EFFICIENT EVALUATION OF REASONING BENCHMARKS WITH LIMITED HISTORICAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accuracy estimates on reasoning benchmarks can be noisy, with uncertainty arising from the small number of benchmark items and the stochastic generation of responses by language models. Repeated evaluation can reduce this uncertainty but requires additional generations, raising the question of how to allocate a generation budget across items to minimize error in estimating benchmark accuracy. We formulate this problem as sequential item selection and develop adaptive sampling policies that leverage intermediate evaluation results from the target model to guide subsequent sampling. Our sampling policies also incorporate structural information about item difficulty from historical evaluations of other models on the benchmark, such as difficulty strata or a relative difficulty order. We provide theoretical analysis showing that informative difficulty strata can reduce estimation error, while a reliable relative difficulty order can tighten an estimation error bound. Experiments show that the resulting policies reduce error in estimating benchmark accuracy across a range of budgets and levels of available historical information. We summarize practical guidelines for choosing an evaluation strategy according to the historical information available, enabling efficient and reliable evaluation.

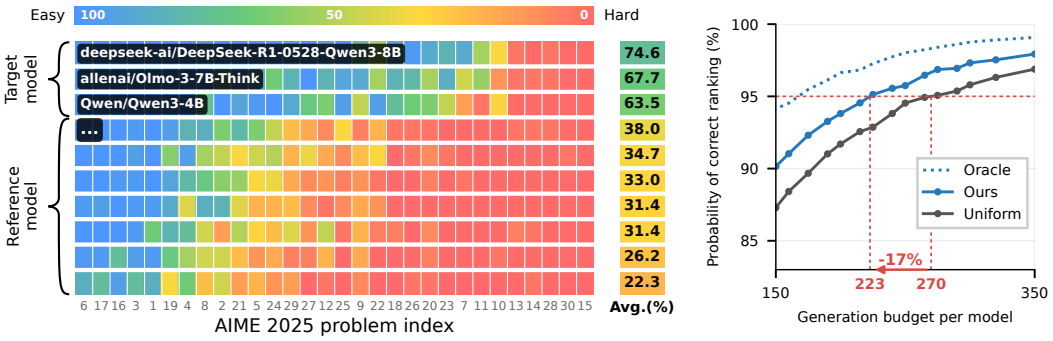
1 INTRODUCTION

Accuracy estimates of language models on reasoning benchmarks can be noisy, because model responses are generated stochastically and typical benchmarks contain few items. Reported accuracy on a benchmark can differ across evaluation settings and random seeds (Blackwell et al., 2024; Hochlehnert et al., 2025), making performance comparisons between models challenging (Biderman et al., 2024; Miller, 2024; Madaan et al., 2024; Lee et al., 2026). Reliable evaluation therefore requires repeated sampling to reduce this uncertainty. For example, Qwen Team (2025) evaluate `Qwen3` models on the AIME 2025 benchmark by sampling each of its 30 problems 64 times and reporting the average accuracy, requiring 1,920 responses per model. The resulting generation cost motivates methods that achieve reliable accuracy estimates with fewer generations.

Given a generation budget, evaluation can be made more efficient by allocating different numbers of responses to different items. Since the target model can have different accuracies across items, an optimal allocation for minimizing error in estimating benchmark accuracy is to assign more evaluations to items whose outcomes are more uncertain (Neyman, 1934). However, the optimal allocation depends on the item-level accuracies that evaluation seeks to estimate. Adaptive sampling policies estimate these item-level accuracies from accumulated target responses and derive allocation priorities from the estimates to decide which item to evaluate next (Carpentier et al., 2015; Saha et al., 2026). At small budgets, too few target responses may be available to estimate these priorities reliably, while at larger budgets, the absolute gain from adaptive allocation becomes smaller.

Historical evaluations of other models on the same benchmark can provide additional information for more efficient evaluation. Existing approaches typically assume access to extensive historical evaluations from previously evaluated models (Vivek et al., 2024; Maia Polo et al., 2024; Zhang et al., 2025). However, such historical information may be limited for newly introduced, updated, or private benchmarks (White et al., 2025; Rajore et al., 2024). Moreover, Zhang et al. (2025) show that when the target model is stronger than the historically evaluated reference models, the methods they evaluate do not consistently outperform evaluation based on uniformly sampled items. These

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



(a) Accuracy for individual items and average accuracy across models

(b) Ranking recovery

Figure 1: **The relative order of items in difficulty is consistent across models and can be used to evaluate language models more efficiently. (a)** For ten models (rows), we report the accuracy of each item (left columns) and the average accuracy (right column) on AIME 2025. Each item is evaluated with 32 generations using different random seeds, and items are ordered from easier to harder based on their average accuracy across models. The relative order is mostly consistent across the ten models. **(b)** We estimate the itemwise accuracies of the three strongest target models under various generation budgets and measure whether they recover the true ranking. We compare our method, which uses the difficulty order from the seven weaker reference models, with uniform allocation, which distributes evaluations evenly across items. Our method reaches 95% correct ranking with approximately 17% fewer generations. See Section 5.4 for details.

limitations motivate evaluation methods that can use historical information even when it is limited or does not transfer reliably to the target model.

Historical evaluations may nonetheless reveal stable relationships among item difficulties that remain informative for a new target model. Figure 1(a) shows similar difficulty patterns across models on AIME 2025: across ten models with substantially different overall accuracies, the relative difficulty pattern across items remains broadly similar. Moreover, benchmark annotations can provide item-level difficulty information that can be grouped into difficulty strata (Ding et al., 2024), while historical evaluations can provide an approximate relative difficulty order of items. We therefore use historical information only through such relationships among items, in the form of difficulty strata or a relative item order, while estimating item-level accuracies from the target model's own responses.

