 - \sigma(z)\right)$, we have

$$\nabla_u g(u) = \frac{1}{N_q} \sum_{j=1}^{N_q} \sigma\left(u^\top v_j \right) \left(1 - \sigma\left(u^\top v_j \right) \right) v_j. \tag{15}$$

By the normal approximation associated with the multivariate Delta method, it follows that

$$\text{Var}\left(\theta^* \mid \mathcal{F}_{t-1}, z_j\right) \approx \left(\nabla_u g(u) \mid_{u=\hat{u}^{(t-1)}}\right)^\top \text{Var}\left(u \mid \mathcal{F}_{t-1}, z_j\right) \left(\nabla_u g(u) \mid_{u=\hat{u}^{(t-1)}}\right).$$

Plugging in $\hat{u}^{(t-1)}$ instead of using the mean parameter update in Equation 14 because we do not have access to z_j yet (as we have not queried our next point yet), by a similar argument as above we have

$$\text{Var}\left(\theta^* \mid \mathcal{F}_{t-1}\right) \approx \left(\nabla_u g(u) \mid_{u=\hat{u}^{(t-1)}}\right)^\top \text{Var}\left(u \mid \mathcal{F}_{t-1}\right) \left(\nabla_u g(u) \mid_{u=\hat{u}^{(t-1)}}\right).$$

Subtracting, we obtain the following approximate variance reduction after substituting $\hat{u}^{(t-1)}$ into Equation 15:

$$\begin{aligned}
&\text{Var}\left(\theta^* \mid \mathcal{F}_{t-1}\right) - \text{Var}\left(\theta^* \mid \mathcal{F}_{t-1}, z_j\right) \\
&\approx \frac{1}{N_q^2} \left(\sum_{j'=1}^{N_q} \hat{w}_{j'}^{(t-1)} v_{j'}\right)^\top \left(\frac{\hat{w}_j^{(t-1)} \hat{\Sigma}^{(t-1)} v_j v_j^\top \hat{\Sigma}^{(t-1)}}{1 + \hat{w}_j^{(t-1)} v_j^\top \hat{\Sigma}^{(t-1)} v_j}\right) \left(\sum_{j'=1}^{N_q} \hat{w}_{j'}^{(t-1)} v_{j'}\right) \\
&= \frac{1}{N_q^2} \frac{\hat{w}_j^{(t-1)} \left(v_j^\top \hat{\Sigma}^{(t-1)} \sum_{j'=1}^{N_q} \hat{w}_{j'}^{(t-1)} v_{j'}\right)^2}{1 + \hat{w}_j^{(t-1)} v_j^\top \hat{\Sigma}^{(t-1)} v_j}.
\end{aligned}$$

Thus, the index I_t that upon being queried most reduces the (approximate) posterior variance of θ^* is

$$I_t = \underset{j}{\text{argmax}} \frac{\hat{w}_j^{(t-1)} \left(v_j^\top \hat{\Sigma}^{(t-1)} \frac{1}{N_q} \sum_{j'=1}^{N_q} \hat{w}_{j'}^{(t-1)} v_{j'}\right)^2}{1 + \hat{w}_j^{(t-1)} v_j^\top \hat{\Sigma}^{(t-1)} v_j}.$$

B Investigating Sampling With/Without Replacement

A natural question is whether we can sample questions *without replacement*, since re-querying the same question appears wasteful. An instinctive ad-hoc implementation is to "zero out" the sampling probabilities of previously-queried questions and renormalize before drawing the next index. However, making this change while keeping the same PAI estimator and variance estimate is not theoretically supported and can be empirically harmful at moderate-to-large budgets. Recall that Theorem 3.1 was proven via a martingale CLT and that our proof treated $\Delta_t := \phi_t - \theta$, with

$$\phi_t = \frac{1}{N_q}\sum_{j=1}^{N_q} \hat{p}_j^{(t-1)} + \frac{1}{N_q}\frac{z_{I_t} - \hat{p}_{I_t}^{(t-1)}}{q_t(I_t)},$$

