Valid Best-Model Identification for LLM Evaluation via Low-Rank Factorization

Elad Tolochinsky[1] **Yaniv Tenzer**[1] **Yaniv Romano**[1,2]

[1]Department of Computer Science, Technion – Israel Institute of Technology
[2]Department of Electrical and Computer Engineering, Technion – Israel Institute of Technology
elad.t@cs.technion.ac.il, yanivt@technion.ac.il, yromano@technion.ac.il

Abstract

Selecting the best large language model (LLM) for a fixed benchmark is often expensive, since exhaustive evaluation requires running every model on every example. Multi-armed bandit (MAB) algorithms can reduce the number of LLM calls by sequentially selecting the next model-example pair to evaluate, thereby avoiding wasted evaluations on clearly underperforming models. Further savings can be achieved by predicting model scores from the partially observed model–example score matrix using low-rank factorization. However, such predictions are not ground truth: they can be biased and may therefore lead to incorrect identification of the best model. In this work, we propose a principled framework that combines MAB with cheap predicted scores without compromising statistical validity. Specifically, we derive doubly robust estimators of each model's performance that use the low-rank predictions to reduce variance. This enables the construction of valid finite-sample confidence intervals in our setting, where models are selected adaptively and examples are sampled without replacement. Empirical results on real-world benchmarks show that our approach reduces the number of required evaluations, yielding meaningful savings in compute and cost while accurately identifying the best-performing model.

1 Introduction

LLMs are employed across a wide range of applications, including customer-facing dialogue systems, code generation, writing assistance, and information retrieval [1–3]. As LLM adoption grows, practitioners often encounter the challenge of choosing the best-performing model and configuration—such as prompts and hyperparameters—for a specific task. A common practice for that purpose is to compare the performances of the models on fixed benchmarks using a score function. For example, consider the task of mathematical problem solving, as studied in benchmarks such as MMLU [4] and MATH [5]. In this case, a score function can be a binary function that indicates whether the model's solution is correct. Unfortunately, exhaustively evaluating every model on every example is resource-intensive: a single evaluation on the GAIA benchmark [6] can cost up to $2,829 [7], and the recent Holistic Agent Leaderboard evaluation spent roughly $40,000 to compare 9 models across 9 benchmarks [8]. Moreover, these are single-run figures: stochastic agents require repeated rollouts to produce reliable estimates, multiplying costs further.

To reduce the evaluation cost, recent works utilize multi-armed bandit (MAB) algorithms, such as UCB-E [9], to adaptively allocate the limited evaluation budget. Ideally, MAB should spend more budget on promising candidates and less on clearly underperforming ones. In this framework, each model is an arm, and pulling an arm corresponds to evaluating the model on examples and observing the scores. The evaluation proceeds sequentially: at each step, the algorithm selects a model and a batch of examples that this model has not yet evaluated, computes their scores, and updates its

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

estimate of the model's mean accuracy. These updated estimates are then used to guide the selection of the next model.

Further computational savings can be achieved by using cheap predicted scores rather than the exact ones. Such predictions can be found in the structure of the model-example score matrix, i.e., the matrix whose (i, j) entry is the score obtained by running the i-th model on the j-th example. Concretely, the work in [10] utilizes the fact that models may exhibit similar performance profiles across examples, and examples may admit similar scores across models. This correlation suggests that the full score matrix may be well approximated by a low-rank matrix. During evaluation, this matrix is only partially observed. Yet, low-rank factorization provides a cheap way to impute the missing entries, producing predicted scores that can guide MAB's evaluation procedure. Indeed, it is shown in [10] that such predictions can accelerate the identification of the best model.

The challenge is that these predicted scores are not ground truth. They can be biased and may therefore lead to incorrect identification of the best model. This raises the need for a principled framework that incorporates predicted scores while maintaining unbiased estimates of the models' mean performance. To this end, we build on doubly robust estimators—including augmented inverse probability weighting (AIPW) [11, 12] and prediction-powered inference (PPI) [13, 14]—that enable the construction of unbiased estimators of population parameters with potentially lower variance by leveraging machine learning (ML) predictions. These tools have been developed for adaptive MAB, contextual bandits, and even best-arm identification [15–18]. However, it is unclear how to extend the existing guarantees for these techniques to the finite-population setting considered in this paper. Beyond sampling examples without replacement from a fixed benchmark, our setting introduces an additional complication: the imperfect predictions are obtained by repeatedly refitting an online low-rank model to the same adaptively observed model–example score matrix, resulting in strong cross-arm dependence.

In this work, we bridge the gap by presenting PULSE: Prediction-Powered Unbiased Low-Rank Sequential Evaluation. Given a fixed budget of LLM calls, PULSE leverages imperfect predicted scores to increase the probability of identifying the best-performing model. Importantly, our approach remains statistically valid even when the predictions are inaccurate. As we show, when predictions are informative, our method accelerates identification; when they are not, it can approach standard MAB behavior that relies solely on observed evaluations.

Our main contributions are as follows:

- Our first key contribution (Section 3) is an estimator of each model's mean accuracy that leverages predictions obtained by low-rank factorization of the score matrix. The resulting estimator is guaranteed to be unbiased, while informative predictions can reduce its variance. We also show how to construct confidence intervals (CIs) for the models' mean accuracies and establish their finite-sample validity.
- Building on our theoretical results, our second key contribution (Section 3) is a MAB algorithm for best-model identification. We establish a finite-sample success guarantee for PULSE and analyze how prediction quality can affect its convergence.
- Finally, we evaluate the performance of our algorithm on 2.2K models and six real-world benchmarks. As shown in Figure 1, our experiments demonstrate that PULSE can save up to 46% of LLM calls relative to UCB-E when targeting 95% identification accuracy. Code: `https://github.com/elad-tolo/pulse-llm-eval`.

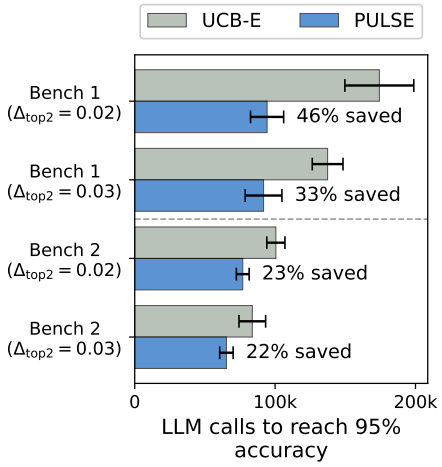


Figure 1: Budget needed for 95% best-model identification accuracy (lower is better). Bench 1 and 2 cover 2.2K models each. Δ_{top2}, the gap between the top two models, controls difficulty. See Section 4.

2 Problem Formulation and Related Work

We let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a fixed dataset of examples, $\mathcal{M} = \{f_1, \ldots, f_m\}$ be a set of LLMs. Let $s(M(x)) \in \{0, 1\}$ be a scoring function used to score the output of the LLMs. Without loss of generality, we assume that $s(f(x)) > s(f'(x))$ indicates that model f has better performance than model f' on example x. We denote by $S \in \{0, 1\}^{m \times n}$ the score matrix so that $S_{ij} := s(f_i(x_j))$. In our setting, S is only partially observed, and we think of its entries as random variables. More precisely, let S_i be the score obtained by model i when tested on an example that is drawn uniformly from $\mathcal{X}$. Under this formulation, $\mu_i = \mathbb{E}[S_i] = \frac{1}{n} \sum_{j=1}^{n} S_{i,j}$, where the expectation is taken with respect to the random choice of the example. In contrast with an infinite population setting, in which the sample average is only an estimate for the population mean, here the sample average is the actual expectation of S^i.

Given a budget of T LLM calls, we consider an adaptive algorithm $\mathcal{A}(T, \mathcal{M}, \mathcal{X})$ that returns an index $\hat{i} \in [m]$, where $[m] = \{1, \ldots, m\}$. Our goal is to maximize the probability of correctly identifying the best model $i^* = \arg\max_{i \in [m]} \mu_i$, which we assume is unique. We denote this probability by $P(\hat{i} = i^*)$, taken over the randomness of the adaptive sampling procedure. Our algorithm follows the UCB-E method [9]. At each round t, the algorithm maintains CIs for the mean accuracy of each model and selects the model with the largest upper confidence bound. It then evaluates the selected model on a batch of previously unseen examples. Using the newly observed scores, the algorithm updates the estimate of the selected model's mean accuracy and its corresponding CI.

The convergence rate of the UCB-E algorithm to the optimal arm (best model) depends on the miscoverage rate of the CIs [9], which decreases with the number of observations. This motivates us to artificially increase the number of samples by predicting the unobserved entries of S using a low-rank factorization that leverages the correlation between the models.

What makes UCB-E nonstandard in our finite-dataset setting is that the quantities used to construct the confidence bounds are both adaptive and non-stationary. The predicted scores are obtained by repeatedly applying a low-rank factorization to the partially observed score matrix, and hence depend on the entire history of previous evaluations. At the same time, the observed scores are also not i.i.d as they are sampled without replacement.

Related Work

There is a rich line of research on the use of doubly robust estimators in the context of MAB [11, 12, 15–23]. Here, we discuss the works that are most closely related to ours. In [24], the authors show how surrogate rewards can improve the sample efficiency of MAB by using PPI. However, they assume that the rewards' distribution of each arm is Gaussian and stationary. In contrast, we assume that the rewards population is finite (i.e., the entries of S), and changes with each step, as we sample scores from S without replacement.

Another closely related work to ours is [25], which studies the problem of best-arm identification while using LLM judges and limited human audits. In contrast to our setting, this work assumes a stream of examples and scores that follows a stationary data-generating distribution, whereas we assume access to a finite number of examples, i.e., the complete dataset. Since we sample example-score pairs without replacement, the score distribution changes over time.

Finally, the FAQ method [26] considers the problem of efficiently evaluating the accuracy of a single candidate model using low-rank factorization. While the unbiased mean accuracy estimator used in our work follows the structure of the one from [26], we differ with respect to three key aspects. First, the FAQ algorithm samples examples with replacement, while we focus on sampling without replacement. Second, we construct CIs using finite-sample concentration inequalities, while the CIs used in FAQ are only asymptotically valid. Finally, we focus on the MAB setup, where the key challenge is to efficiently identify the best model, whereas [26] focuses on evaluating a single model.

3 Method

In this section, we present our sequential approach for identifying the best-performing model while safely utilizing predicted scores to reduce the number of LLM calls. Our sequential method alter-

Algorithm 1 PULSE: Prediction-Powered Unbiased Low-Rank Sequential Evaluation

Require: Test models $\mathcal{M} = \{f_1, \ldots, f_m\}$, examples $\mathcal{X} = \{x_1, \ldots, x_n\}$, scoring function $s(\cdot) \in \{0, 1\}$; evaluation budget T, batch size B, exploration parameter a; low-rank model $\mathcal{M}$ with rank r, retrain period τ; init pulls per arm T_0; side-information score matrix S^{train}

Ensure: Prediction $\hat{i}^*$ for the best model i^*.

Stage 1: Initialization

Fit $(U_{\text{train}}, V) \leftarrow \mathcal{M}(S^{\text{train}}; r)$ using Step 1; freeze V and discard U_{train}.
for $i = 1, \ldots, m$ **do**
    Sample $J_i \subset [n]$ with $|J_i| = B \times T_0$ and set $S_{i,j} \leftarrow s(f_i(x_j))$ and $O_{i,j} \leftarrow 1 \; \forall j \in J_i$.
end for
$b \leftarrow T - mBT_0$
Refit U on the observed test entries with V frozen; set $\hat{S}^0 \leftarrow \sigma(U^T V)$.
For each i: initialize $\hat{\mu}_i$ and set $t, \hat{\lambda}_i^0 \leftarrow 0$.

Stage 2: bandit loop

for $t = 1, \ldots, b/B$ **do**
    $i \leftarrow \arg\max_{i:|O_i|<n} \hat{\mu}_i + \sqrt{a/|O_i|}$.
    if $t \bmod \tau = 0$ **then**
        Refit U on the current observation mask O with V frozen; set $\hat{S}^t \leftarrow \sigma(U^T V)$
    end if
    compute $\hat{\lambda}_{i_t}^t$ using Step 3.
    sample $B_t \subset U_i^t$ with $|B_t| = B$
    $S_{i,j} \leftarrow s(f_i(x_j))$ and $O_{i,j} \leftarrow 1$ for every $j \in B_t$.
    Update $\hat{\mu}_i^t$
end for
return $\hat{i}^* \leftarrow \arg\max_i \hat{\mu}_i$.

nates between two key steps: (i) predicting the missing entries of the score matrix S via low-rank factorization that utilizes the data evaluated thus far, and (ii) an adaptive MAB step that selects a model–example pair to evaluate, followed by updating the estimates of the models' mean accuracies μ_i and their corresponding CIs. For ease of exposition, we describe the method in the setting where a single example is sampled at each step. Later, we describe how to extend it to batch sampling.

Step 1: Low-Rank Factorization

Given a binary matrix $S \in \{0,1\}^{m \times n}$ and an observation mask $O \in \{0,1\}^{m \times n}$ (with observed index set $\Omega = \{(i,j) : O_{ij} = 1\}$), we model the probability that model $i \in [m]$ answers correctly a question $j \in [n]$ by

$$\hat{S}_{i,j} = \sigma\left(u_i^\top v_j\right), \qquad \sigma(z) = \frac{1}{1 + e^{-z}},$$

where $u_i, v_j \in \mathbb{R}^r,\ 1 \le i \le m,\ 1 \le j \le n$ are learned low-dimensional latent representations of model i and question j, respectively. The inner product $u_i^\top v_j$ is used to predict the log-odds that model i answers question j correctly.