In this paper, we formulate evaluation as sequential item selection given a generation budget and study adaptive sampling policies that use intermediate results from the target model together with limited historical difficulty information. Difficulty strata group items with similar difficulty and use their responses together to estimate which groups need more evaluations, while a relative difficulty order uses the expected ordering of item difficulties to refine accuracy estimates before deciding what to sample next. Figure 1(b) shows that even a difficulty order obtained from weaker reference models can reduce the number of generations needed to reliably rank three strong models.

We provide practical guidelines for selecting an evaluation strategy according to the available historical information in Figure 2, and we summarize our contributions as follows.

- In Sections 2 and 3, we formulate benchmark evaluation as sequential item selection and develop adaptive sampling policies that use intermediate target responses together with limited historical information represented by difficulty strata or a relative difficulty order.
- In Section 4, we provide theoretical analysis showing that informative difficulty strata can reduce estimation error, while a reliable relative difficulty order can tighten an estimation error bound.
- In Section 5, we show on four reasoning benchmarks that our adaptive sampling policies reduce error in estimating benchmark accuracy relative to uniform allocation across a range of budgets and levels of available historical information. When reference models are few or weaker than the target model, our policies also achieve lower error than existing methods that rely on historical evaluations. This reduction remains even when the difficulty order is obtained from benchmark metadata alone, without any historical evaluations.

An extended discussion of related work is provided in Appendix A.

2 PRELIMINARIES

We consider evaluating a target model on a benchmark consisting of n items, indexed by $[n] := \{1, \cdots, n\}$. Because the model may generate different responses to the same item across repeated trials, we represent its correctness on item i by a binary random variable $Y_i \in \{0, 1\}$. We define the item-level accuracy as $\theta_i := \mathbb{E}[Y_i]$, representing the probability that the model answers item i correctly under a given generation setting. The model's accuracy on the benchmark is then

$$\theta := \frac{1}{n} \sum_{i \in [n]} \theta_i = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}[Y_i]. \tag{1}$$

Our goal is to estimate the accuracy θ using a total budget of m generations. At each generation, an evaluation procedure selects an item to evaluate, possibly based on previously observed outcomes. After m generations, let $m_i \geq 0$ denote the total number of times item i has been evaluated. For any item with $m_i > 0$, we denote the observed correctness of its jth trial by $y_{i,j} \in \{0, 1\}$, where $j \in \{1, \cdots, m_i\}$. Thus, some items may remain unevaluated while others are evaluated multiple times. For each evaluated item, we estimate its item-level accuracy by $\hat{\theta}_i := \frac{1}{m_i} \sum_{j \in [m_i]} y_{i,j}$. When every item is evaluated, the benchmark accuracy is estimated by

$$\hat{\theta} := \frac{1}{n} \sum_{i \in [n]} \hat{\theta}_i. \tag{2}$$

When some items are not evaluated, we instead average $\hat{\theta}_i$ over the evaluated items.

Allocation objective and oracle. Under a generation budget m, we can choose the number of evaluations m_i assigned to each item, subject to $\sum_{i \in [n]} m_i = m$. These counts determine how many responses are observed for each item and therefore affect the estimation error of $\hat{\theta}$, defined as

$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}\big[(\hat{\theta} - \theta)^2\big]. \tag{3}$$

If the item-level accuracies θ_i were known, these counts m_i could be chosen to minimize the error in (3) (Neyman, 1934; Carpentier et al., 2015). The oracle allocation[1] is

$$m_i^\star := m \, \frac{\sqrt{\theta_i(1 - \theta_i)}}{\sum_{i' \in [n]} \sqrt{\theta_{i'}(1 - \theta_{i'})}}. \tag{4}$$

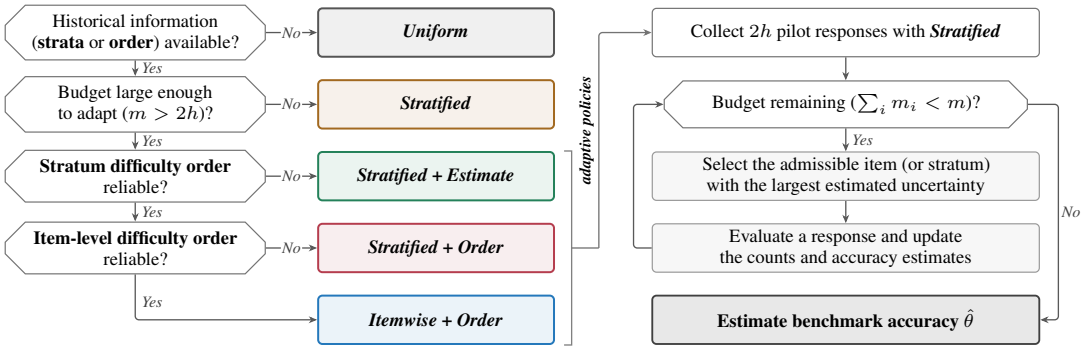
Thus, the oracle allocates more evaluations to items with greater response uncertainty. In practice, the item-level accuracies θ_i are unknown and the oracle allocation cannot be computed directly.

Historical information. Historical evaluations on the same benchmark may provide information about the structure of the unknown item-level accuracies θ_i. We consider two forms of such information: *difficulty strata* and a *relative difficulty order*.