as martingale differences/increments with respect to the filtration $\mathcal{F}_t$. With with-replacement sampling, at each round we draw our next question $I_t \sim q_t(\cdot)$ (c.f. our hybrid sampling policy in Eq. 4) over the full question bank, and the inverse-probability correction in ϕ_t uses $q_t(I_t)$. This alignment is what enables our martingale structure and makes Assumptions A.1–A.2 plausible: as our model-factor estimate $(\hat{u}^{(t)}, \hat{\Sigma}^{(t)})$ stabilizes as we observe more question labels, the predicted probabilities $\hat{p}^{(t)}$ and, by extension, the sampling policy $q_t(\cdot)$, also stabilize. This is important because now it is reasonable that the average conditional variance $n_b^{-1}\sum_{t=1}^{n_b} \mathrm{Var}(\Delta_t \mid \mathcal{F}_{t-1})$ converges to a finite limit v^2—the crux of Assumption A.1.

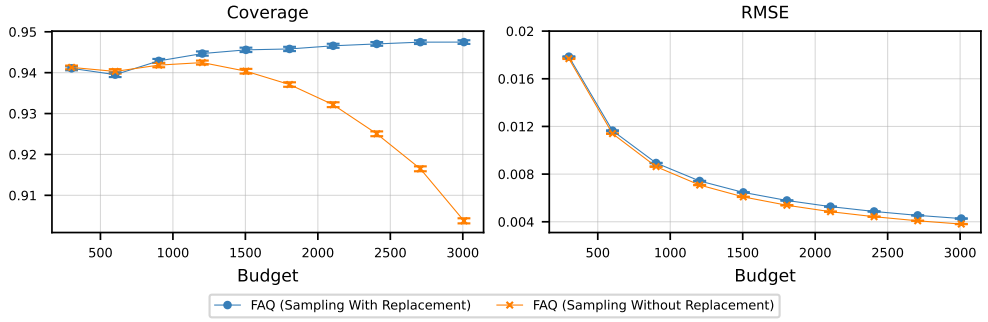

Figure 6: **Coverage and RMSE of FAQ with/without replacement on MMLU-Pro with fully-observed historical data. (Left)** coverage and **(Right)** RMSE of $\hat{\theta}_{n_b}$ as a function of budget n_b, for FAQ sampling with/without replacement. Standard errors are shown, but negligible on the plot scale. Targeted nominal coverage for all experiments shown was 95%.

Under the ad-hoc without-replacement modification, we are no longer drawing the next question from q_t: rather, we are drawing from the restriction of q_t to the remaining unseen questions R_{t-1}—let us call this restriction $\tilde{q}_t$. The key point is that $\tilde{q}_t(\cdot)$ is no longer aligned with PAI's estimator and variance estimate formulas—which both still use $q_t(I_t)$. As such, we no longer have martingale differences,[13] thus hindering our martingale CLT argument and undermining our regularity conditions in Assumptions A.1–A.2. For one, with our ad-hoc without-replacement modification, our $q_t(\cdot)$ are no longer lower-bounded, leading to unbounded inverse-probability corrections $1/q_t(\cdot)$ and challenging Assumption A.2. Without access to the martingale CLT, our resulting normal approximation and variance estimate can become miscalibrated.

Empirically, Figure 6 (Left) corroborates this: at moderate-to-large budgets, the ad-hoc without-replacement variant suffers from substantial miscoverage at moderate to large budgets. At low budgets, the difference in coverage is smaller because repeat sampling of questions is rarer, and thus the two sampling schemes behave similarly. We also observe from Figure 6 (Right) that the ad-hoc without-replacement modification yields little improvement in RMSE (point estimate performance).

For these reasons, we recommend sampling with replacement when using FAQ/PAI to preserve theoretically-guaranteed asymptotic uncertainty quantification, with minimal loss in point-estimation performance.