We learn u_i and v_j by optimizing the following objective:

$$\mathcal{L}(U, V) = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \text{BCE}\left(S_{ij}, \sigma(u_i^\top v_j)\right) + \frac{\lambda}{2(m+n)} \left(\|U\|_F^2 + \|V\|_F^2\right), \tag{1}$$

where, for a label $y \in \{0,1\}$ and predicted probability $p \in (0,1)$, $\text{BCE}(y,p) = -y \log p - (1-y)\log(1-p)$. Equivalently, this is the negative log-likelihood of a Bernoulli model in which $\mathbb{P}(S_{ij} = 1 \mid u_i, v_j) = \sigma(u_i^\top v_j)$. Thus, the first term in (1) fits the observed entries of the score matrix using a low-rank logistic model. The regularization term prevents overfitting to the observed entries and promotes generalization to the unobserved entries, which is crucial in our setting since only a fraction of the score matrix is observed. The parameter λ controls the trade-off between fitting the observed scores and regularizing the latent factors.

Step 2: Estimating mean accuracies and constructing CIs

We first fix a model $i \in [m]$ and show how to construct, at each time step t, a conditionally unbiased estimator of μ_i using the data collected during the sequential procedure. We begin by introducing the required notation. Let $O_i^{t-1} \subset [n]$ denote the set of indices of examples evaluated by model i by the end of step $t-1$. Let $U_i^t = [n] \setminus O_i^{t-1}$ denote the set of indices of examples that model i has not evaluated before step t. We denote by $\hat{S}_{i,j}^t$ the predicted score of model i on example j at step t. The superscript t emphasizes that the low-rank factorization is updated over time, so the predicted scores may also change from one step to the next. Finally, since the next example is sampled uniformly from U_i^t, we write $\pi_i^t = 1/|U_i^t|$ for the probability of selecting any particular index $j \in U_i^t$.

To state the estimator and its unbiasedness property precisely, we also need to specify what information is available before the new example is sampled. This is captured by the natural filtration. Formally, let $\{\mathcal{F}_t\}_{t\geq 0}$ denote the filtration, where $\mathcal{F}_t$ is the sigma-algebra generated by all observations collected up to step t. A random variable X is said to be $\mathcal{F}_t$-measurable if it is completely determined by the data available up to step t.

Now suppose that, at step t, model i is selected and an index $j_t \in U_i^t$ is sampled uniformly. We define the following estimator for μ_i:

$$\hat{\theta}_i^t = \frac{1}{n}\Big[\sum_{j \in O_i^{t-1}} S_{i,j} + \sum_{j \in U_i^t} \lambda_i^t \hat{S}_{i,j}^t + \underbrace{\frac{S_{i,j_t} - \lambda_i^t \hat{S}_{i,j_t}^t}{\pi_i^t}}_{Z_i^t} \Big]. \tag{2}$$

The first term sums the scores alow-rankeady observed for model i before step t. The second term uses the predicted scores for the entries not yet observed for model i, scaled by a weight $\lambda_i^t \in [0,1]$ that can be tuned using past data, as discussed below. The final term, denoted by Z_i^t, is an inverse-probability residual correction that removes the bias introduced by using potentially inaccurate predicted scores for the unobserved entries. As the next lemma shows, this correction makes $\hat{\theta}_i^t$ a conditionally unbiased estimator for μ_i.

Lemma 1. *Assume π_i^t, λ_i^t and $\hat{S}_{ij}^t$ are $\mathcal{F}_{t-1}$ measurable, then for each $t \geq 1$, $\mathbb{E}[\hat{\theta}_i^t \mid \mathcal{F}_{t-1}] = \mu_i$*

Importantly, our algorithm satisfies the condition of Lemma 1 by construction. Indeed, at the beginning of step t, both π_i^t and $\hat{S}_{ij}^t$ are fully determined by the data available up to the end of step $t-1$. Hence, both quantities are $\mathcal{F}_{t-1}$-measurable. Similarly, as we discuss below, the weight λ_i^t is chosen using only past data, before observing the new score at step t. Therefore, λ_i^t is also $\mathcal{F}_{t-1}$-measurable.

To better understand the role of λ_i^t, consider the ideal case in which the predictions are exact, i.e., $\hat{S}_{ij}^t = S_{ij}$ for all $j \in U_i^t$. Then, setting $\lambda_i^t = 1$ yields the perfect estimator $\hat{\theta}_i^t \equiv \mu_i$.[1] Since μ_i is a finite-population mean, this perfect estimator has zero variance. At the other extreme, if the predictions are uninformative, the estimator remains conditionally unbiased by Lemma 1, but using such poor predictions can increase variance. In this case, it is better to discard the predictions by setting $\lambda_i^t = 0$. In Section 3, we formalize this intuition by showing that the variance of the estimator is controlled by the variance of the residual correction Z_i^t, which can be reduced by tuning λ_i^t.

The preceding construction gives a conditionally unbiased estimator $\hat{\theta}_i^t$ whenever model i is selected at step t. To aggregate the information collected across multiple selections of the same model, we average these one-step estimators. Formally, let $A_i(t)$ be the set of steps in which the i-th model was selected. For example, if by time $t = 10$ model $i = 5$ was selected at steps $s = 1$ and $s = 7$, then $A_5(10) = \{1, 7\}$. We define the running estimator of μ_i at step t as follows:

$$\hat{\mu}_i^t := \frac{1}{|A_i(t)|} \sum_{s \in A_i(t)} \hat{\theta}_i^s. \tag{3}$$

We next show how the estimator $\hat{\mu}_i^t$ can be used to construct a valid confidence interval for μ_i. A key complication is that, even for a fixed model i, the sampled scores are not i.i.d. This is because

[1] For $\lambda_i^t = 1$ and $S_{i,j} = \hat{S}_{i,j}^t$ we have that $Z_i^t = 0$ and $\hat{\theta}_i^t = \frac{1}{n}\left[\sum_{j \in O_i^{t-1}} S_{i,j} + \sum_{j \in U_i^t} S_{i,j}\right] = \frac{1}{n}\sum_j S_{i,j} = \mu_i$.

examples are sampled without replacement: once model i is evaluated on example j, this index is removed from the future sets U_i^t. Moreover, the predicted scores $\hat{S}_{ij}^t$ may change over time because the low-rank factorization is updated during the sequential procedure. As a result, the one-step estimators $\{\hat{\theta}_i^s : s \in A_i(t)\}$ are neither independent nor identically distributed. We therefore use martingale theory to derive a concentration bound on the estimation error.

Theorem 1. *Fix model i and step t, and assume that $|A_i(t)| = K > 0$. Define $V_i = \frac{1}{|A_i(t)|}\sum_{s\in A_i(t)} Var(Z_i^s|\mathcal{F}_{s-1})$. Assume that there exist a constant $v > 0$ such that $V_i \le v$. Then, under the assumptions of Lemma 1, for every $\epsilon > 0$,*

$$P\big(|\hat{\mu}_i^t - \mu_i| \ge \epsilon,\ |A_i(t)| = K\big) \ \le\ 2\exp\left(-\frac{\epsilon^2}{2\big(\frac{v}{Kn^2} + \frac{2\epsilon}{3K}\big)}\right).$$

The above theorem clarifies how our CIs are *prediction-powered*. Specifically, it bounds the miscoverage rate of the interval $[\hat{\mu}_i^t - \epsilon, \hat{\mu}_i^t + \epsilon]$ and establishes that it decreases with the conditional variance of the residual. Importantly, constructing this confidence interval does not require knowing the upper bound v. As we show next, the miscovergae rate depends on both (i) how informative the predicted scores are about the true scores, and (ii) the choice of the weighting parameter λ_i^t.

Step 3: Tuning λ_i^t

We now study the effect of λ_i^t on the estimator's variance and derive a principled tuning approach. We begin by comparing our estimator to the observation-only estimator:

$$\hat{\mu}^{\text{obs}} = \frac{1}{|A_i(t)|}\sum_{s\in A_i(t)} S_{i,j_s}. \tag{4}$$

Finite-sample confidence intervals for the observation-only estimator can be constructed using the Bernstein–Serfling inequality [27]. The miscoverage rate of the CI for $\hat{\mu}_i^t$ is driven by the conditional variance appearing in Theorem 1, while the miscoverage rate of the CI for $\hat{\mu}^{\text{obs}}$ is driven by the finite-population (unconditional) variance $\mathrm{Var}(S_i)$; we therefore compare these two variances directly. The following lemma shows that an appropriate choice of λ_i^t reduces the conditional variance relative to the observation-only baseline.

Lemma 2. *Fix a model i. Suppose that step t is the K-th selection of model i, so that $|A_i(t)| = K$. The conditional variance of the residual correction Z_i^t is minimized by choosing*

$$(\lambda_i^t)^* = \frac{\mathrm{Cov}\Big(S_{i,j_t},\ \hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)}{\mathrm{Var}\Big(\hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)}. \tag{5}$$

At this choice of $(\lambda_i^t)^*$,

$$Var(\hat{\mu}_i^t|\mathcal{F}_{t-1}) \le (1-\rho_t^2)Var(\hat{\mu}^{obs}),$$

where $\rho_t = \mathrm{Corr}\Big(S_{i,j_t},\ \hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)$ is the conditional correlation between the true and predicted scores over the currently unobserved examples.

Since both Theorem 1 (applied to $\hat{\mu}_i^t$) and the Bernstein–Serfling inequality (applied to $\hat{\mu}^{\text{obs}}$) yield miscoverage rates driven by the variances bounded above, Lemma 2 implies that PULSE's CI miscoverage rate is not greater than the observation-only CI.[2] Moreover, the variance reduction improves as the conditional correlation ρ_t between the true and predicted scores increases. Hence, more informative predictions yield confidence intervals with larger coverage rate through a smaller residual variance.

In practice, computing $(\lambda_i^t)^*$ is not possible before observing the new score S_{i,j_t}. However, to preserve the conditional unbiasedness guarantee, λ_i^t must be chosen using only data available in $\mathcal{F}_{t-1}$, which does not include S_{i,j_t}. We therefore use a predictable plug-in approximation. The idea is to approximate the conditional variance

$$\mathrm{Var}(Z_i^t \mid \mathcal{F}_{t-1}) = \mathbb{E}\Big[\big(Z_i^t - \mathbb{E}[Z_i^t \mid \mathcal{F}_{t-1}]\big)^2 \,\Big|\, \mathcal{F}_{t-1}\Big]$$

[2] Up to constants of the respective concentration inequalities.

using quantities that are measurable with respect to $\mathcal{F}_{t-1}$. Specifically, we estimate the conditional mean $\mathbb{E}[Z_i^t \mid \mathcal{F}_{t-1}]$ by the average of the previously observed residual corrections,

$$\bar{Z}_i^{t-1} = \frac{1}{|A_i(t-1)|} \sum_{s \in A_i(t-1)} Z_i^s,$$

and approximate the unknown future score S_{i,j_t} by its current prediction $\hat{S}_{i,j_t}^t$. This yields the following predictable choice of λ_i^t:

$$\hat{\lambda}_i^t = \mathrm{clip}\left(1 - \frac{\left(\sum_{j \in U_i^t} \hat{S}_{ij}^t\right) \bar{Z}_i^{t-1}}{|U_i^t| \sum_{j \in U_i^t} (\hat{S}_{ij}^t)^2}, 0, 1\right). \tag{6}$$

Putting it all together

We now assemble the components above into the full PULSE procedure, summarized in Algorithm 1. At each round, the algorithm selects a model for evaluation, optionally refits the low-rank factorization model from Step 1, updates $\hat{\lambda}_i^t$ using Step 3, samples a new example for the selected model, and updates the running estimator $\hat{\mu}_i^t$ and its confidence interval using Step 2.