Difficulty strata group items whose item-level accuracies are expected to be similar. We denote a partition of the benchmark index set $[n]$ into h strata by $I_1, \cdots, I_h$. For any $i, i' \in I_k$ with $k \in [h]$, we expect θ_i and $\theta_{i'}$ to be close. A relative difficulty order instead provides information about the ordering of the item-level accuracies. After reindexing items from easier to harder, a perfectly accurate ordering for the target model would imply $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$. Neither relationship needs to hold exactly for the target model. How well these two forms of historical information transfer to the target model depends on the degree of within-stratum homogeneity and the agreement of the supplied order with the ordering of the item-level accuracies, respectively.

Problem statement. Given a generation budget m and available historical information, our objective is to minimize the estimation error in (3) by using the historical structure and observed target responses to determine which items should receive additional evaluations. Specifically, our allocation strategies follow the oracle principle in (4) by assigning more evaluations to items estimated to have greater response uncertainty. Section 3 develops adaptive sampling strategies for this setting.

[1] For the oracle allocation, we relax the integer constraint and allow m_i to take continuous values.



(a) Choose a sampling policy given available historical information **(b)** Evaluate with an adaptive policy

Figure 2: **Overview of the evaluation framework.** We estimate a model's accuracy on n benchmark items from m generated responses. **(a)** Historical information is given as h difficulty strata or a relative difficulty order. *Uniform* and *Stratified* allocate evaluations equally across items and strata, respectively. The three adaptive policies estimate the accuracy of each item (or stratum) from intermediate evaluation results; in the estimation, *Stratified + Order* and *Itemwise + Order* have an additional constraint that the estimates have to be consistent with the historical difficulty order. Adaptive policies require a budget larger than the $2h$ pilot responses (two per stratum) needed to initialize these estimates ($m > 2h$); otherwise, the pilot uses the entire budget and they reduce to *Stratified*. **(b)** After the pilot, an adaptive policy selects the admissible item (or stratum) with the largest estimated uncertainty in response correctness for each evaluation and updates its estimates with the result. See Section 3 for details.

3 Adaptive Sampling with Limited Historical Information

The oracle allocation in (4) is unavailable in practice because it depends on the unknown uncertainty in response correctness. We therefore approximate it sequentially with a *sampling policy*, a rule that selects the next item to evaluate based on current estimates of this uncertainty. These estimates use intermediate results of the target model, regularized by historical difficulty information when available. We consider five policies. *Uniform* and *Stratified* allocate evaluations equally across items and strata, respectively, without using these estimates. The three *adaptive* policies (*Stratified + Estimate*, *Stratified + Order*, and *Itemwise + Order*) prioritize items with larger estimated uncertainty. Figure 2 summarizes which policy to use given the available information.

Limiting imbalance in evaluation counts. All policies share a constraint that limits how unequally evaluations can be allocated across items, so that approaching the oracle allocation does not come at the expense of covering every item. Without such a constraint, an adaptive policy can fail in two ways. First, priorities estimated from only a few early results can be inaccurate, and the policy may commit much of the budget based on them. Second, items with low priority may never be evaluated, and the benchmark accuracy estimate, which then excludes them, would be biased.

We index generations in order by $t \in [m]$. We define $m_i(t)$ as the number of evaluations of item i up to generation t and the *imbalance tolerance* as an integer $\ell \geq 1$ that bounds the gap between the largest and smallest evaluation counts; that is, every policy maintains

$$\max_{i \in [n]} m_i(t) - \min_{i \in [n]} m_i(t) \leq \ell \quad \text{for all } t. \tag{5}$$

An item is *admissible* if evaluating it next preserves (5), and every policy selects only admissible items. For example, with $\ell = 1$, every item must be evaluated once before any item is evaluated twice; with $\ell = 2$, some items can be evaluated twice before others are evaluated at all. A larger ℓ thus gives a policy more freedom to allocate evaluations unequally. In particular, when $\ell = 1$ and $m = n$, every item is evaluated exactly once, so all policies yield the same allocation.

Item selection by priority. Each policy assigns a *priority* to every item (or stratum) and selects the admissible item (or stratum) with the largest priority for the next evaluation. For all policies except *Uniform*, ties in priority are broken by selecting a candidate with the fewest evaluations, with remaining ties broken uniformly at random.

Uniform samples uniformly among admissible items. When $\ell = 1$, this is equivalent to sampling uniformly among the least-evaluated items. *Stratified* assigns equal priority to all admissible strata and selects one with the fewest evaluations, breaking ties uniformly at random. The adaptive policies follow the principle of the oracle allocation in (4), which assigns more evaluations to items with greater uncertainty: the priority is the estimated standard deviation $\sqrt{p(1-p)}$ of the correctness of a single response, where p is the policy's current accuracy estimate for that item (or stratum). The adaptive policies differ only in how they estimate p, and they update this estimate after each evaluation, starting from a short pilot described in Section 3.4.

3.1 Sampling without historical information

Uniform. *Uniform* assigns the same priority to every item. It is used when no historical information is available. It estimates benchmark accuracy by (2) in Section 2.

Averaging uniformly sampled items remains a strong baseline for efficient benchmark evaluation, especially when the target model is more accurate than previously evaluated models (Zhang et al., 2025). We therefore compare all other policies against *Uniform*.

3.2 Sampling with difficulty strata

Items in a difficulty stratum are expected to have similar item-level accuracies, so we regularize the estimates by pooling the responses within each stratum into a single stratum-level accuracy. For the policies that select strata, we define a stratum as admissible when it contains at least one admissible item; after a stratum is selected, we evaluate an admissible item with the smallest m_i within that stratum and choose uniformly at random if several such items exist. These policies estimate benchmark accuracy by a weighted average of the stratum-level accuracies:

$$\hat{\theta} := \sum_{k\in[h]} \frac{|I_k|}{n} \frac{\sum_{i\in I_k}\sum_{j\in[m_i]} y_{i,j}}{\sum_{i\in I_k} m_i}. \tag{6}$$

Stratified. *Stratified* assigns the same priority to every stratum, so it allocates evaluations equally across strata. It is used when difficulty strata are available but the budget is too small to adapt.