[13] A consequence of this is also that $\hat{\theta}_{n_b}$ is no longer unbiased for θ.

C Additional Experimental Details

Datasets We compile our historical data from the HuggingFace Open LLM Leaderboard [Fourrier et al., 2024, Myrzakhan et al., 2024], which is licensed under Apache-2.0. We create one benchmark from MMLU-Pro [Wang et al., 2024] itself with $N_q = 12032$ questions, and aggregate the remaining datasets—BBH [Suzgun et al., 2023], GPQA [Rein et al., 2024], IFEval [Zhou et al., 2023], MATH [Hendrycks et al., 2021], and MuSR [Sprague et al.]—into a second benchmark with $N_q = 9574$ questions.

Historical splits and missingness For non-cold-start experiments comparing FAQ to the strongest baselines and uniform sampling, we consider the fully-observed historical data setting and the following $(n_{\text{full-obs}}, p_{\text{obs}})$ pairs: $(50, 0.1)$, $(200, 0.1)$, $(800, 0.1)$, and a particularly extremely-sparse $(0, 0.001)$ setting.

For our traditional active inference ablation experiments, we consider the fully-observed data setting and the following $(n_{\text{full-obs}}, p_{\text{obs}})$ pairs: $(0, 0.01)$, $(0, 0.001)$, $(0, 0.0001)$, and $(0, 0.00001)$. We chose this range of missingness settings after observing the results of FAQ against the strongest baselines and uniform sampling, and desired missingness-settings that would best distinguish between FAQ and the traditional active inference ablation.

Factor model selection For non-cold-start experiments, we use the AdamW optimizer with weight decay λ, learning rate 10^{-2}, and maximum iterations 2000, implemented in PyTorch [Paszke et al., 2019], with all experiments run on either a single NVIDIA A100 (40 GB) or NVIDIA H100 (80 GB). We sweep over latent factor dimension $k \in \{8, 16, 32, 64\}$ and weight decay $\lambda \in \{10^1, 10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

FAQ hyperparameters For all experiments, we sweep over the maximum tempering exponent $\beta_0 \in \{0.25, 0.5, 0.75, 1.0\}$, the tempering governor $\gamma \in \{0.0, 0.05, 0.25, 0.5, 0.75\}$, the exploration-to-exploitation governor $\rho \in \{0.0, 0.05, 0.25, 0.5, 0.75\}$, and the uniform-mixing strength $\tau \in \{0.05, 0.25, 0.5, 0.75\}$.

Baselines and active inference ablation We sweep over uniform mixing strength $\tau \in \{0.05, 0.25, 0.5, 0.75\}$.

Cold-start and cross-suite transfer Because no pre-existing target-suite (i.e., GPQA (All)) historical matrix is available for factor-model selection, we fit the initial factor model only on fully-observed MMLU-Pro historical data, using a fixed a priori configuration with factor dimension $k = 8$, weight decay $\lambda = 1.0$, and 2000 AdamW iterations. We then freeze the MMLU-Pro-fitted historical model factors u_i and use their empirical mean and covariance to initialize the prior for new models' factors on GPQA (All). To initialize GPQA (All) question factors, we compute embeddings of all MMLU-Pro and GPQA (All) question texts using OpenAI's `text-embedding-3-large` model. For each GPQA (All) question, we find its 8 nearest MMLU-Pro questions using cosine similarity in embedding space, and set its question factor to the corresponding similarity-weighted average of their MMLU-Pro question factors. FAQ sampling hyperparameters are tuned analogously to the procedures described above and in the main text, using a held-out set of 443 validation models that we assume are available to query, and minimizing CI width only. These validation models are excluded from final evaluation: coverage and efficiency are evaluated only on the 2.2K test models.

All source code and analysis routines for reproducing all results in this paper can be found at `https://github.com/skbwu/efficiently-evaluating-llms`.