The following theorem gives a finite-sample success guarantee for the algorithm, despite adaptive sampling, sampling without replacement, and time-varying predictions:

Theorem 2. *Let $i^* = \arg\max_{i \in [m]} \mu_i$ be the unique best model, and define $\Delta_i = \mu_{i^*} - \mu_i$ for $i \neq i^*$ and $\Delta_{i^*} = \min_{i \neq i^*} \Delta_i$. Let $H_1 = \sum_i \Delta_i^{-2}$ and $\hat{i}_T$ be the output of Algorithm 1 when executed with a budget T. Suppose that the exploration parameter a satisfies $0 \leq a \leq \frac{25(T-m)}{36H_1}$, and assume that a constant $v_{i,K}$ in Theorem 1 can be chosen for every model $i \in [m]$ and every pull counts $K \leq T$. Then*

$$\mathbb{P}(\hat{i}_T = i^*) \geq 1 - \sum_{i=1}^{m} \sum_{K=1}^{T} 2\exp\left(-\frac{\epsilon_K^2}{2\left(\frac{v_{i,K}}{Kn^2} + \frac{2\epsilon_K}{3K}\right)}\right), \quad \epsilon_K = \frac{1}{5}\sqrt{\frac{a}{K}}.$$

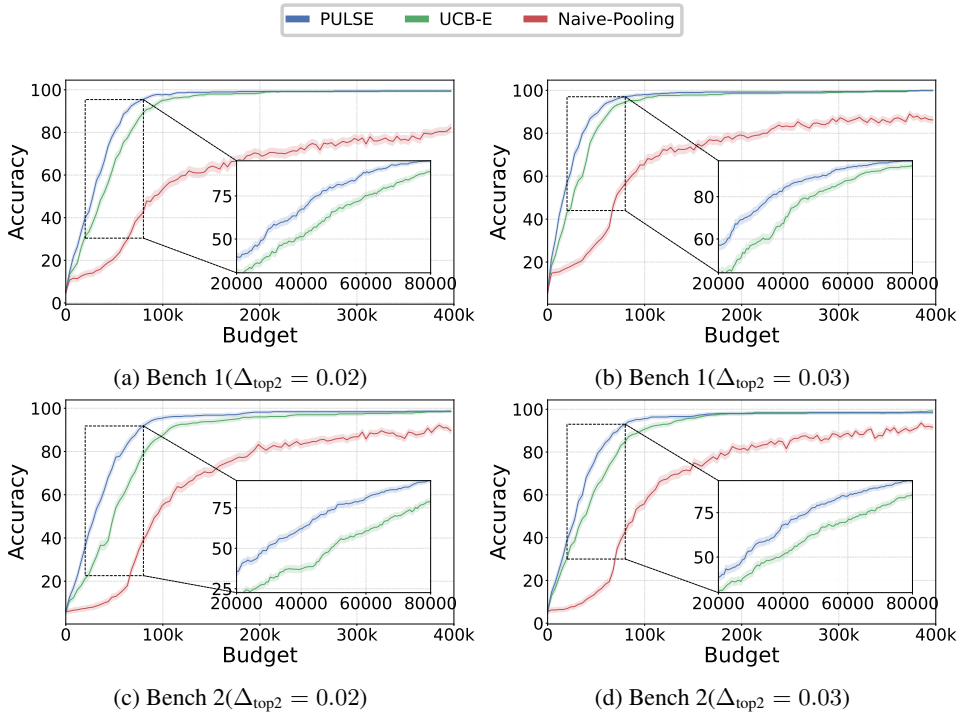
The theorem shows that the success probability is controlled by two main quantities. The first is the exploration parameter a, which determines the width of the confidence bounds used by Algorithm 1. Larger values of a lead to larger confidence bounds, encourage more exploration, and can improve convergence. However, taking a too large could violate the theorem condition, while taking a too small could reduce the success probability.

The second quantity affecting the success probability is the variance term v appearing in the confidence bound. Hence, informative predictions, together with an appropriate choice of λ_i^t, can reduce the number of LLM evaluations needed to reach a target success by decreasing v. Conversely, when the predictions are uninformative, our tuning rule is designed to downweight the predicted scores by choosing $\hat{\lambda}_i^t$ close to zero.

From single-example to batch evaluation. Since GPUs are most efficient when computations are batched, it is natural to evaluate multiple examples for the selected model at each step rather than a single example. This motivates a batch-mode extension of PULSE. The extension only requires modifying the residual correction term. Specifically, suppose that at step t we sample a batch $B_t \subset U_i^t$ of new example indices for model i, uniformly without replacement. If each index $j \in U_i^t$ has inclusion probability π_i^t, then we replace the single-example correction Z_i^t by $\tilde{Z}_i^t = \sum_{j \in B_t} (S_{ij} - \lambda_i^t \hat{S}_{ij}^t)/\pi_i^t$. For uniform batch sampling with $|B_t| = B$, we have $\pi_i^t = B/|U_i^t|$. All other components, $\hat{\theta}_i^t$, $\hat{\mu}_i^t$, and the tuning rule for λ_i^t, are unchanged.

4 Experiments

We evaluate our algorithm on real-world benchmarks with the goal of assessing two key aspects: (i) the benefit of incorporating predicted scores, and (ii) the effect of correcting their bias when estimating model accuracies and constructing valid CIs.



(a) Bench 1($\Delta_{\text{top2}} = 0.02$)

(b) Bench 1($\Delta_{\text{top2}} = 0.03$)

(c) Bench 2($\Delta_{\text{top2}} = 0.02$)

(d) Bench 2($\Delta_{\text{top2}} = 0.03$)

Figure 2: Accuracy versus budget. Shaded areas around each curve depict standard error.

Baselines. We compare PULSE against two baselines that share the UCB-E backbone but differ in how they estimate each model's mean accuracy: (i) **UCB-E**, the standard UCB-E algorithm, which ignores the low-rank predictions and uses the observation-only estimator (4); and (ii) **Naive-Pooling**, a UCB-E variant that does not account for prediction bias and treats predicted scores as ground truth: $\hat{\mu}_{i,t}^{\text{pred}} = \frac{1}{n}\left(\sum_{j \in O_i^{t-1}} S_{i,j} + \sum_{j \in U_i^t} \hat{S}_{i,j}^t\right)$.

The first baseline isolates the value of using predictions: any gap between PULSE and UCB-E reflects variance reduction attributed to informative predictions. The second isolates the value of preserving statistical validity: any gap between PULSE and Naive-Pooling reflects the benefit of bias correction.

All methods use batch size 64 and total budget $T = 540{,}800$ LLM calls ($T_0 = 1$ warmup batch per model plus $400{,}000$ for the bandit loop). PULSE and Naive-Pooling refit the low-rank factorization every $1{,}000$ steps. We run 500 independent repetitions and report, at each step, the fraction in which the method identifies the true best model. Full parameters in Appendix C.

Performance metrics. We consider the following metrics: (i) Accuracy@t: the probability of correctly identifying the best model at step t, $1 \le t \le T$ (ii) Budget needed to obtain 95% accuracy: the budget that each method requires to reach 95% level of accuracy.

Datasets We use the evaluation dataset released by [26], which provides binary correctness scores for $4.4K$ models across six benchmarks with $21.5K$ total number of questions. Following [26], we organize the data into two benchmark datasets: Bench 1 which contains *MMLU-Pro* [4], with 12K questions; and Bench 2 - a composite dataset of BBH [28], GPQA [29], IFEval [30], MATH [31], and MuSR [32], with 9.5K questions. For each dataset, we split the models chronologically according to the models release date: the older half is used as training data to warm-start the low-rank factorization model (as described next), and the remaining half forms the test pool over which best-model identification is performed. To assess our method under varying levels of difficulty, we consider the gap in accuracy between the top-2 models, denoted by Δ_{top2}, and examine different levels of $\Delta_{\text{top2}} \in \{0.01, 0.02, 0.03\}$. For each target Δ_{top2}, we construct a subset of 1000 models as follows. We first include the top-performing model from the full score matrix. We then select as

the second-best model the one whose accuracy gap with respect to the top model is closest to Δ_{top2}. The remaining 998 models are chosen as the next highest-performing models in descending order of mean accuracy. Smaller values of Δ_{top2} correspond to more challenging instances of the problem.

Low-rank model We use the train split of each dataset to fit an initial low-rank factorization $(U_{\text{train}}, V_{\text{train}})$. At test time, we initialize the question representations by setting $V \leftarrow V_{\text{train}}$ and keep V fixed throughout the run, as described in Stage 1 of Algorithm 1. The model representations U, in contrast, are reinitialized and updated online using the observed test-time scores via the refitting step in the algorithm. Specifically, after each batch of evaluations, we refit U while keeping V fixed. This design leverages the fact that question representations are shared between the train and test splits and can therefore be estimated from historical data. In contrast, the representations of the test models must be inferred from test-time observations, since these models are evaluated only during the sequential procedure.

Hyperparameters of the factorization—the regularization strength λ, rank r, learning rate, and number of training steps—are selected with the train split via 5-fold cross-validation procedure. Concretely, we use the 1-SE rule [33] that favors more parsimonious configurations among those within one standard error of the best cross-validated loss.

4.1 Results

Figure 1 shows the minimum budget required for 95% identification accuracy on both datasets with $\Delta_{\text{top2}} \in \{0.02, 0.03\}$: as can be seen, PULSE requires up to 46% fewer evaluations than UCB-E. Figure 2 reports identification accuracy across evaluation budgets, with full results in sections B.2 and B.3. As shown, PULSE achieves higher accuracy than UCB-E across all configurations, demonstrating the benefit of informative predictions. Importantly, PULSE's computational overhead, primarily due to low-rank factorization, is negligible: a run on MMLU-Pro with $T = 540$K takes approximately 60 seconds. Comparing PULSE to Naive-Pooling highlights the value of retaining statistical validity: indeed, Naive-Pooling performs substantially worse than both PULSE and UCB-E across all configurations, demonstrating that incorporating predictions without correcting their bias can be more harmful than relying solely on observed evaluations. We assess the significance of PULSE's accuracy gains over both baselines using one-sided McNemar tests with Holm–Bonferroni correction; the p-values appear in Appendix B.7.

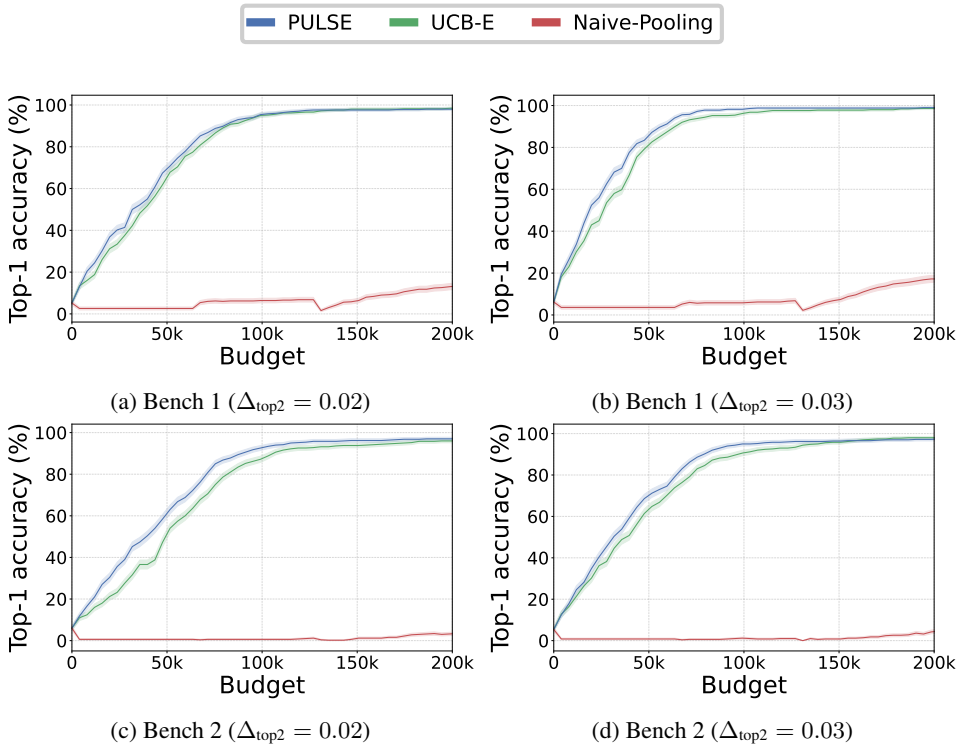
Given PULSE's strong performance, a natural question is whether similar results hold in a cold-start setting, where no historical data are available. We examine this question through an ablation study in which we randomly initialize U and V and fit them using only the data observed during the run. As shown in Figure 3, PULSE maintains an accuracy advantage over UCB-E across the full budget range. See Appendix B.1 for full details.

Next, we assess PULSE along three dimensions: its miscoverage rate, the correlation between predicted and observed scores, and the performance of the plug-in weight $\hat{\lambda}_{t,i}$ in (6) relative to that of the theoretically optimal weight in (5). Appendix B.4 compares the miscoverage rates of all methods. As can be seen, PULSE's miscoverage rate is no higher than that of UCB-E and is substantially lower than that of Naive-Pooling at all reported pull counts. Turning to predictive accuracy, we assess the correlation between predicted and observed scores and find a mean Pearson correlation coefficient of $\rho \approx 0.7$ across all configurations (Appendix B.6). Lastly, Appendix B.5 shows that the adaptive plug-in rule reaches 95% identification accuracy with fewer evaluations than any fixed value of λ tested, while achieving performance close to that of the optimal-weight benchmark.

Finally, we examine whether PULSE's gains persist across different choices of the exploration parameter a. As shown in Appendix B.9, PULSE requires fewer evaluations than UCB-E to reach 95% identification accuracy at every tested value of a.

5 Conclusions and Future Work

We studied the problem of identifying the best-performing LLM under a limited evaluation budget and proposed PULSE, a sequential framework that combines adaptive sampling with low-rank score prediction while preserving statistical validity. By leveraging AIPW, our method constructs unbiased



(a) Bench 1 ($\Delta_{\text{top2}} = 0.02$)

(b) Bench 1 ($\Delta_{\text{top2}} = 0.03$)

(c) Bench 2 ($\Delta_{\text{top2}} = 0.02$)

(d) Bench 2 ($\Delta_{\text{top2}} = 0.03$)

Figure 3: Accuracy versus budget without using historical data

estimators and valid CIs in a non-i.i.d. setting. Empirically we demonstrated that PULSE significantly reduces the number of required LLM evaluations while maintaining high identification accuracy.