Stratified + Estimate. *Stratified + Estimate* estimates the accuracy p of each stratum from its pooled responses. It is used when difficulty strata are available but their order is not reliable. Specifically, it adds half a correct and half an incorrect pseudo-response to the observed responses of each stratum and sets the priority accordingly:

$$\hat{p}_k := \frac{\sum_{i\in I_k}\sum_{j\in[m_i]} y_{i,j} + 1/2}{\sum_{i\in I_k} m_i + 1}, \qquad q_k := \sqrt{\hat{p}_k(1-\hat{p}_k)}. \tag{7}$$

The pseudo-responses keep $\hat{p}_k$ strictly between zero and one even when all responses in a stratum are correct or all are incorrect (Krichevsky & Trofimov, 1981), so no stratum's priority collapses to zero.

3.3 Sampling with a relative difficulty order

A relative difficulty order implies that item-level accuracies should not increase from easier to harder items. The two policies in this section therefore regularize their accuracy estimates by constraining them to follow the order through weighted isotonic regression, which pools adjacent estimates that violate the order (Ayer et al., 1955; Barlow et al., 1972).

Stratified + Order. *Stratified + Order* extends *Stratified + Estimate* by adjusting its stratum-level accuracy estimates to follow a difficulty order of the strata. It is used when this stratum order is reliable but the order of individual items is not. The stratum order may be given directly or obtained by grouping consecutive items in a relative difficulty order. The policy fits the estimates in (7) to this order, giving more weight to strata with more evaluations:

$$\tilde{\boldsymbol{p}} := \arg\min_{1\ge u_1\ge\cdots\ge u_h\ge 0} \sum_{k\in[h]} \Big(\sum_{i\in I_k} m_i + 1\Big)(\hat{p}_k - u_k)^2. \tag{8}$$

It then uses $\tilde{p}_k$ in place of $\hat{p}_k$ in the priority (7).

Itemwise + Order. *Itemwise + Order* applies the order at the item level, fitting the item-level accuracies to the order. It is used when the item-level order is reliable, and it requires only this order. The policy applies the fit in (8) with each item as its own stratum, using the item-level accuracy estimates $\hat{\theta}_i$ of the evaluated items with weights m_i. For items with $m_i = 0$, the accuracies are estimated by linear interpolation between neighboring evaluated items in the order and by constant extension beyond the evaluated range. Thus every item receives a fitted accuracy $\tilde{\theta}_i$, including items not yet evaluated, which serves as its accuracy estimate p. Unlike the other adaptive policies, *Itemwise + Order* also uses the fitted item-level accuracies to estimate the final benchmark accuracy, $\hat{\theta} := \frac{1}{n}\sum_{i\in[n]} \tilde{\theta}_i$. This allows items with $m_i = 0$ to contribute to the final estimate through the order.

3.4 Practical choices

Pilot. The adaptive policies need some evaluation results before computing priorities. Following prior work (Carpentier et al., 2015), each adaptive policy therefore starts with a *pilot* of $2h$ generations allocated by *Stratified*, two per stratum. Since *Itemwise + Order* is given only an item order, it forms the h strata for its pilot by splitting the ordered items into h consecutive groups of nearly equal size. The pilot evaluations count toward the budget m and are retained for subsequent allocation and estimation. If the budget is no larger than the pilot, the pilot uses the entire budget.

Number of strata. The number of strata h controls the amount of pooling. Fewer strata combine observations from more items, while more strata preserve information about differences in difficulty. We use $h = 5$ by default and $h = 20$ for benchmarks with many items, such as GPQA Diamond.

Imbalance tolerance. A larger imbalance tolerance ℓ in (5) allows allocation closer to the oracle, whereas a smaller ℓ better protects against the failure modes described above. We use $\ell = 1$ for *Uniform* and $\ell = \max\{1, \lfloor m/n \rfloor\}$ for the other policies, so the permitted imbalance grows with the budget per item. A tolerance proportional to m/n permits an allocation that retains a fixed fraction of the oracle gain while assigning every item a positive number of evaluations.

4 Theoretical Analysis

We analyze when historical difficulty information reduces the error in estimating benchmark accuracy, in a simplified setting where the allocation proportions are estimated from preliminary responses of the target model and are independent of the responses used for the final estimate.[2] The benefit depends on whether items within each stratum have similar accuracies for the target model and whether the difficulty order agrees with its item-level accuracies. We further show how the imbalance tolerance should scale with the budget. Proofs are provided in Appendix E.

Why reliable strata improve estimation. Policies that use difficulty strata allocate evaluations at the stratum level. For clarity, we consider balanced strata with $|I_k| = n/h$ for all $k \in [h]$ in the main text. The general case for arbitrary stratum sizes is given in Appendix E.

These policies correspond to an estimator $\widehat{\theta}_{\mathrm{strat}}$. Let $\widehat{\pi}_k$ be the fraction of the budget m allocated to stratum I_k, estimated from the preliminary responses, so that $\sum_{k\in[h]} \widehat{\pi}_k = 1$. The estimator assigns $m\widehat{\pi}_k$ evaluations to stratum I_k, distributes them equally within the stratum, and estimates benchmark accuracy by (6). We assume that each item receives a positive integer number of evaluations and compare $\widehat{\theta}_{\mathrm{strat}}$ with the estimate $\widehat{\theta}_{\mathrm{unif}}$ from *Uniform*.