D Traditional Active Inference Ablation

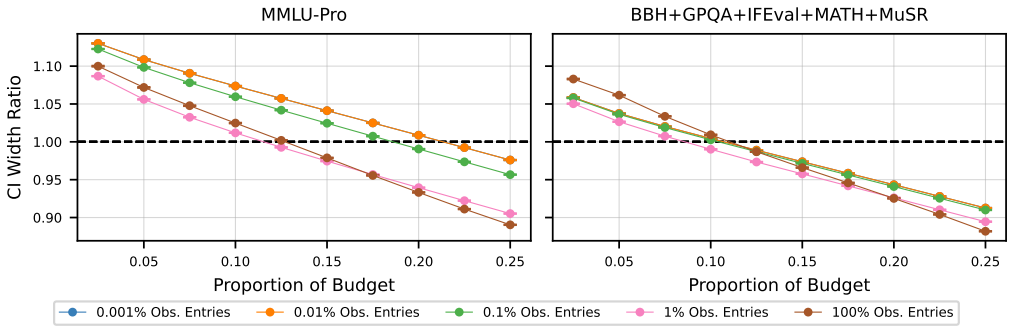

Figure 7: **CI-width ratio (post-hoc best traditional active inference ablation / FAQ) across budgets/missingness** Missingness uses $n_{\text{full-obs}} = 0$ and MCAR p_{obs} (legend values) on the remaining entries. Standard errors are shown, but negligible at the plot scale. **Width ratios > 1 favor FAQ (narrower CIs).**

To isolate the gains of PAI and our hybrid sampling (beyond the factor model), we compare FAQ to a "traditional" active inference ablation that follows Zrnic and Candès [2024]'s sequential ("nature's order")[14] procedure while using the *same* factor-model predictions. We report CI-width ratios (rather than ESS) since they compare tightness directly across budgets without converting to a uniform-equivalent "budget savings" scale. Figure 7 reveals a clear phase shift. At low-budgets (e.g, $< 10\%$ of the bank), FAQ yields substantially narrower CIs; at larger budgets, the traditional ablation eventually overtakes FAQ by similar margins. This is not discouraging: large-scale evaluation (1K+ models, 10K+ questions) typically emphasizes the low-budget regime where FAQ is strongest. It also suggests a natural extension—a three-stage hybrid policy: active-learning early, oracle-style variance reduction mid-budget, and traditional active inference at the end—combining the best of both worlds.

Three mechanisms likely drive the shift. First, FAQ can proactively target informative questions early on via its active-learning component, whereas traditional active inference can only label/skip items in a fixed stream and its $\pi_t(t) \propto \min(\hat{p}_t^{(t-1)}, 1 - \hat{p}_t^{(t-1)})$ rule contains no explicit active-learning objective—explaining the low-budget gap. Second, at large budgets, FAQ's sampling with replacement can resample questions, reducing efficiency relative to stream-based procedures that touch each question at most once. Third, our tuned τ-mixing suggests variance control dominates at high budgets: the best post-hoc traditional ablation consistently uses $\tau = 0.75$ (near-uniform query/label probabilities). By contrast, FAQ's train/val tuning often selects much smaller τ (e.g., 0.05 for many budgets $\geq 10\%$), yielding spikey/highly-concentrated q_t. After the factor model has largely converged, this extra "targeting" offers little learning benefit, but increases AIPW variance through larger weights $1/q_t(I_t)$, helping explain why the traditional ablation can overtake FAQ at large budgets. Because FAQ targets cost savings, we focus on the low-budget regime.

[14] Here, "nature's order" refers to the benchmarks' native question orderings in HuggingFace's Open LLM Leaderboard.