Our approach has several limitations. First, its effectiveness depends on the quality of the low-rank factorization; improving the prediction model - for example, by incorporating richer contextual information—could further enhance performance. Second, our analysis focuses on binary scoring functions, and extending the framework to more general (non-binary) scores is an important direction, which we expect can be achieved with relatively minor modifications. Finally, PULSE currently samples examples uniformly at each step. Extending the method to non-uniform, potentially adaptive sampling is a promising direction. Intuitively, for a given model, some examples may be more informative than others for estimating its mean accuracy. Recent work [26] explores this idea in the single-model setting via adaptive question selection, and integrating such strategies into our framework could further improve efficiency by jointly optimizing over models and examples.

6 Acknowledgments

This research was supported the European Union (ERC, SafetyBounds, 101163414). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The authors also thank Matteo Sesia for his valuable feedback and insightful discussions.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35 (2):1–72, 2026.

[3] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[4] 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.

[5] 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. *NeurIPS*, 2021.

[6] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 9025–9049, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/file/25ae35b5b1738d80f1f03a8713e405ec-Paper-Conference.pdf`.

[7] Franck Ndzomga. Efficient benchmarking of ai agents. *arXiv preprint arXiv:2603.23749*, 2026.

[8] Sayash Kapoor et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. *arXiv preprint arXiv:2510.11977*, 2025.

[9] Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

[10] Jin Peng Zhou, Christian K Belardi, Ruihan Wu, Travis Zhang, Carla P Gomes, Wen Sun, and Kilian Q Weinberger. On speeding up language model evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[11] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[12] Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

[13] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[14] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

[15] Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

[16] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

[17] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

[18] Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021.

[19] Masahiro Kato, Kenichiro McAlinn, and Shota Yasui. The adaptive doubly robust estimator and a paradox concerning logging policy. *Advances in neural information processing systems*, 34:1351–1364, 2021.

[20] Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van Der Laan. Post-contextual-bandit inference. *Advances in neural information processing systems*, 34:28548–28559, 2021.

[21] Ruohan Zhan, Vitor Hadad, David A Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135, 2021.

[22] Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Wonyoung Kim, Gi-Soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling with linear payoffs. *Advances in neural information processing systems*, 34:15830–15840, 2021.

[24] Wenlong Ji, Yihan Pan, Ruihao Zhu, and Lihua Lei. Multi-armed bandits with machine learning-generated surrogate rewards. *arXiv preprint arXiv:2506.16658*, 2025.

[25] Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with llm judges and limited human audits. *Available at SSRN 6147806*, 2026.

[26] Skyler Wu, Yash Nair, and Emmanuel J Candés. Efficient evaluation of llm performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

[27] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 2015.

[28] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, and Quoc V Le. H chi, denny zhou, et al. 2022. challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[29] 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. URL `https://openreview.net/forum?id=Ti67584b98`.

[30] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[31] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

[32] Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 14670–14728, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/file/3f8c7eb848ffec848f3ed2b7ca44915d-Paper-Conference.pdf`.

[33] Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22, 2010.

[34] Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

[35] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

A Proofs

A.1 Proof of Lemma 1

Lemma 3. *Assume π_i^k, λ_i^k and $\hat{S}_{ij}^k$ are $\mathcal{F}_{k-1}$ measurable, then for each $k \geq 1$, $\mathbb{E}[\hat{\theta}_i^k \mid \mathcal{F}_{k-1}] = \mu_i$*

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_i^k \mid \mathcal{F}_{k-1}] =& \mathbb{E}\left[\frac{1}{n}\left[\sum_{j\in O_i^{k-1}} S_{i,j} + \sum_{j\in U_i^k} \lambda_i^k \hat{S}_{i,j}^k + \left(\frac{S_{i,j_k} - \lambda_i^k \hat{S}_{i,j_k}^k}{\pi_i^k}\right)\right] \,\middle|\, \mathcal{F}_{k-1}\right] \\
=& \frac{1}{n}\sum_{j\in O_i^{k-1}} S_{i,j} + \frac{1}{n}\sum_{j\in U_i^k} \lambda_i^k \hat{S}_{i,j}^k + \frac{1}{n}\mathbb{E}\left[\frac{S_{i,j_k} - \lambda_i^k \hat{S}_{i,j_k}^k}{\pi_i^k} \,\middle|\, \mathcal{F}_{k-1}\right] \\
=& \frac{1}{n}\sum_{j\in O_i^{k-1}} S_{i,j} + \frac{1}{n}\sum_{j\in U_i^k} \lambda_i^k \hat{S}_{i,j}^k + \frac{1}{n}\sum_{j\in U_i^k} \pi_i^k \left(\frac{S_{i,j} - \lambda_i^k \hat{S}_{i,j}^k}{\pi_i^k}\right) \\
=& \frac{1}{n}\sum_{j\in O_i^{k-1}} S_{i,j} + \frac{1}{n}\sum_{j\in U_i^k} S_{i,j} = \frac{1}{n}\sum_{j=1}^{n} S_{ij} = \mu_i
\end{aligned}$$

□

A.2 Proof of Theorem 1

Theorem 3. *Fix model i and step t, and assume that $|A_i(t)| = K > 0$. Define $V_i = \frac{1}{|A_i(t)|}\sum_{s\in A_i(t)} Var(Z_i^s|\mathcal{F}_{s-1})$. Assume that there exist a constant $v > 0$ such that $V_i \leq v$. Then, under the assumptions of Lemma 1, for every $\epsilon > 0$,*

$$P\big(|\hat{\mu}_i^t - \mu_i| \geq \epsilon,\ |A_i(t)| = K\big) \ \leq\ 2\exp\left(-\frac{\epsilon^2}{2\big(\frac{v}{Kn^2} + \frac{2\epsilon}{3K}\big)}\right).$$

Proof. For $s \in A_i(t)$, Define $D_i^s = \hat{\theta}_i^s - \mathbb{E}[\hat{\theta}_i^s|\mathcal{F}_{s-1}]$. Recall that:

$$\hat{\theta}_i^s = \frac{1}{n}\sum_{j\in O_i^{s-1}} S_{i,j} + \frac{1}{n}\sum_{j\in U_i^s} \lambda_i^s \hat{S}_{i,j}^s + \frac{1}{n}\left(\frac{S_{i,j_s} - \lambda_i^s \hat{S}_{i,j_s}^s}{\pi_i^s}\right).$$

Given $\mathcal{F}_{s-1}$, the only random part of $\hat{\theta}_i^s$ is $Z_i^s = \frac{S_{i,j_s} - \lambda_i^s \hat{S}_{i,j_s}^s}{\pi_i^s}$. Thus:

$$\mathbb{E}[\hat{\theta}_i^s|\mathcal{F}_{s-1}] = \frac{1}{n}\sum_{j\in O_i^{s-1}} S_{i,j} + \frac{1}{n}\sum_{j\in U_i^s} \lambda_i^s \hat{S}_{i,j}^s + \frac{1}{n}\mathbb{E}[Z_i^s|\mathcal{F}_{s-1}].$$

Therefore:

$$\hat{\theta}_i^s - \mathbb{E}[\hat{\theta}_i^s|\mathcal{F}_{s-1}] = \frac{1}{n}(Z_i^s - \mathbb{E}[Z_i^s|\mathcal{F}_{s-1}]).$$

Next, since $\lambda_i^s, \hat{S}_{i,j}^s, S_{i,j} \in [0,1]$ and since $\pi_i^s = 1/|U_i^s|$ we have that $-|U_i^s| \leq Z_i^s \leq |U_i^s|$. Thus $-2|U_i^s| \leq Z_i^s - \mathbb{E}[Z_i^s|\mathcal{F}_{s-1}] \leq 2|U_i^s|$. Since $|U_i^s| \leq n$, we have that:

$$|\hat{\theta}_i^s - \mathbb{E}[\hat{\theta}_i^s|\mathcal{F}_{s-1}]| \leq 2. \tag{7}$$

Additionally, the following holds:

$$\sum_{s\in A_i(t)} \mathbb{E}[(D_i^s)^2 \mid \mathcal{F}_{s-1}] = \sum_{s\in A_i(t)} \mathbb{E}[(\hat{\theta}_i^s - \mathbb{E}[\hat{\theta}_i^s|\mathcal{F}_{s-1}])^2 \mid \mathcal{F}_{s-1}] = \sum_{s\in A_i(t)} \text{Var}(\hat{\theta}_i^s \mid \mathcal{F}_{s-1}).$$

As before, given $\mathcal{F}_{s-1}$, the only random part of $\hat{\theta}_i^s$ is Z_i^s. Thus:

$$\sum_{s\in A_i(t)} \text{Var}(\hat{\theta}_i^s \mid \mathcal{F}_{s-1}) = \frac{1}{n^2} \sum_{s\in A_i(t)} \text{Var}(\hat{Z}_i^s \mid \mathcal{F}_{s-1}).$$

Combining this with the assumption that $V_i < v$ we get:

$$\sum_{s\in A_i(t)} \text{Var}(\hat{\theta}_i^s \mid \mathcal{F}_{s-1}) \leq \frac{vK}{n^2} \tag{8}$$

Finally, It holds that:

$$\mathbb{E}[D_i^s|\mathcal{F}_{s-1}] = 0. \tag{9}$$

Combining (7), (8), and (9) we get that $(D_i^s)_{s\in\mathbb{N}}$ is a martingale difference sequence (M.D.S) and $S_i^K = \sum_{s\in A_i(t)} D_i^s$ is a martingale with bounded differences.

Equivalently, define $\tilde{D}_i^s = D_i^s \cdot 1\{i_s = i\}$ for $s = 1, \ldots, t$; since $1\{i_s = i\}$ is $\mathcal{F}_{s-1}$-measurable by construction, $(\tilde{D}_i^s)_{s=1}^t$ is a martingale difference sequence w.r.t. $(\mathcal{F}_s)$ with $S_i^K = \sum_{s=1}^t \tilde{D}_i^s$, and the bound and conditional-variance estimates above carry over. With slight abuse of notation, we drop the tilde and write D_i^s for $\tilde{D}_i^s$ in what follows.

By Freedman's inequality [34, 35] applied to the martingale $(\tilde{D}_i^s)_{s\geq 1}$ at the deterministic time t, with $|\tilde{D}_i^s| \leq 2$ from (7) and predictable quadratic variation $\langle S^{(i)}\rangle_t = \sum_{s=1}^t \mathbb{E}[(\tilde{D}_i^s)^2 \mid \mathcal{F}_{s-1}]$, for every $\epsilon' > 0$,

$$P\big(|S_t^{(i)}| \geq \epsilon',\ \langle S^{(i)}\rangle_t \leq Kv/n^2\big) \ \leq\ 2\exp\left(-\frac{(\epsilon')^2}{2\big(Kv/n^2 + 2\epsilon'/3\big)}\right).$$

We now translate this into a bound on the event of interest. On $\{|A_i(t)| = K\}$, the padding indicators yield $S_t^{(i)} = \sum_{s\in A_i(t)} D_i^s = K(\hat{\mu}_i^t - \mu_i)$, and the assumption $V_i \leq v$ gives $\langle S^{(i)}\rangle_t = \sum_{s\in A_i(t)} \text{Var}(\hat{\theta}_i^s \mid \mathcal{F}_{s-1}) \leq Kv/n^2$ by (8). Hence

$$\{|\hat{\mu}_i^t - \mu_i| \geq \epsilon,\ |A_i(t)| = K\} \ \subseteq\ \big\{|S_t^{(i)}| \geq K\epsilon,\ \langle S^{(i)}\rangle_t \leq Kv/n^2\big\}.$$

Setting $\epsilon' = K\epsilon$ in Freedman's bound and combining with this inclusion yields

$$P\big(|\hat{\mu}_i^t - \mu_i| \geq \epsilon,\ |A_i(t)| = K\big) \ \leq\ 2\exp\left(-\frac{\epsilon^2}{2\big(v/(Kn^2) + 2\epsilon/(3K)\big)}\right),$$

which completes the proof.

□

A.3 Proof of Lemma 2

Lemma 4. *Fix a model i and a step t. Assume model i is selected at step t and $|A_i(t)| = K$. The conditional variance of the residual correction Z_i^t is minimized by choosing*

$$(\lambda_i^t)^* = \frac{\text{Cov}\Big(S_{i,j_t},\ \hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)}{\text{Var}\Big(\hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)}. \tag{10}$$

At this choice of $(\lambda_i^t)^*$,

$$\textit{Var}(\hat{\mu}_i^t|\mathcal{F}_{t-1}) \leq (1-\rho_t^2)\textit{Var}(\hat{\mu}^{\textit{obs}})$$

where $\rho_t = \text{Corr}\Big(S_{i,j_t},\ \hat{S}_{i,j_t}^t \,\Big|\, \mathcal{F}_{t-1}\Big)$. is the conditional correlation between the true and predicted scores over the currently unobserved examples.

Proof. Recall that:

$$Z_i^t = \frac{S_{i,j_t} - \lambda_i^t \hat{S}_{i,j_t}^t}{\pi_i^t}.$$

Thus:

$$\text{Var}(Z_i^t \mid \mathcal{F}_{t-1}) = \mathbb{E}[(Z_i^t - \mathbb{E}[Z_i^t|\mathcal{F}_{t-1}])^2 \mid \mathcal{F}_{t-1}].$$

Given $\mathcal{F}_{t-1}$, λ_i^t, π_i^t are constant, thus: $\mathbb{E}[Z_i^t|\mathcal{F}_{t-1}] = \frac{1}{\pi_i^t}(\mathbb{E}[S_{i,j_t} \mid \mathcal{F}_{t-1}] - \lambda_i^t \mathbb{E}[\hat{S}_{i,j_t}^t \mid \mathcal{F}_{t-1}])$.