Proposition 1. *Assume $p_k \in (0,1)$ and $\widehat{\pi}_k > 0$ for all $k \in [h]$. Then we have*

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}_{\mathrm{strat}}) \geq \frac{1}{m}\left[\frac{1}{h}\sum_{k\in[h]}(\sigma_k - \bar{\sigma})^2 - \frac{1}{n}\sum_{k\in[h]}\sum_{i\in I_k}(\theta_i - p_k)^2 - \bar{\sigma}^2 \sum_{k\in[h]} \frac{(\widehat{\pi}_k - \pi_k^\star)^2}{\widehat{\pi}_k}\right], \quad (9)$$

where $p_k := \frac{h}{n}\sum_{i\in I_k}\theta_i$ is the stratum-level accuracy, $\sigma_k := \sqrt{p_k(1-p_k)}$, $\bar{\sigma} := \frac{1}{h}\sum_{k\in[h]}\sigma_k$, and $\pi_k^\star := \sigma_k/(h\bar{\sigma})$ is the oracle fraction when items within each stratum share the same accuracy p_k.

[2] Independence makes $\hat{\theta}$ unbiased given the allocation, so its estimation error in (3) equals its variance. As in (4), evaluation counts are treated as continuous. In experiments, our policies use all observed responses for both allocation and estimation, so that every response is used for the final estimate.

Proposition 1 gives a condition for historical strata to reduce estimation error. The first term in (9) is the gain from differences in the uncertainty in response correctness across strata, determined by the benchmark and the target model. The second term decreases as items within each stratum have more similar accuracies, and the third term decreases as $\widehat{\pi}_k$ approaches $\pi_k^\star$. When the first term exceeds the other two, the bound guarantees a reduction in estimation error.

Why a reliable order tightens an estimation error bound. Policies that use a difficulty order regularize their accuracy estimates by constraining them to follow the order. Given estimates $\hat{\theta}_i$ from m_i responses to item i, let $\widetilde{\theta}_i$ denote the fitted estimate obtained by weighted isotonic regression with weights m_i as in Section 3.3. These policies correspond to $\widehat{\theta}_{\mathrm{iso}}$, which uses (4) with θ_i replaced by $\widetilde{\theta}_i$, and compare it with $\widehat{\theta}_{\mathrm{raw}}$, which uses $\widehat{\theta}_i$.

To measure how reliable the order is for the target model, we define the order mismatch $\delta := \min_{u_1 \geq \cdots \geq u_n} \|\boldsymbol{\theta} - \boldsymbol{u}\|_{\boldsymbol{m}}$, where boldface denotes vectors, such as $\boldsymbol{\theta} = (\theta_i)_{i\in[n]}$, and $\|\boldsymbol{v}\|_{\boldsymbol{m}}^2 := \sum_{i\in[n]} m_i v_i^2$. A smaller δ indicates a more reliable order, and $\delta = 0$ when the order is correct.

Lemma 1. *Assume $m_i > 0$ for all $i \in [n]$ and $2\delta \leq \|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}$. Then $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2 \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2$.*

Proposition 2. *Assume $m_i > 0$ and $\widehat{\theta}_i \in (0,1)$ for all $i \in [n]$ and $2\delta \leq \|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}$. Then we have*

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{iso}}) - \mathrm{MSE}(\widehat{\theta}^\star) \leq \underbrace{c\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2}_{\text{upper bound on } \mathrm{MSE}(\widehat{\theta}_{\mathrm{raw}}) - \mathrm{MSE}(\widehat{\theta}^\star)} - c\big[\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}^2 - 2\delta\,\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}\big], \tag{10}$$

where $\widehat{\theta}^\star$ uses the oracle allocation and $c > 0$ is a constant independent of $\widetilde{\boldsymbol{\theta}}$, specified in Appendix E.

Proposition 2 shows that a reliable difficulty order can tighten an upper bound on the excess estimation error over the oracle. The first term in (10) is the corresponding bound based on the estimates $\hat{\theta}_i$, while the second term captures the improvement from the isotonic fit and is positive when $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}} > 2\delta$.

Why the tolerance should scale with the budget. We examine whether a tolerance proportional to the budget per item $\frac{m}{n}$ can permit an allocation that retains part of the oracle gain. We compare the estimator $\hat{\theta}_{\mathrm{unif}}$ under equal allocation $m_i = \frac{m}{n}$ with the oracle estimator $\hat{\theta}^\star$.

Proposition 3. *Assume $\frac{1}{n}\sum_{i\in[n]} \sqrt{\theta_i(1-\theta_i)} \geq c$ for some $c > 0$. Then the oracle counts satisfy*

$$\max_{i\in[n]} m_i^\star - \min_{i\in[n]} m_i^\star \leq \frac{m}{2cn}. \tag{11}$$

Moreover, there exists an allocation whose final counts differ by at most $m/(2n)$ and give every item at least $(1-c)m/n$ evaluations, and whose estimator $\widehat{\theta}_{\mathrm{mix}}$ satisfies

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}_{\mathrm{mix}}) \geq c\left[\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}^\star)\right]. \tag{12}$$

Thus, a tolerance $\ell \geq m/(2n)$ suffices for the imbalance constraint to permit an allocation that retains a fraction c of the oracle gain, which motivates scaling ℓ with the budget per item in Section 3.4.