E Additional Results

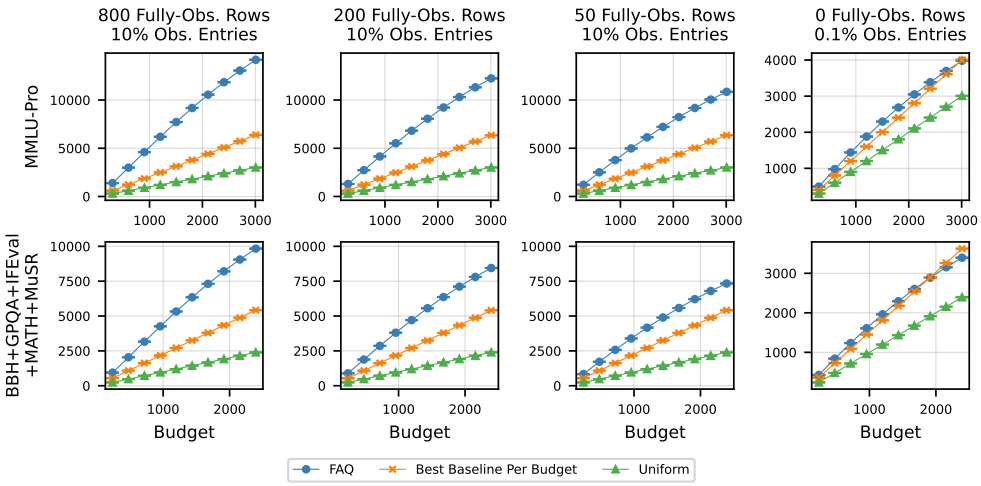

Figure 8: **ESS under missing historical data on MMLU-Pro and BBH+GPQA+IFEval +MATH+MuSR.** ESS vs. budget for FAQ, the strongest baseline per budget (selected post-hoc), and uniform sampling, per missingness setting. Missingness is parameterized by $n_{\text{full-obs}}$ fully-observed historical rows (out of 2.2K) and MCAR entrywise observation probability p_{obs} on the remaining rows (e.g., $n_{\text{full-obs}} = 50$ and $p_{\text{obs}} = 0.1 \sim 12\%$ observed). Standard errors are shown, but negligible on the plot scale. This figure shows the full results corresponding to Figure 3 in the main text.

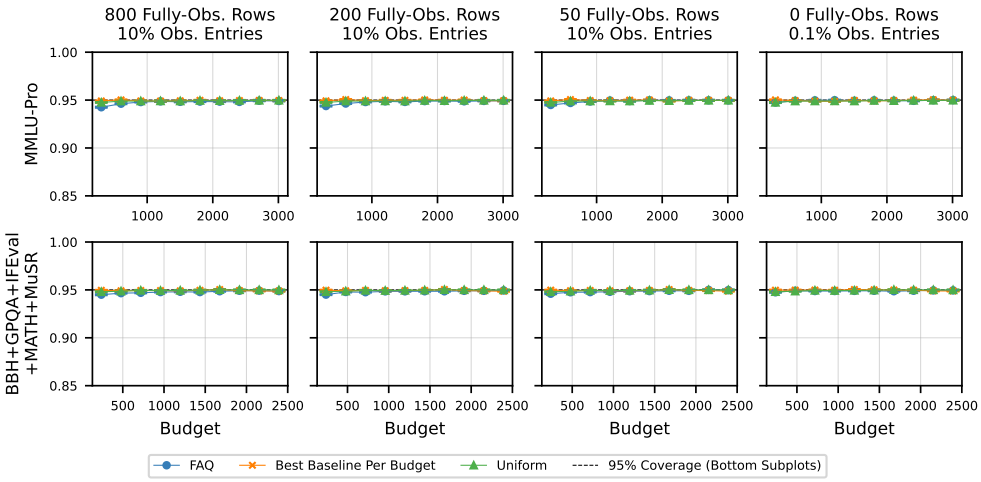

Figure 9: **Coverage under missing historical data on MMLU-Pro and BBH+GPQA+IFEval +MATH+MuSR.** Coverage vs. budget for FAQ, the strongest baseline per budget (selected post-hoc), and uniform sampling, per missingness setting. Missingness is parameterized by $n_{\text{full-obs}}$ fully-observed historical rows (out of 2.2K) and MCAR entrywise observation probability p_{obs} on the remaining rows (e.g., $n_{\text{full-obs}} = 50$ and $p_{\text{obs}} = 0.1 \sim 12\%$ observed). Standard errors are shown, but negligible on the plot scale. This figure confirms that the results in Figure 3 in the main text were achieved with valid coverage.