Therefore:

$$(Z_i^t - \mathbb{E}[Z_i^t|\mathcal{F}_{t-1}])^2 = \left[\frac{1}{\pi_i^t}(S_{i,j_t} - \lambda_i^t \hat{S}_{i,j_t}^t) - \frac{1}{\pi_i^t}(\mathbb{E}[S_{i,j_t} \mid \mathcal{F}_{t-1}] - \lambda_i^t \mathbb{E}[\hat{S}_{i,j_t}^t \mid \mathcal{F}_{t-1}])\right]^2 = \frac{1}{(\pi_i^t)^2}((S_{i,j_t} - \mathbb{E}[S_{i,j_t} \mid \mathcal{F}_{t-1}]) - \lambda_i^t(\hat{S}_{i,j_t}^t - \mathbb{E}[\hat{S}_{i,j_t}^t \mid \mathcal{F}_{t-1}]))^2$$

Denote $\eta = \frac{1}{\pi_i^t}(S_{i,j_t} - \mathbb{E}[S_{i,j_t} \mid \mathcal{F}_{t-1}])$, and $\xi = \frac{1}{\pi_i^t}(\hat{S}_{i,j_t}^t - \mathbb{E}[\hat{S}_{i,j_t}^t \mid \mathcal{F}_{t-1}])$.

We have that: $(Z_i^t - \mathbb{E}[Z_i^t|\mathcal{F}_{t-1}])^2 = \eta^2 + (\lambda_i^t)^2\xi^2 - 2\lambda_i^t\eta\xi$.

Thus:

$$\text{Var}(Z_i^t \mid \mathcal{F}_{t-1}) = \mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}] + (\lambda_i^t)^2\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}] - 2\lambda_i^t\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}]. \tag{11}$$

This is a quadratic equation in λ_i^t, which optimum is:

$$(\lambda_i^t)^* = \frac{\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}]}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]} \tag{12}$$

Substituting η and ξ we get

$$(\lambda_i^t)^* = \frac{\text{Cov}\left(\frac{S_{i,j_t}}{\pi_i^t}, \frac{\hat{S}_{i,j_t}^t}{\pi_i^t} \mid \mathcal{F}_{t-1}\right)}{\text{Var}\left(\frac{\hat{S}_{i,j_t}^t}{\pi_i^t} \mid \mathcal{F}_{t-1}\right)}.$$

Given $\mathcal{F}_{t-1}$, π_i^t is a constant, thus we can pull it out of the variance and covariance to get:

$$(\lambda_i^t)^* = \frac{\text{Cov}\left(S_{i,j_t}, \hat{S}_{i,j_t}^t \,\middle|\, \mathcal{F}_{t-1}\right)}{\text{Var}\left(\hat{S}_{i,j_t}^t \,\middle|\, \mathcal{F}_{t-1}\right)}.$$

Next, plug in (12) in (11) to get:

$$\begin{aligned}
\text{Var}(Z_i^t \mid \mathcal{F}_{t-1})\big|_{(\lambda_i^t)^*} =& \mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}] + \left(\frac{\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}]}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]}\right)^2 \mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}] - 2\frac{\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}]}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]}\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}] = \\
& \mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}] + \frac{(\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}])^2}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]} - 2\frac{(\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}])^2}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]} = \\
& \mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}] - \frac{(\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}])^2}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]} = \mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}]\left(1 - \frac{(\mathbb{E}[\eta\xi \mid \mathcal{F}_{t-1}])^2}{\mathbb{E}[\xi^2 \mid \mathcal{F}_{t-1}]\mathbb{E}[\eta^2 \mid \mathcal{F}_{t-1}]}\right) = \\
& \text{Var}\left(\frac{S_{i,j_t}}{\pi_i^t} \mid \mathcal{F}_{t-1}\right)(1 - \rho_t^2) = |U_i^t|^2\text{Var}\left(S_{i,j_t} \mid \mathcal{F}_{t-1}\right)(1 - \rho_t^2).
\end{aligned}$$

Where the last transition is due to $1/\pi_i^t = |U_i^t|$.

Next, since

$$\hat{\theta}_i^t = \frac{1}{n} \sum_{j \in O_i^{t-1}} S_{i,j} + \frac{1}{n} \sum_{j \in U_i^t} \lambda_i^s \hat{S}_{i,j}^t + \frac{1}{n} Z_i^t.$$

We have that:

$$\text{Var}(\hat{\theta}_i^t \mid \mathcal{F}_{t-1}) = \frac{1}{n^2} \text{Var}(Z_i^t \mid \mathcal{F}_{t-1})$$

Thus:

$$\text{Var}(\hat{\theta}_i^t \mid \mathcal{F}_{t-1}) = \frac{|U_i^t|^2}{n^2} \text{Var}\left(S_{i,j_t} \mid \mathcal{F}_{t-1}\right)(1-\rho_t^2)$$

Finally, since $\hat{\mu}_i^t := \frac{1}{|A_i(t)|} \sum_{s \in A_i(t)} \hat{\theta}_i^s$ and $|A_i(t)| = K$ we have that:

$$\begin{aligned}
\text{Var}(\hat{\mu}_i^t | \mathcal{F}_{t-1}) &= \text{Var}\left(\frac{1}{|A_i(t)|} \sum_{s \in A_i(t)} \hat{\theta}_i^s \mid \mathcal{F}_{t-1} \right) && (13) \\
&= \frac{1}{K^2} \text{Var}\left(\hat{\theta}_i^t \mid \mathcal{F}_{t-1} \right) && (14) \\
&= \frac{1}{K^2} \frac{|U_i^t|^2}{n^2} \text{Var}\left(S_{i,j_t} \mid \mathcal{F}_{t-1}\right)(1-\rho_t^2) && (15) \\
&\leq \frac{1}{K^2} \frac{|U_i^t|^2}{n^2} \frac{n}{|U_i^t|} \cdot \text{Var}(S_i)(1-\rho_t^2). && (16)
\end{aligned}$$

Where the last inequality stems from Lemma 5.

The variance of the fully observed estimator is:

$$\text{Var}(\hat{\mu}^{\text{obs}}) = \frac{\text{Var}(S_i)}{K} \cdot \frac{n-K}{n-1},$$

where $\text{Var}(S_i) = \frac{1}{n} \sum_j (S_{i,j} - \mu_i)^2$ is the finite-population variance [27].

Thus:

$$\text{Var}(S_i) = \text{Var}(\hat{\mu}^{\text{obs}}) K \frac{n-1}{n-K}. \tag{17}$$

Combining (16) and (17) yields:

$$\text{Var}(\hat{\mu}_i^t | \mathcal{F}_{t-1}) \leq \frac{1}{K^2} \frac{|U_i^t|^2}{n^2} \frac{n}{|U_i^t|} \cdot \text{Var}(\hat{\mu}^{\text{obs}}) K \frac{n-1}{n-K} (1-\rho_t^2)$$

Substitute $|U_i^t| = n - K + 1$ we have that:

$$\text{Var}(\hat{\mu}_i^t | \mathcal{F}_{t-1}) \leq \frac{|U_i^t|}{|U_i^t| - 1} \frac{n-1}{nK} \text{Var}(\hat{\mu}^{\text{obs}})(1-\rho_t^2) \tag{18}$$

For $K = 1$ we have that $|U_i^t| = n$ and:

$$\begin{aligned}
\text{Var}(\hat{\mu}_i^t | \mathcal{F}_{t-1}) &\leq \frac{|U_i^t|}{|U_i^t| - 1} \frac{n-1}{nK} \text{Var}(\hat{\mu}^{\text{obs}})(1-\rho_t^2) = && (19) \\
\frac{n}{n-1} \frac{n-1}{n} \text{Var}(\hat{\mu}^{\text{obs}})(1-\rho_t^2) &= \text{Var}(\hat{\mu}^{\text{obs}})(1-\rho_t^2). && (20)
\end{aligned}$$

For $K \leq n-1$, we have that $|U_i^t| \geq 2$ which implies that $\frac{|U_i^t|}{|U_i^t|-1} \leq 2$. Thus:

$$\mathrm{Var}(\hat{\mu}_i^t|\mathcal{F}_{t-1}) \le \frac{2(n-1)}{nK}\,\mathrm{Var}(\hat{\mu}^{\mathrm{obs}})(1-\rho_t^2) \le \frac{2}{K}\,\mathrm{Var}(\hat{\mu}^{\mathrm{obs}})(1-\rho_t^2)$$

As such, for $K \ge 2$, $\mathrm{Var}(\hat{\mu}_i^t|\mathcal{F}_{t-1}) \le \mathrm{Var}(\hat{\mu}^{\mathrm{obs}})(1-\rho_t^2)$. Combining this with (20) yields that for any $K \in [1, n-1]$, $\mathrm{Var}(\hat{\mu}_i^t|\mathcal{F}_{t-1}) \le \mathrm{Var}(\hat{\mu}^{\mathrm{obs}})(1-\rho_t^2)$

□

Lemma 5. *For every $t \ge 1$ and $i \in [m]$, the conditional variance of S_{i,j_t} given $\mathcal{F}_{t-1}$ is bounded by*

$$\mathrm{Var}(S_{i,j_t} \mid \mathcal{F}_{t-1}) \;\le\; \frac{n}{|U_i^t|} \cdot \mathrm{Var}(S_i). \tag{21}$$

Proof. Let $\bar{S}_{U_i^t} := \frac{1}{|U_i^t|}\sum_{j\in U_i^t} S_{i,j}$ denote the empirical mean of the scores of model i over the currently unobserved examples. Since j_t is sampled uniformly from U_i^t, we have

$$\mathrm{Var}(S_{i,j_t} \mid \mathcal{F}_{t-1}) = \frac{1}{|U_i^t|}\sum_{j\in U_i^t}\big(S_{i,j} - \bar{S}_{U_i^t}\big)^2.$$

Using the identity

$$\sum_{j\in U_i^t}\big(S_{i,j} - \bar{S}_{U_i^t}\big)^2 = \sum_{j\in U_i^t}(S_{i,j}-\mu_i)^2 - |U_i^t|\big(\bar{S}_{U_i^t} - \mu_i\big)^2,$$

and dropping the non-negative second term, we obtain

$$\mathrm{Var}(S_{i,j_t} \mid \mathcal{F}_{t-1}) \le \frac{1}{|U_i^t|}\sum_{j\in U_i^t}(S_{i,j}-\mu_i)^2.$$

Since each summand is non-negative, extending the sum to all $j \in [n]$ only increases it:

$$\frac{1}{|U_i^t|}\sum_{j\in U_i^t}(S_{i,j}-\mu_i)^2 \le \frac{1}{|U_i^t|}\sum_{j=1}^{n}(S_{i,j}-\mu_i)^2 = \frac{n}{|U_i^t|}\cdot\mathrm{Var}(S_i),$$

which completes the proof. □

A.4 Proof of Theorem 2

Theorem 4. *Let $i^* = \arg\max_{i\in[m]}\mu_i$ be the unique best model, and define $\Delta_i = \mu_{i^*} - \mu_i$ for $i \ne i^*$ and $\Delta_{i^*} = \min_{i\ne i^*}\Delta_i$. Let $H_1 = \sum_i \Delta_i^{-2}$. Let $\hat{i}_T$ be the output of Algorithm 1 when executed with a budget T, and let a satisfy $0 \le a \le \frac{25(T-m)}{36H_1}$, and assume that a constant $v_{i,K}$ in Theorem 1 can be chosen for every model $i \in [m]$ and every pull counts $K \le T$. Then*

$$\mathbb{P}(\hat{i}_T = i^*) \ge 1 - \sum_{i=1}^{m}\sum_{K=1}^{T} 2\exp\left(-\frac{\epsilon_K^2}{2\left(\frac{v}{Kn^2} + \frac{2\epsilon_K}{3K}\right)}\right), \quad \epsilon_K = \frac{1}{5}\sqrt{\frac{a}{K}}.$$

Proof. For each i and $K \ge 1$, let $\tau_i^{(K)} := \inf\{t : |A_i(t)| = K\}$ denote the wall-clock time at which arm i is selected for the K-th time (with $\tau_i^{(K)} = \infty$ if this never occurs). Since $\hat{\mu}_i^t$ is piecewise constant in t between consecutive pulls of arm i, the running estimator at any $t \in [T]$ with $|A_i(t)| = K$ equals $\hat{\mu}_i^{\tau_i^{(K)}}$.