5 Experiments

We examine the sampling policies in Section 3 on reasoning benchmarks through four questions. Section 5.1 asks whether historical difficulty information reduces the error in estimating benchmark accuracy under a given generation budget. Section 5.2 asks whether this benefit remains with only a few reference models and how it compares with existing methods that rely on historical evaluations. Section 5.3 asks how the benefit depends on the reliability of the difficulty order and whether an order from benchmark metadata alone suffices—without any historical evaluations. Section 5.4 asks whether more accurate estimates reduce the number of generations needed to rank models.

Table 1: Percentage MAE (with standard error) in estimating benchmark accuracy on AIME 2025. Bold and underlined values mark the lowest and second-lowest errors in each row.

m	m/n	*Uniform*	*Stratified*	*Stratified + Estimate*		*Stratified + Order*		*Itemwise + Order*	
		$\ell = 1$	$\ell = 1$	$\ell = 1$	$\ell = 2$	$\ell = 1$	$\ell = 2$	$\ell = 1$	$\ell = 2$
15	0.50	7.84 (0.06)	6.14 (0.05)	6.10 (0.05)	6.18 (0.05)	5.73 (0.04)	5.84 (0.04)	<u>5.66</u> (0.04)	**5.58** (0.04)
20	0.67	6.25 (0.05)	5.25 (0.04)	5.03 (0.04)	5.48 (0.04)	**4.81** (0.04)	4.96 (0.04)	<u>4.87</u> (0.04)	**4.81** (0.04)
30	1.00	4.08 (0.03)	4.08 (0.03)	4.08 (0.03)	4.04 (0.03)	4.08 (0.03)	**3.85** (0.03)	4.08 (0.03)	<u>3.91</u> (0.03)
40	1.33	3.70 (0.03)	3.63 (0.03)	3.34 (0.03)	3.41 (0.03)	<u>3.30</u> (0.03)	**3.28** (0.02)	3.35 (0.03)	3.37 (0.03)
50	1.67	3.30 (0.02)	3.21 (0.02)	3.01 (0.02)	2.99 (0.02)	<u>2.97</u> (0.02)	**2.93** (0.02)	3.04 (0.02)	3.01 (0.02)

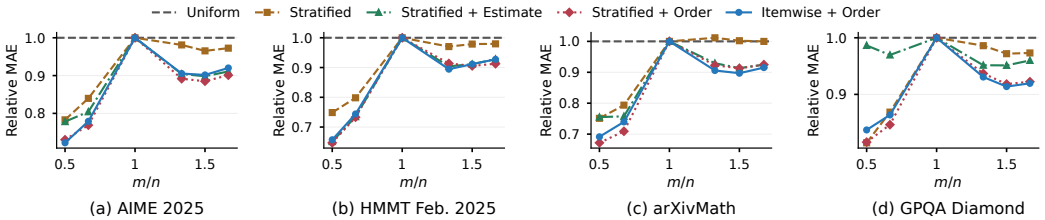


Figure 3: Ratio of each policy's MAE to that of *Uniform* against the budget per item m/n.

Experimental setup. We use four reasoning benchmarks: AIME 2025, HMMT February 2025, arXivMath, and GPQA Diamond. Since our experiments need repeated responses from several models to every item, we use both publicly released responses and responses from our own evaluations, obtaining 4 to 11 models per benchmark and 8 to 128 responses per item as detailed in Appendix B.

Since regenerating responses is expensive, we treat the fraction of correct stored responses as the true item-level accuracy θ_i and simulate responses from it. We report the mean absolute error (MAE) over 1,000 Monte Carlo simulations per target model, and Appendix C gives further details.

We construct historical difficulty information from reference models, by default all models other than the target model, such as nine models for AIME 2025 and three for GPQA Diamond. We order items by the average item-level accuracy of the reference models and split them into h consecutive nearly equal-size strata. Our policies use only this relative difficulty order or these difficulty strata.

5.1 Estimation under a given budget

We compare the sampling policies on AIME 2025 at the same generation budget in Table 1. For the adaptive policies, we report results for $\ell = 1$ and $\ell = 2$, where the imbalance tolerance ℓ in (5) bounds the gap between the largest and smallest evaluation counts. The larger tolerance tends to be beneficial once the budget per item reaches $m/n \geq 1$ and is less so at smaller budgets, which supports increasing ℓ with m/n as in Section 3.4. Using difficulty strata or a difficulty order generally reduces MAE relative to *Uniform*.

Figure 3 shows that using difficulty strata or a difficulty order also reduces the estimation error on the other benchmarks, where each policy's MAE is shown relative to that of *Uniform* and the tolerance follows the practical choice in Section 3.4. At $m = n$, every item is evaluated exactly once, so all policies produce the same allocation and the same final estimate. At other budgets, the order-based policies reduce the error relative to *Uniform* on every benchmark. These results suggest that even limited historical information, such as only a relative difficulty order, can substantially reduce the error relative to *Uniform*. Appendix D.1 further reports the results at $\ell = 1, 2$ for all four benchmarks.

5.2 Estimation with few reference models

We compare our methods, *Stratified + Order* and *Itemwise + Order*, with existing efficient evaluation methods that rely on historical evaluations in the two settings motivated in the introduction, where reference models are few or weaker than the target model. The considered methods are Anchor-Points-Weighted/Predictor (*AP-W*/*AP-P*) (Vivek et al., 2024), (Generalized) Performance Item Response Theory (*P-IRT*/*GP-IRT*) (Maia Polo et al., 2024), and Augmented Inverse Propensity Weighting (*AIPW*) (Zhang et al., 2025). On AIME 2025, we use one or nine reference models for all ten target models, and the three weakest models as references for the three strongest.

Table 2: Percentage MAE (with standard error) in estimating benchmark accuracy on AIME 2025.