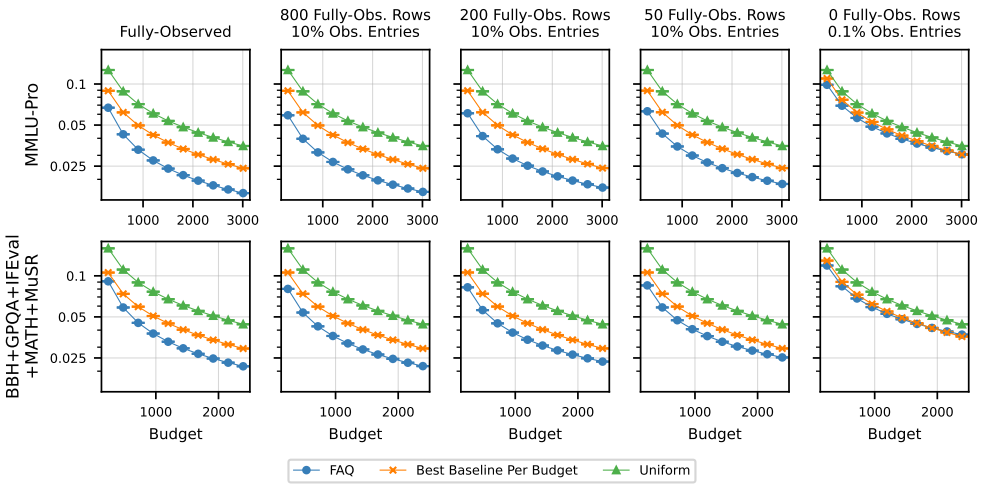

Figure 10: **Confidence interval (CI) widths under missing historical data on MMLU-Pro and BBH+GPQA+IFEval+MATH+MuSR.** CI width vs. budget for FAQ, the strongest baseline per budget (selected post-hoc), and uniform sampling, per missingness setting. Missingness is parameterized by $n_{\text{full-obs}}$ fully-observed historical rows (out of 2.2K) and MCAR entrywise observation probability p_{obs} on the remaining rows (e.g., $n_{\text{full-obs}} = 50$ and $p_{\text{obs}} = 0.1 \sim 12\%$ observed). Standard errors are shown, but negligible on the plot scale. **Smaller CI widths indicate stronger performance.**

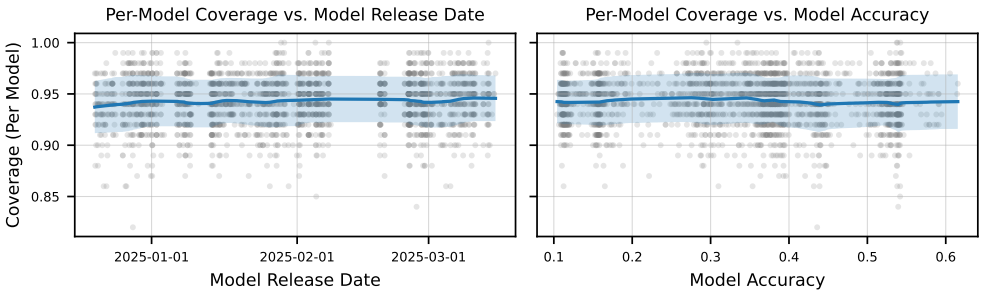

Figure 11: **Per-model coverages on MMLU-Pro with fully-observed historical data at budget** 7.5%**.** Per-model coverages (averaged over 100 seeds) vs. **(Left)** model release date and **(Right)** true model accuracy. Gray dots: individual models; solid-blue curve: locally-weighted mean with a $K = 501$ nearest-neighbor Gaussian kernel; light-blue band: local ± 1 SD band. This figure shows the MMLU-Pro counterpart to Figure 4 in the main text.