For every $1 \le i \le m, 1 \le K \le T$ Consider the events:

$$\xi_{i,K} \;=\; \left\{\tau_i^{(K)} > T\right\} \cup \left\{|\hat{\mu}_i^{\tau_i^{(K)}} - \mu_i| < \tfrac{1}{5}\sqrt{a/K}\right\}, \qquad \xi = \bigcap_{i=1}^{m}\bigcap_{K=1}^{T}\xi_{i,K}.$$

From [9] we know that for every $0 \le a \le \frac{25(T-m)}{36H_1}$ on ξ we have that $\hat{i}_T = i^*$. To conclude the proof we show that $P(\xi) \ge 1 - \sum_{K=1}^{T}\sum_{i=1}^{m} \alpha_{i,K}$. Fix i and K. From Theorem 1 we have that:

$$P\big(|\hat{\mu}_i^t - \mu_i| \ge \epsilon,\ |A_i(t)| = K\big) \ \le\ 2\exp\left(-\frac{\epsilon^2}{2\big(\frac{v_{i,K}}{Kn^2} + \frac{2\epsilon}{3K}\big)}\right).$$

To control $\xi_{i,K}^c$, apply the argument of Theorem 1 to the stopped martingale $S_t^{(i,K)} := S_{t\wedge\tau_i^{(K)}}^{(i)}$. Stopping preserves the martingale property, the increment bound $|\Delta S_t^{(i,K)}| \le 2$, and the predictable quadratic variation bound $\langle S^{(i,K)}\rangle_T \le K v_{i,K}/n^2$. On $\{\tau_i^{(K)} \le T\}$, $S_T^{(i,K)} = K(\hat{\mu}_i^{\tau_i^{(K)}} - \mu_i)$, so Freedman applied at the deterministic time T gives

$$P\big(\tau_i^{(K)} \le T,\ |\hat{\mu}_i^{\tau_i^{(K)}} - \mu_i| \ge \epsilon_K\big) \ \le\ \alpha_{i,K}.$$

Thus:

$$P\left(\xi_{i,K}\right) \ge 1 - \alpha_{i,K}$$

where $\alpha_{i,K} = 2\exp\left(-\frac{\epsilon^2}{2(\frac{v_{i,K}}{Kn^2} + \frac{2}{K}\epsilon/3)}\right)$

Taking the union bound over $\xi_{i,K}$ we get $P(\xi) \ge 1 - \sum_{K=1}^{T}\sum_{i=1}^{m} \alpha_{i,K}$.

□

A.5 Derivation of $\hat{\lambda}_i^k$

Our goal is to find the weighting parameter λ_i^k that minimizes the approximate optimization objective:

$$\mathbb{E}\left[\left(Z_i^k - \bar{Z}_i^{<k}\right)^2 \,\middle|\, \mathcal{F}_{k-1}\right], \tag{22}$$

where $\bar{Z}_i^{<k}$ is a predictable proxy for $\mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}]$ formed from past residuals. To simplify notation, let $n_k = |U_i^{k-1}|$ be the number of unobserved examples for model i at step k, and define the historical residual mean as the average over the steps at which model i was previously selected,

$$\bar{Z}_i^{<k} \ =\ \frac{1}{|A_i(k-1)|}\sum_{s\in A_i(k-1)} Z_i^s,$$

matching the convention used in Step 3 of the main text. Since $A_i(k-1)$ and the residuals $\{Z_i^s : s \in A_i(k-1)\}$ are $\mathcal{F}_{k-1}$-measurable, $\bar{Z}_i^{<k}$ is treated as a constant throughout the conditional expectation below.

By expanding the squared difference and adding/subtracting $\mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}]$, we can decompose the objective into variance and bias components:

$$\begin{aligned}\mathbb{E}\left[\left(Z_i^k - \bar{Z}_i^{<k}\right)^2 \,\middle|\, \mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\left(Z_i^k - \mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}] + \mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}] - \bar{Z}_i^{<k}\right)^2 \,\middle|\, \mathcal{F}_{k-1}\right] \\ &= \mathrm{Var}(Z_i^k \mid \mathcal{F}_{k-1}) \ +\ \left(\mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}] - \bar{Z}_i^{<k}\right)^2, \end{aligned} \tag{23}$$

where the cross-term vanishes because $\mathbb{E}[(Z_i^k - \mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}]) \mid \mathcal{F}_{k-1}] = 0$ and the remaining terms are $\mathcal{F}_{k-1}$ measurable.

Variance and Expectation Proxies. Because j_k is sampled uniformly from U_i^{k-1}, the inclusion probability is $\pi_i^k = 1/n_k$. The true conditional variance is:

$$\begin{aligned}\mathrm{Var}(Z_i^k \mid \mathcal{F}_{k-1}) &= n_k^2\,\mathrm{Var}(S_{i,j_k} - \lambda_i^k \hat{S}_{i,j_k}^k \mid \mathcal{F}_{k-1}) \\ &= n_k \sum_{j\in U_i^{k-1}} \left((S_{i,j} - \lambda_i^k \hat{S}_{i,j}^k) - \frac{1}{n_k}\sum_{j'\in U_i^{k-1}} (S_{i,j'} - \lambda_i^k \hat{S}_{i,j'}^k)\right)^2.\end{aligned}$$

Since the true scores $S_{i,j}$ for $j \in U_i^{k-1}$ are unknown, we substitute them with the predictable plug-in estimates $\hat{S}_{i,j}^k$. The term $(S_{i,j} - \lambda_i^k \hat{S}_{i,j}^k)$ is therefore approximated by $(1-\lambda_i^k)\hat{S}_{i,j}^k$. This yields the variance proxy:

$$\widetilde{\mathrm{Var}}(Z_i^k \mid \mathcal{F}_{k-1}) \;=\; n_k(1-\lambda_i^k)^2 \sum_{j \in U_i^{k-1}} (\hat{S}_{i,j}^k - \bar{S}_i^k)^2 \;=\; n_k(1-\lambda_i^k)^2 \mathrm{SS}_i^k, \tag{24}$$

where $\bar{S}_i^k = \frac{1}{n_k} \sum_{j \in U_i^{k-1}} \hat{S}_{i,j}^k$ and $\mathrm{SS}_i^k := \sum_{j \in U_i^{k-1}} \left(\hat{S}_{i,j}^k - \bar{S}_i^k\right)^2$

Similarly, we approximate the predictable expectation. The true conditional expectation is $\mathbb{E}[Z_i^k \mid \mathcal{F}_{k-1}] = \sum_{j \in U_i^{k-1}} (S_{i,j} - \lambda_i^k \hat{S}_{i,j}^k)$. Substituting $S_{i,j} \approx \hat{S}_{i,j}^k$, the predictable expectation proxy becomes:

$$\widetilde{\mathbb{E}}[Z_i^k \mid \mathcal{F}_{k-1}] \;=\; (1-\lambda_i^k) \sum_{j \in U_i^{k-1}} \hat{S}_{i,j}^k \;=\; (1-\lambda_i^k) F_i^k. \tag{25}$$

Closed form for $\hat{\lambda}_i^k$. Plugging these proxies back into Eq. (23), our approximate objective becomes:

$$\widetilde{\mathrm{Var}}(Z_i^k \mid \mathcal{F}_{k-1}) \;+\; \left(\widetilde{\mathbb{E}}[Z_i^k \mid \mathcal{F}_{k-1}] - \bar{Z}_i^{<k}\right)^2 = n_k(1-\lambda_i^k)^2 \mathrm{SS}_i^k \;+\; \left((1-\lambda_i^k)F_i^k - \bar{Z}_i^{<k}\right)^2$$
$$= (1-\lambda_i^k)^2 \left(n_k \mathrm{SS}_i^k + (F_i^k)^2\right) - 2(1-\lambda_i^k) F_i^k \bar{Z}_i^{<k} + (\bar{Z}_i^{<k})^2.$$

By applying the algebraic identity from Lemma 6, we know that $n_k \mathrm{SS}_i^k + (F_i^k)^2 = n_k \Phi_i^k$, where $\Phi_i^k = \sum_{j \in U_i^{k-1}} (\hat{S}_{i,j}^k)^2$. Thus, the objective simplifies to:

$$(1-\lambda_i^k)^2 n_k \Phi_i^k - 2(1-\lambda_i^k) F_i^k \bar{Z}_i^{<k} + (\bar{Z}_i^{<k})^2. \tag{26}$$

This is a simple quadratic in $(1-\lambda_i^k)$. Taking the derivative with respect to $(1-\lambda_i^k)$ and setting it to zero yields:

$$2(1-\lambda_i^k) n_k \Phi_i^k - 2F_i^k \bar{Z}_i^{<k} = 0 \quad \Longrightarrow \quad 1-\lambda_i^k = \frac{F_i^k \bar{Z}_i^{<k}}{n_k \Phi_i^k}. \tag{27}$$

Solving for λ_i^k and projecting the result back into the valid range $[0,1]$ yields the final closed-form update rule used in the algorithm:

$$\hat{\lambda}_i^k = \mathrm{clip}\left(1 - \frac{F_i^k \bar{Z}_i^{<k}}{n_k \Phi_i^k},\; 0,\; 1\right), \tag{28}$$

which matches the formula presented in Step 3.

Lemma 6. *Let U_i^{k-1} be a finite index set of size $n := |U_i^{k-1}|$, and let $\hat{S}_{i,j}^k$ be a real-valued sequence indexed by $j \in U_i^{k-1}$. Define*

$$F_i^k \;:=\; \sum_{j \in U_i^{k-1}} \hat{S}_{i,j}^k, \qquad \Phi_i^k \;:=\; \sum_{j \in U_i^{k-1}} (\hat{S}_{i,j}^k)^2, \qquad \bar{S}_i^k \;:=\; \frac{F_i^k}{n}.$$

Define the sum of squared deviations

$$\mathrm{SS}_i^k \;:=\; \sum_{j \in U_i^{k-1}} \left(\hat{S}_{i,j}^k - \bar{S}_i^k\right)^2.$$

Then

$$n \cdot \mathrm{SS}_i^k \;=\; n\,\Phi_i^k \;-\; (F_i^k)^2. \tag{29}$$

Proof. Expand the squared deviation:

$$\left(\hat{S}_{i,j}^k - \bar{S}_i^k\right)^2 \;=\; (\hat{S}_{i,j}^k)^2 \;-\; 2\hat{S}_{i,j}^k \bar{S}_i^k \;+\; (\bar{S}_i^k)^2.$$

Sum over $j \in U_i^{k-1}$:

$$\mathrm{SS}_i^k \;=\; \sum_{j \in U_i^{k-1}} (\hat{S}_{i,j}^k)^2 \;-\; 2\bar{S}_i^k \sum_{j \in U_i^{k-1}} \hat{S}_{i,j}^k \;+\; \sum_{j \in U_i^{k-1}} (\bar{S}_i^k)^2.$$

Identify each term:

- $\sum_{j \in U_i^{k-1}} (\hat{S}_{i,j}^k)^2 = \Phi_i^k$ by definition.
- $\sum_{j \in U_i^{k-1}} \hat{S}_{i,j}^k = F_i^k$, so the middle term is $-2\bar{S}_i^k \, F_i^k$.
- $\sum_{j \in U_i^{k-1}} (\bar{S}_i^k)^2 = n(\bar{S}_i^k)^2$ since $\bar{S}_i^k$ is constant in j and the sum has n terms.

Therefore

$$\mathrm{SS}_i^k \;=\; \Phi_i^k \;-\; 2\bar{S}_i^k \, F_i^k \;+\; n(\bar{S}_i^k)^2.$$

Substituting $\bar{S}_i^k = F_i^k / n$:

$$\mathrm{SS}_i^k \;=\; \Phi_i^k \;-\; 2 \cdot \frac{F_i^k}{n} \cdot F_i^k \;+\; n \cdot \frac{(F_i^k)^2}{n^2} \;=\; \Phi_i^k \;-\; \frac{2(F_i^k)^2}{n} \;+\; \frac{(F_i^k)^2}{n} \;=\; \Phi_i^k \;-\; \frac{(F_i^k)^2}{n}.$$

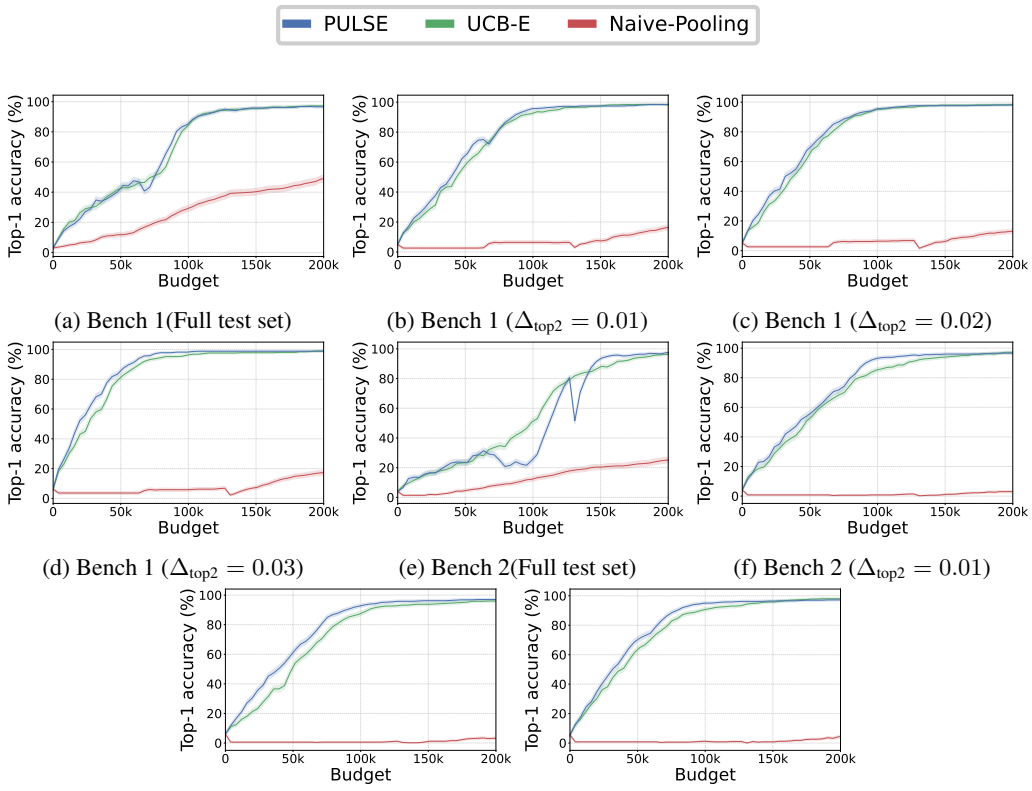
Multiplying both sides by n yields (29). □

B Additional Results

B.1 Evaluation without Historical Data

We repeat the experiment without historical side information, using the same configuration as the side-information experiment except for proxy training. For each seed, we initialize a fresh rank-100 matrix factorization and train both model and question factors using only the initial observations from the test models. Both factors remain trainable during subsequent updates, and we use a retraining learning rate of 0.01 instead of 0.1. We compare PULSE, UCB-E, and Naive-Pooling with a periodically retrained proxy, keeping the rest of the configuration unchanged.