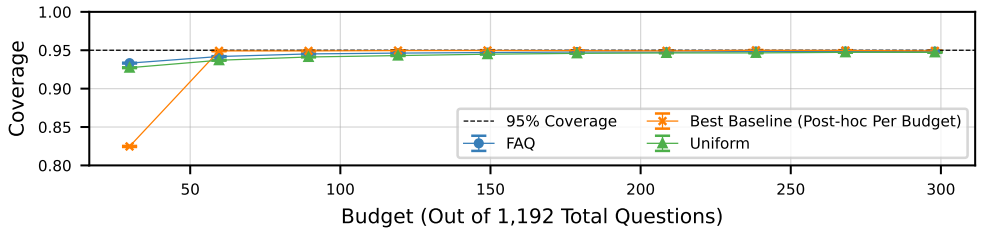

Figure 12: **Coverage for cold-start FAQ on GPQA (All).** Coverage vs. budget for cold-start FAQ *with zero target-suite historical data for factor-model fitting*, the strongest post-hoc baseline per budget *with access to* 0.1% *target-suite historical data*, and uniform sampling. Standard errors are shown, but negligible on the plot scale. This figure confirms that main text Figure 5's results were achieved with valid coverage.

F Additional Exploratory Robustness Experiments

In this section, we report additional exploratory experiments examining FAQ's robustness to structured (as opposed to MCAR) historical-data missingness and historical-label corruption, as well as its sensitivity to factor-model predictive accuracy. Given their exploratory nature, we focus on the BBH+GPQA+IFEval+MATH+MuSR suite ($N_q = 9574$) at three budgets, where applicable: 2.5%, 7.5%, and 12.5%. For the historical-data perturbation experiments, we retain the full FAQ pipeline with proper train/validation/test splits, using the same perturbed historical data H for all baselines, and 100 randomized seeds per setting. We organize our specific experiments by research question.

How does FAQ perform under historical-data missingness by model family? We removed all fine-tuned Qwen2.5 and QwQ (i.e., Qwen2.5's thinking/reasoning variant) variants from H. We chose the Qwen2.5/QwQ family because it was the largest base model family, compared to Llama-3.1 and Gemma-2. This constituted the removal of 131 of the 2.2K historical models. On the 321 Qwen2.5/QwQ-family test models at budgets 2.5%, 7.5%, and 12.5%, FAQ achieved ESS multipliers over uniform sampling of 2.54, 3.31, and 3.57, respectively, outperforming the best post-hoc selected baseline, whose ESS multipliers never exceeded 2.02, while maintaining coverage within 0.6% of the 95% target. Thus, FAQ can pool information across families to remain effective when a major family is absent from H.

How does FAQ perform under historical missingness by release date? We sorted H by release date and removed its newest 50% of models, an extreme perturbation given that our original historical/test split was already chronological. Nonetheless, across all 2.2K test models at budgets 2.5%, 7.5%, and 12.5%, FAQ achieved ESS multipliers over uniform sampling of 3.32, 3.59, and 3.58, respectively, outperforming the best post-hoc selected baseline, whose ESS multipliers never exceeded 2.23, while maintaining coverage within 0.6% of the 95% target.

How does FAQ perform under random label-flipping corruption? We uniformly flipped 5% of labels in the historical data as a pessimistic stress test. Across all 2.2K test models at budgets 2.5%, 7.5%, and 12.5%, FAQ achieved ESS multipliers over uniform sampling of 3.85, 4.30, and 4.36, respectively, outperforming the best post-hoc selected baseline, whose ESS multipliers never exceeded 2.28, while maintaining coverage within 0.5% of the 95% target. These results are slightly better than those in Figure 2. One possible explanation is that random label-flipping regularizes the factor-model fit and/or tempers extreme probabilistic predictions, improving test-set probability calibration and hence sampling efficiency.

How does FAQ perform under benchmark-specific label corruption? We concentrated the same 5% overall corruption on BBH, the largest benchmark (5761 questions), corresponding to 8.30% corruption within BBH. Across all 2.2K test models at budgets 2.5%, 7.5%, and 12.5%, FAQ achieved ESS multipliers over uniform sampling of 3.75, 4.17, and 4.22, respectively, outperforming the best post-hoc selected baseline, whose ESS multipliers never exceeded 2.27, while maintaining coverage within 0.5% of the 95% target. These results are also slightly better than those in Figure 2, plausibly for the same reason.

How does FAQ perform under model-family-specific label corruption? We concentrated the 5% overall corruption rate on the 131 historical Qwen2.5/QwQ-family models, corresponding to 84.3% corruption within this family. On the 321 Qwen2.5/QwQ-family test models at budgets 2.5%, 7.5%, and 12.5%, FAQ achieved ESS multipliers over uniform sampling of 2.69, 3.47, and 3.71, respectively, outperforming the best post-hoc selected baseline, whose ESS multipliers never exceeded 2.03, while maintaining coverage within 0.6% of the 95% target.

How robust are FAQ's empirical efficiency gains under factor model misspecification? Because the true response model is unknown, we quantify practical misspecification using pre-query binary prediction accuracy: the fraction of question outcomes correctly predicted before observing any new-model labels. On BBH+GPQA+IFEval+MATH+MuSR with fully observed H, we stratified the 2.2K test models into quartiles by this accuracy. Predictions use the empirical normal prior mean computed from the historical model factors: lower accuracy therefore indicates weaker transfer to the new model. At a 7.5% budget, FAQ achieved mean ESS multipliers over uniform of 2.64 in the worst-predicted quartile and 5.43 in the best-predicted quartile, averaged over models within each quartile.

Their mean pre-query accuracies were 66.21% and 80.20%, respectively. FAQ also outperformed the strongest post-hoc selected baseline in every quartile. Across quartiles, coverage ranged from 94.64% to 94.84%, near the 95% target. Thus, poorer factor predictions reduce attainable efficiency gains, but FAQ remains more efficient than the strongest baselines while preserving near-nominal coverage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see our abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see our "Empirical Results" and "Discussion and Conclusion" sections.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see our “Method" section and full proofs in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see our “Experimental Setup" section and Appendix C. We also include full source code to reproduce all experimental results in our supplementary ZIP file.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Please see our full source code for all experiments and curated dataset files in our supplementary ZIP file.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see our full source code for all experiments and curated dataset files in our supplementary ZIP file, which also contains detailed documentation. All results figures in our paper can be reproduced directly via our source code. Please also see written experimental details in our “Experimental Setup" section and in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: When possible, all results figures contain error bars and/or confidence bands.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see our "Empirical Results" section and the README of our supplementary ZIP file.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully read the NeurIPS Code of Ethics and have made sure that our research follows it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper introduces a statistical method for adaptively selecting questions for evaluating and estimating LLM performance. As such, it does not have direct paths to negative societal impacts, aside from the general interactions of machine learning, AI, and statistical methods with society.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not include any data or models that have high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see our Appendix C.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please see our supplementary ZIP file. Specifically, please see our README.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The results in this paper do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The results in this paper do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our experiments are mainly from publicly-available data from HuggingFace's Open LLM Leaderboard. We use OpenAI's `text-embedding-3-large` model for our cold-start and cross-suite transfer experiment, which we detail in Appendix C and our supplementary ZIP file.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.