Figure 4 depicts the results. We can see that for all Δ_{top2} datasets PULSE achieves higher accuracy than UCB-E. For the full dataset, PULSE reaches high accuracy faster than UCB-E, for low accuracy zones, PULSE may lag behind UCB-E. This is due to noise from training from scratch that may lead to uncorrelated predictions in low-budget regions.



(a) Bench 1(Full test set) (b) Bench 1 ($\Delta_{\text{top2}} = 0.01$) (c) Bench 1 ($\Delta_{\text{top2}} = 0.02$)

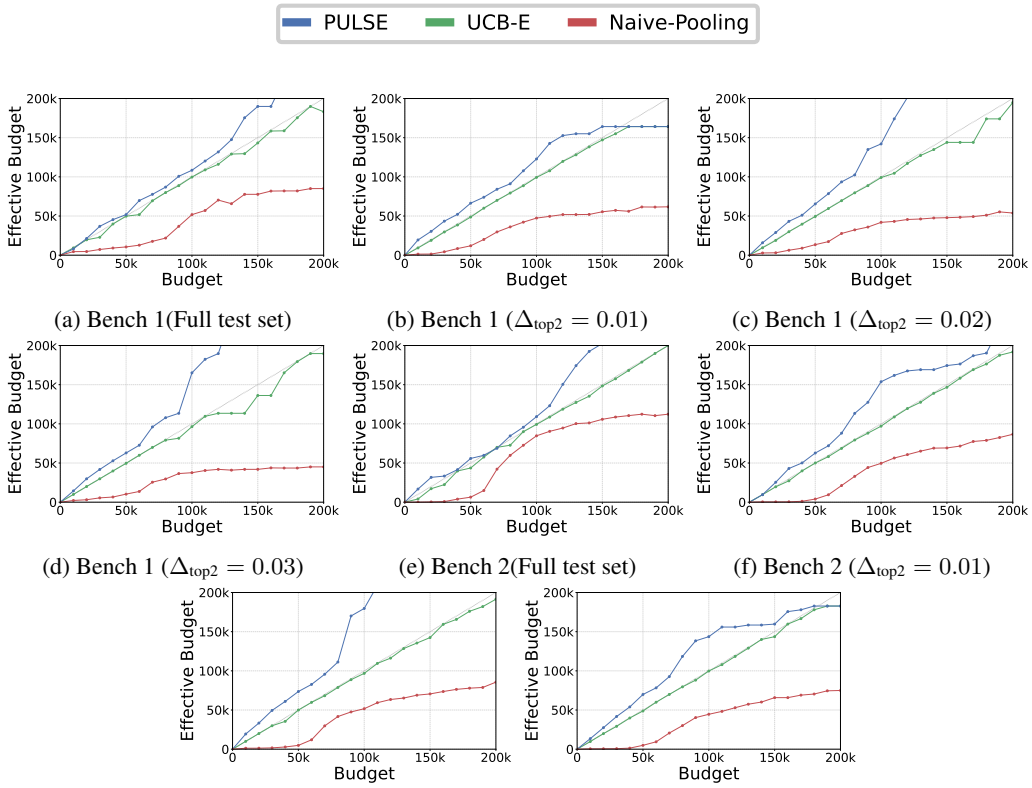
(d) Bench 1 ($\Delta_{\text{top2}} = 0.03$) (e) Bench 2(Full test set) (f) Bench 2 ($\Delta_{\text{top2}} = 0.01$)

(g) Bench 2 ($\Delta_{\text{top2}} = 0.02$) (h) Bench 2 ($\Delta_{\text{top2}} = 0.03$)

Figure 4: Accuracy versus budget without using historical data

B.2 Effective Budget

Figure 5 report the effective budget compared to UCB-E, defined as the number of LLM calls required by each method to achieve the level of accuracy achieved by UCB-E when called with Budget B. As can be seen, across all test settings, PULSE requires fewer evaluations than UCB-E to reach the same accuracy level.



(a) Bench 1(Full test set) (b) Bench 1 ($\Delta_{\text{top2}} = 0.01$) (c) Bench 1 ($\Delta_{\text{top2}} = 0.02$)

(d) Bench 1 ($\Delta_{\text{top2}} = 0.03$) (e) Bench 2(Full test set) (f) Bench 2 ($\Delta_{\text{top2}} = 0.01$)

(g) Bench 2 ($\Delta_{\text{top2}} = 0.02$) (h) Bench 2 ($\Delta_{\text{top2}} = 0.03$)

Figure 5: Effective Budget

B.3 Full Accuracy Results

Figure 6 (b) and (d) show the Accuracy@t curve of Bench 1 and 2, respectively (i.e., when evaluated on the entire datasets). Figure 6 (a) and (c) show the Accuracy@t curve for a subset of 1K models selected from Bench 1 and 2, respectively, so that $\Delta_{\text{top2}} = 0.01$.

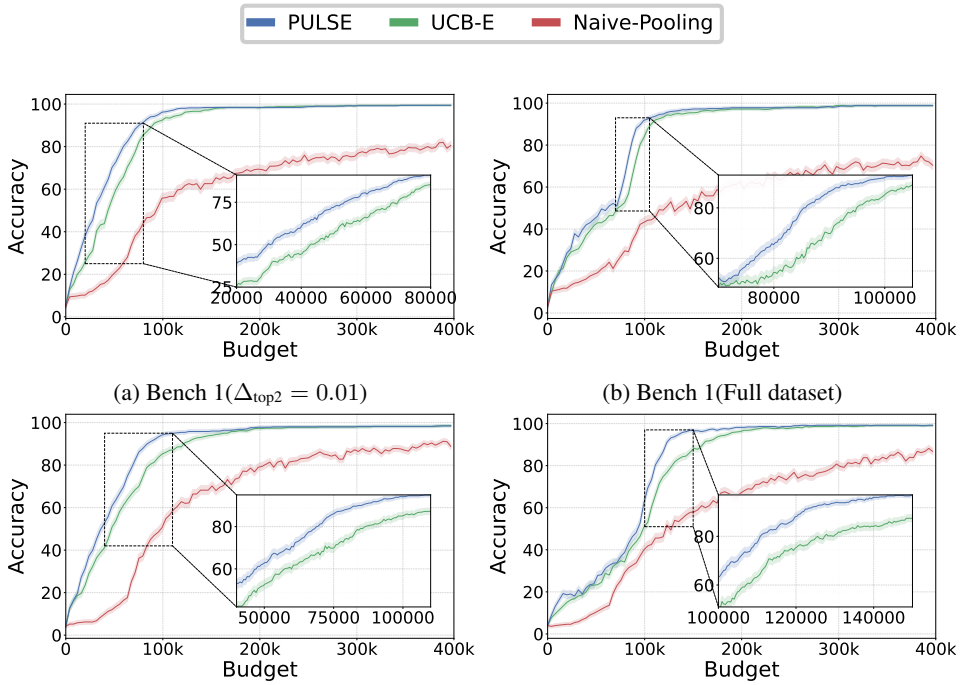
B.4 Empirical Miscoverage

There are two conclusions from Theorem 1 which are crucial to our method: (i) The interval is valid and its miscoverage is bounded, where the bound tightens as the pull-count K grows. (ii) The bound improves as the conditional variance of the estimator decreases.

We next validate these conclusions. Concretely, for each pull-count, we report the miscoverage rate (lower is better) over the 500 repetitions; Table 1 presents the results. As expected, for PULSE and UCB-E the rate broadly declines as K grows. On the other hand, Naive-Pooling instead miscovers at 11–17%, roughly an order of magnitude higher than PULSE, and its rate does not improve with K. This is to be expected due to the inherent bias in the Naive-Pooling estimator. Additionally, PULSE's miscoverage is never higher than UCB-E's, and is often lower. This effect is aligned with Theorem 1 that shows that miscoverage decreases with the variance, which according to Lemma 2 is minimized via the tuning of λ.

B.5 λ Ablations

we conducted an ablation study on all configurations (8 in total) comparing six choices: ($\lambda_t = 0$), $\lambda = 1/3$, $\lambda = 2/3$, ($\lambda_t = 1$), our clipped plug-in rule, and the offline oracle choice which computes the value of $(\lambda_i^t)^*$. Table 2 shows the number of configurations each method won (exlucding the



(a) Bench 1($\Delta_{\text{top2}} = 0.01$)

(b) Bench 1(Full dataset)

(c) Bench 2($\Delta_{\text{top2}} = 0.01$)

(d) Bench 2(Full dataset)

Figure 6: Accuracy versus budget. Shaded areas around each curve depict standard error.

Table 1: Comparison of UCB-E, PULSE, and Naive-Pooling for different values of K.

Pull Count	UCB-E	PULSE	Naive-Pooling
128	2.01%	0.80%	12.88%
256	2.21%	0.81%	13.71%
512	1.21%	1.21%	11.11%
1024	1.21%	0.20%	12.76%
2048	0.80%	0.20%	14.92%
4096	0.20%	0.20%	17.18%

oracle). As we can see, the adaptive plugin rule wins in most of the configurataions. This shows that while some fixed values may occasionaly outperform the plugin rule, an adaptive rule is required to be able to achieve good performence across a wide range of settings. Table 3 lists the required budget for each method to reach 95% accuracy.

Table 2: Number of settings achieving the lowest post-initialization budget to first reach 95% mean top-1 accuracy, with and without side information (SI).

Coefficient	With SI	Without SI
Adaptive PULSE	**5/8**	**4/8**
$\lambda = 1$	2/8	3/8
$\lambda = 2/3$	1/8	0/8
$\lambda = 1/3$	0/8	1/8
$\lambda = 0$	0/8	0/8

Table 3: Post-initialization evaluation budget required to first reach 95% mean top-1 accuracy.

Suite	Setting	Adaptive	$\lambda = 0$	$\lambda = 1/3$	$\lambda = 2/3$	$\lambda = 1$	λ^*
With side information							
Composite	Full pool	**134,784**	195,712	170,176	147,840	158,592	146,368
Composite	$\Delta = 0.01$	107,264	159,296	113,344	103,744	**96,320**	102,656
Composite	$\Delta = 0.02$	**91,200**	141,696	137,280	100,224	92,416	92,544
Composite	$\Delta = 0.03$	89,088	148,096	114,752	**82,816**	92,480	92,608
MMLU-Pro	Full pool	**119,744**	128,768	129,024	125,376	145,024	137,088
MMLU-Pro	$\Delta = 0.01$	**93,568**	115,520	94,272	104,448	114,048	104,896
MMLU-Pro	$\Delta = 0.02$	**76,032**	123,456	85,952	77,056	93,120	71,168
MMLU-Pro	$\Delta = 0.03$	66,496	78,336	68,032	59,968	**58,688**	58,880
Without side information							
Composite	Full pool	**154,432**	195,712	156,288	155,392	172,736	163,456
Composite	$\Delta = 0.01$	**121,920**	159,296	131,328	128,320	145,856	117,312
Composite	$\Delta = 0.02$	114,048	141,696	115,136	128,448	**107,392**	115,136
Composite	$\Delta = 0.03$	99,072	148,096	113,728	104,512	**97,280**	108,544
MMLU-Pro	Full pool	126,976	128,768	126,336	139,392	**125,568**	138,240
MMLU-Pro	$\Delta = 0.01$	**97,472**	115,520	109,632	97,600	114,816	110,016
MMLU-Pro	$\Delta = 0.02$	100,736	123,456	**82,112**	91,584	108,672	103,872
MMLU-Pro	$\Delta = 0.03$	**66,112**	78,336	68,352	67,008	69,376	63,616

Table 4: Mean ρ at the first retrain and accuracy at 100K. The accuracy gap is PULSE minus UCB-E in percentage points.

Dataset	Retrain-1 mean ρ	PULSE @100K	UCB-E @100K	Accuracy gap
bench1 / $\delta = 0.02$	0.7414	97.80%	95.00%	+2.80 pts
bench1 / $\delta = 0.03$	0.7415	98.00%	96.20%	+1.80 pts
bench2 / $\delta = 0.02$	0.6790	95.60%	87.20%	+8.40 pts
bench2 / $\delta = 0.03$	0.6792	95.80%	90.80%	+5.00 pts

B.6 Prediction Quality

Table 4 report the average correlation across all models at T=100K. We can see that the correlation is high enough to drive approximately 50% variance reduction. Specifically, the table below shows that this correlation yields an approximate $1 - 0.7^2 = 51\%$ reduction in variance. The advantage of PULSE is also evident from the experiments: it achieves a 2-8% accuracy improvement compared to standard UCB-E.

B.7 Statistical Comparisons

We compare between the accuracy of each baseline and the accuracy of PULSE, at a fixed budget. Concretely, for each budget we ran a one-sided McNemar test, where the null hypothesis is (H_0: PULSE $\leq$ baseline) and the alternative is: (H_1: PULSE $>$ baseline). We run the tests at four budgets that correspond to different accuracy thresholds: for each accuracy threshold $\tau \in \{0.90, 0.93, 0.95, 0.96\}$, we use the budget at which PULSE first reaches accuracy τ. Table 5 and Table 6 report the absolute accuracy gap $\Delta = \mathrm{acc}(\text{PULSE}) - \mathrm{acc}(\text{baseline})$ at the specific budget. In parentheses shown are the resulting p-values adjusted using Holm–Bonferroni correction (16 tests). As can be seen, all 32 comparisons are significant at $\alpha = 0.05$. We will add this analysis to the paper.

B.8 Comparison with MLA-UCB

We compare our method to MLA-UCB [24], which to our knowledge is the closest prediction-powered bandit method applicable to our setting. Concretely, we implement the batched-reward algorithm of MLA-UCB (Section 4.2 in [24]), which builds a different PPI-based mean estimator than ours. They designed the estimator to handle Gaussian reward, plus they assume the model producing

Table 5: PULSE vs. UCB-E.

Benchmark	90%	93%	95%	96%
Bench 1 $\Delta_{\text{top2}} = 0.02$	$+13.6\%$ (8.29×10^{-8})	$+12.2\%$ (3.77×10^{-8})	$+8.0\%$ (3.21×10^{-5})	$+6.0\%$ (4.71×10^{-4})
Bench 1 $\Delta_{\text{top2}} = 0.03$	$+7.4\%$ (1.60×10^{-3})	$+7.4\%$ (5.49×10^{-4})	$+3.8\%$ (1.45×10^{-2})	$+3.8\%$ (1.45×10^{-2})
Bench 2 $\Delta_{\text{top2}} = 0.02$	$+14.8\%$ (2.07×10^{-9})	$+12.2\%$ (4.98×10^{-8})	$+10.0\%$ (2.10×10^{-7})	$+5.8\%$ (4.54×10^{-4})
Bench 2 $\Delta_{\text{top2}} = 0.03$	$+10.6\%$ (2.72×10^{-5})	$+8.2\%$ (1.71×10^{-4})	$+6.8\%$ (5.49×10^{-4})	$+4.6\%$ (5.69×10^{-3})

Table 6: PULSE vs Naive-Pooling

Benchmark	90%	93%	95%	96%
Bench 1 $\Delta_{\text{top2}} = 0.02$	$+60.2\%$ (2.59×10^{-74})	$+61.2\%$ (8.11×10^{-78})	$+59.2\%$ (1.75×10^{-78})	$+56.8\%$ (2.59×10^{-74})
Bench 1 $\Delta_{\text{top2}} = 0.03$	$+60.6\%$ (3.19×10^{-76})	$+61.8\%$ (1.51×10^{-78})	$+56.8\%$ (4.89×10^{-72})	$+56.2\%$ (4.89×10^{-72})
Bench 2 $\Delta_{\text{top2}} = 0.02$	$+65.6\%$ (1.02×10^{-84})	$+65.4\%$ (2.64×10^{-85})	$+63.6\%$ (7.98×10^{-88})	$+60.6\%$ (1.48×10^{-83})
Bench 2 $\Delta_{\text{top2}} = 0.03$	$+65.4\%$ (3.20×10^{-86})	$+66.8\%$ (3.08×10^{-90})	$+64.6\%$ (8.68×10^{-92})	$+57.6\%$ (3.40×10^{-79})

the surrogate rewards is fixed; in our context, this means they cannot update the imputations as the algorithm progresses.

Since MLA-UCB requires the surrogate rewards to be i.i.d., we make the following adjustments: (i) we fit the low-rank predictor on historical data and hold it fixed for the remainder of the run. (2) We sample questions with replacements. These two modifications provide a faithful implementation of their method.

We run MLA-UCB with the same budget as PULSE and we don't count previously observed questions towards MLA-UCB budget. Since MLA-UCB samples with replacements, it may never use the entire LLM budget, thus we terminate the experiments after a number of questions equal to 10 times the original budget has been observed. Figure 7 shows the results. As can be observed, PULSE outperforms MLA-UCB in every setting.

B.9 Exploration Parameter Ablation

We ran UCB-E and PULSE on Bench1 $\Delta_{top2} = 0.02$ with different values of the exploration parameter a. The results are depicted in Figure 8. As we can see, PULSE outperforms UCB-E for every value of a which was tested.

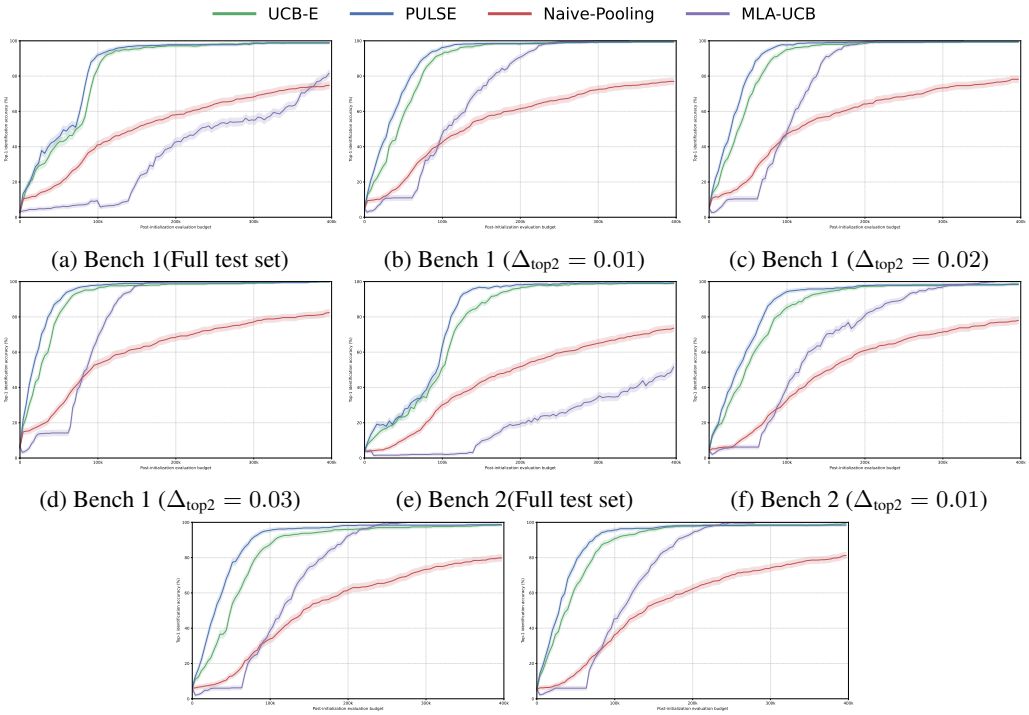
C Experiment Details

The dataset we use is taken from [26] and was downloaded from the paper's GitHub repository[3] at commit `857ee18`. The dataset is licensed under the Apache 2.0 license.

C.1 Hyperparameter

To fit the low-rank factorization model, we used the AdamW optimizer. Hyperparameters of the factorization—the regularization strength λ, rank r, learning rate, and number of training steps—are selected with the train split via 5-fold cross-validation procedure. Concretely, we use the 1-SE

[3] `https://github.com/skbwu/efficiently-evaluating-llms/tree/main/data/processed`

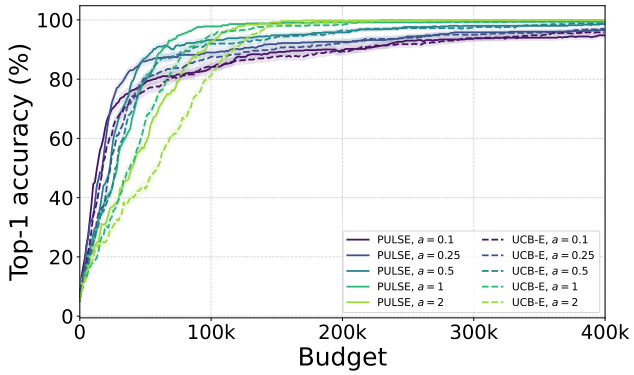


(a) Bench 1(Full test set) (b) Bench 1 ($\Delta_{\text{top2}} = 0.01$) (c) Bench 1 ($\Delta_{\text{top2}} = 0.02$)

(d) Bench 1 ($\Delta_{\text{top2}} = 0.03$) (e) Bench 2(Full test set) (f) Bench 2 ($\Delta_{\text{top2}} = 0.01$)

(g) Bench 2 ($\Delta_{\text{top2}} = 0.02$) (h) Bench 2 ($\Delta_{\text{top2}} = 0.03$)

Figure 7: Accuracy versus budget comparing to MLA



(a) Bench 1($\Delta_{\text{top2}} = 0.01$)

Figure 8: Exploration Parameter ablation

rule [33] that favors more parsimonious configurations among those within one standard error of the best cross-validated loss. The cross-validation resulted in the following parameters:

Bench 1. When fitting the low-rank model on the training data we use: $r = 100$, $\lambda = 0.001$, learning-rate=0.01, and 5 epochs. For the initialization phase we use: $\lambda = 0.01$, learning-rate=0.01,epochs=50. During the bandit loop we use learning-rate=0.1, $\lambda = 0.01$, epochs=10.

Bench 2. When fitting the low-rank model on the training data we use: $r = 100$, $\lambda = 0.01$, learning-rate=0.01, and 5 epochs. For the initialization phase we use: $\lambda = 0.01$, learning-rate=0.01,epochs=50. During the bandit loop we use learning-rate=0.1, $\lambda = 0.01$, epochs=10.

All methods used $a = 1$ to run the algorithm.

C.2 Experiments compute resources

Experiments were conducted on cpus and GPUs. We use Nvidia A40. A single run of our algorithm with budget $540,800$ took 60 seconds on average.

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: The abstract and the introduction's contribution bullets accurately reflect Lemma 1, Theorem 1, and Theorem 2, and are corroborated by the experiments in Section 4; the "up to 46% savings" claim appears in the introduction is bounded to the 95% accuracy target on ∼2.2K models across six benchmarks, matching the empirical scope.

   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: Section 5 contains a paragraph beginning "Our approach has several limitations" that explicitly discusses (i) the dependence of the speedup on low-rank-factorization quality, (ii) the restriction to binary scoring functions, and (iii) the use of uniform within-model example sampling.

   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: All five theoretical results (Lemma 1, Theorem 1, Lemma 2, Theorem 2, Lemma 6) are numbered, cross-referenced, with their assumptions stated in the theorem statements, and proved 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: Section 4 states the data splits and the bandit-side settings (T, B, a, τ, T_0, 500 seeds, 1-SE CV rule), Algorithm 1 fully specifies the two-stage proxy and bandit step, and Appendix C reports the optimizer and the per-stage rank, λ, learning rate, and epochs for both benchmarks. The score matrix from [26] is publicly available with URL and commit hash given in Appendix C, and an anonymized code repository at `https://anonymous.4open.science/r/PULSE-37E6/` contains the PULSE implementation.

   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: An anonymized code repository at `https://anonymous.4open.science/r/PULSE-37E6/` (introduced in Section 1) contains the PULSE implementation; the underlying score-matrix dataset is publicly available, with URL and commit hash given in Appendix C.

   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: Section 4 states the data splits, T_0, τ, B, T, a, 500 seeds, and the 1-SE CV rule used to select the LR hyperparameters; Appendix C lists the optimizer and the per-stage rank, λ, learning rate, and epochs for both benchmarks.

   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: All accuracy-versus-budget curves (Figure 2, Figure 6) include standard error in a shaded areas around each curve. The Figure 1 includes error bars.

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: The "Experiments compute resources" subsection of Appendix C states that experiments were conducted on CPUs and Nvidia A40 GPUs and that a single run of PULSE at budget $T = 540{,}800$ took 60 seconds on average.

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: The work is purely algorithmic and statistical, uses only existing public benchmarks (MMLU-Pro, BBH, GPQA, IFEval, MATH, MuSR, accessed via the score-matrix release of [26]), involves no human subjects, and releases no new high-risk asset; we have reviewed the NeurIPS Code of Ethics and the research conforms to 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: Our works is a statistical algorithm focused only on efficient evaluation of machine learning models.

    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: PULSE is a statistical best-arm-identification algorithm operating on existing public benchmark score matrices; the paper releases no pre-trained models, generative models, or scraped datasets, so no high-misuse asset requires safeguards.

    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: The score-matrix dataset we actually consume is the release of [26], downloaded from the URL and commit hash given in Appendix C and noted as Apache 2.0; the constituent benchmarks (MMLU-Pro [4], BBH [28], GPQA [29], IFEval [30], MATH [31], MuSR [32]) are properly credited via citation.

    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: [N/A]

    Justification: The paper does not introduce or release any new dataset, model, or benchmark; PULSE is an algorithm fully described in the manuscript, and no code is currently being released as a supplemental asset.

    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 paper involves no crowdsourcing or human-subjects research; experiments rely solely on public LLM benchmarks (originally curated by their authors) and the score-matrix release of [26], with no human participants recruited or studied.

    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: This work evaluates language-model benchmarking algorithms on existing public benchmark score matrices; there is no crowdsourcing, no human-subjects data collection, and no participants involved, so IRB review does not apply.

    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: [N/A]

    Justification: LLMs were only used to assist with coding and writing.

    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.