References	m	m/n	*Uniform*	*Stratified + Order*	*Itemwise + Order*	*AP-W*	*AP-P*	*P-IRT*	*GP-IRT*	*AIPW*
1 reference	15	0.50	7.98 (0.06)	6.13 (0.05)	**6.10** (0.05)	——— Requires at least two references ———				
	50	1.67	3.30 (0.03)	**3.02** (0.02)	3.05 (0.02)					
9 references	15	0.50	7.88 (0.06)	5.78 (0.04)	**5.64** (0.04)	8.89 (0.06)	15.05 (0.11)	8.32 (0.06)	7.09 (0.05)	6.63 (0.05)
	20	0.67	6.36 (0.05)	**4.79** (0.04)	4.84 (0.04)	6.36 (0.04)	11.87 (0.10)	6.76 (0.05)	5.83 (0.04)	5.49 (0.04)
	50	1.67	3.34 (0.03)	**3.02** (0.02)	3.08 (0.02)	——— Identical to *Uniform* (when $m > n$) ———				
3 lowest	15	0.50	7.91 (0.11)	**6.81** (0.09)	6.94 (0.10)	18.97 (0.13)	64.47 (0.64)	9.45 (0.12)	8.66 (0.11)	7.61 (0.10)
	20	0.67	6.29 (0.09)	**5.46** (0.07)	5.61 (0.08)	18.57 (0.13)	54.37 (0.67)	7.35 (0.10)	6.73 (0.09)	6.06 (0.08)
	50	1.67	3.28 (0.04)	3.04 (0.04)	**3.01** (0.04)	——— Identical to *Uniform* (when $m > n$) ———				

In Table 2, our two methods attain the two lowest MAEs in every condition, even with a single reference model. The gap is largest when the reference models are weaker than the target models, where most existing methods, which predict accuracy from the reference models, perform worse than *Uniform*. Our methods use the reference models only to order items and estimate accuracies from the target model's responses. Appendix D.2 reports additional results on GPQA Diamond.

5.3 Estimation under unreliable and metadata-based difficulty orders

We next examine how different the order can be from the target model's difficulty while remaining useful. For *Itemwise + Order* on AIME 2025 with $m = 40$, we perturb the historical order so that its rank correlation with the unperturbed order decreases from 1 to 0 and its rank correlation with the target models' order decreases from 0.79 to 0. Table 3 reports the MAE reduction, defined as the MAE of *Uniform* minus that of *Itemwise + Order* in percentage points. The MAE reduction decreases as the correlation with the target models decreases, but remains positive down to a correlation of 0.4.

Table 3: MAE reduction of *Itemwise + Order* over *Uniform* under perturbed and metadata-based orders.

	Kendall rank correlation		MAE
Ordering	Reference	Target	reduction
Historical	1.00	0.79	$+\mathbf{0.35}$
	0.80	0.71	$+\mathbf{0.34}$
↓	0.60	0.56	$+\mathbf{0.21}$
	0.40	0.38	$+\mathbf{0.06}$
	0.20	0.19	-0.11
Random	0.00	0.00	-0.23
Problem number	–	0.52	$+\mathbf{0.22}$

Metadata alone provides a useful difficulty order. In AIME 2025, the problems in each of the two 15-problem examinations are arranged roughly from easiest to hardest, so the problem number itself gives a relative difficulty order. We use this order in place of the historical order for *Itemwise + Order*, treating items with a smaller problem number as easier. As shown in the final row of Table 3, this order reduces MAE by 0.22 percentage points without any historical evaluations, comparable to a perturbed historical order with a similar correlation. Our methods can therefore be applied even to a benchmark with no evaluation history, using only its metadata.

5.4 Estimation for model ranking

We rank the three strongest AIME 2025 target models using the relative difficulty order from the seven weaker models. Figure 1(b) shows that *Itemwise + Order*, denoted as Ours, reaches a 95% probability of recovering the correct ranking with approximately 17% fewer generations than *Uniform*.

6 Conclusion

We studied how to allocate a generation budget across benchmark items to estimate the accuracy of a language model on a reasoning benchmark. We developed adaptive sampling policies that use intermediate target responses and difficulty strata or a relative difficulty order from historical evaluations. In our experiments, using difficulty strata reduces estimation error relative to *Uniform*, and using a relative difficulty order reduces it further. These gains remain when only a single reference model is available, and the order-based policies achieve lower error than existing methods that rely on historical evaluations. Although the benefit decreases as the difficulty order becomes less reliable, even an order from benchmark metadata, without any historical evaluations, improves over *Uniform*. Whether our sampling policies can be combined with existing methods that correct estimates using historical evaluations, particularly when richer historical information is available, remains an open question for future work. We hope that our policies and the guidelines in Figure 2 help make the evaluation of reasoning models more efficient and reliable, even with limited historical information.

AI USE STATEMENT

We used generative AI tools to assist with identifying relevant prior work, implementing experiments, visualizing experimental results, improving the clarity and grammar of our writing, and checking mathematical proofs. We have reviewed all AI-assisted work mentioned above. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The experimental setup and evaluation protocols are described in the main paper and appendix. Complete proofs are provided in the appendix. Code for reproducing the experiments is available at `https://anonymous.4open.science/r/efficient-eval`.

REFERENCES

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, 1972. ISBN 0471049700.

P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

Paul N. Bennett and Vitor R. Carvalho. Online stratified sampling: Evaluating classifiers at web-scale. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 1581–1584. Association for Computing Machinery, 2010.

Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Thomas Rainforth. Scaling up active testing to large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Robert E Blackwell, Jon Barry, and Anthony G Cohn. Towards reproducible llm evaluation: Quantifying uncertainty in llm benchmark scores. *arXiv preprint arXiv:2410.03492*, 2024.

Sam Bowyer, Acyr Locatelli, and Kris Cao. Efficient benchmarking is just feature selection and multiple regression. *arXiv preprint arXiv:2605.25773*, 2026.

A. Carpentier, R. Munos, and A. Antos. Adaptive strategy for stratified Monte Carlo sampling. *Journal of Machine Learning Research*, 16(68):2231–2271, 2015.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977. ISBN 978-0-471-16240-7.

J. Dekoninck, N. Jovanović, T. Gehrunger, K. Rögnvaldsson, I. Petrov, C. Sun, and M. Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. Easy2Hard-Bench: Standardized difficulty labels for profiling LLM performance and generalization. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024.

Pierre Étoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12(3):335–360, 2010.

R. Fogliato, P. Patil, M. Monfort, and P. Perona. A framework for efficient model evaluation through stratification, sampling, and estimation. In *European Conference on Computer Vision (ECCV)*, 2024.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling*, 2025.

J. Kossen, S. Farquhar, Y. Gal, and T. Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Anurag Kumar and Bhiksha Raj. Classifier risk estimation under limited labeling resources. In Dinh Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi (eds.), *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Proceedings, Part I*, volume 10937 of *Lecture Notes in Computer Science*, pp. 3–15, Cham, 2018. Springer.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-Judge evaluations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

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

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tanmay Rajore, Nishanth Chandran, Sunayana Sitaram, Divya Gupta, Rahul Sharma, Kashish Mittal, and Manohar Swaminathan. TRUCE: Private benchmarking to prevent contamination and improve comparative evaluation of LLMs. *arXiv preprint arXiv:2403.00393*, 2024.

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. *arXiv preprint arXiv:2311.12022*, 2023.

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988. ISBN 0471917877.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. LLM-as-a-Judge on a budget. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4375–4383. PMLR, 2026.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601. Association for Computational Linguistics, 2024.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A Extended Related Work

Benchmark prediction from historical evaluations. Anchor Points (Vivek et al., 2024) and tinyBenchmarks (Maia Polo et al., 2024) use item-level information across reference models to reduce benchmark evaluation cost, with tinyBenchmarks including methods based on item response theory. Bowyer et al. (2026) further combine subset selection with regression. Notably, Zhang et al. (2025) show that benchmark prediction becomes less effective for target models with higher accuracy than previously evaluated models and develop an estimator based on augmented inverse propensity weighting. These methods use richer numerical historical information than the strata or order.

Estimation under order restrictions. Isotonic regression estimates parameters under a prescribed order constraint (Ayer et al., 1955; Barlow et al., 1972; Robertson et al., 1988). We use it to regularize estimates of target-model item-level accuracy for allocation and final accuracy estimation.

***Stratified* sampling and adaptive evaluation.** Classical stratified sampling combines stratum means with population weights. The optimal allocation minimizes the estimator variance when stratum variances are known (Cochran, 1977). Adaptive stratified sampling instead uses accumulated observations to guide subsequent allocation (Étoré & Jourdain, 2010; Carpentier et al., 2015). Stratified sampling has also been applied to classifier evaluation under limited labeling budgets (Bennett & Carvalho, 2010; Kumar & Raj, 2018), while Fogliato et al. (2024) jointly study stratification, sampling, and estimation for model evaluation. Active testing also uses adaptive example selection to reduce evaluation cost (Kossen et al., 2021; Berrada et al., 2025). Our adaptive stratum policies pool target responses within historical strata and select evaluations using estimated uncertainty.

B Evaluation Data

B.1 AIME 2025 model set

Our main experiments use the open-weight checkpoints listed in Table 4, with 32 stored responses for each of the 30 AIME 2025 problems. Five runs come from `pmahdavi/aime2025-merging-leaderboard`; we collect the other five. For our evaluations, we use vLLM (Kwon et al., 2023), temperature 0.6, and top-p 0.95. We include only developer-released checkpoints with complete response coverage.

Model	Source	Acc. (%)
`deepseek-ai/DeepSeek-R1-0528-Qwen3-8B`	ours	74.6
`allenai/Olmo-3-7B-Think`	ours	67.7
`Qwen/Qwen3-4B`	ours	63.5
`microsoft/Phi-4-mini-reasoning`	ours	34.7
`deepseek-ai/DeepSeek-R1-Distill-Llama-8B`	ours	31.4
`allenai/Olmo-3.1-7B-RL-Zero-Math`	public	38.0
`allenai/Olmo-3.1-7B-RL-Zero-Code`	public	33.0
`allenai/Olmo-3-7B-RL-Zero-Math`	public	31.4
`allenai/Olmo-3-7B-RL-Zero-Code`	public	26.2
`allenai/Olmo-3-1025-7B`	public	22.3

Table 4: AIME 2025 checkpoints and accuracy.

B.2 Benchmark coverage and preprocessing

Table 5 summarizes the four datasets. We retain every problem in each benchmark and require complete item coverage for each retained model configuration.

Benchmark	Items n	Models	Trials/item
AIME 2025	30	10	32
HMMT Feb. 2025	30	5	32
arXivMath	40	11	8 or 16
GPQA Diamond	198	4	128

Table 5: Evaluation datasets.

HMMT February 2025. We use the same five checkpoints marked "ours" in Table 4, with 32 responses per problem.

arXivMath. We combine the 17 December 2025 and 23 January 2026 MathArena problems, retaining the 11 model configurations present in both releases (Dekoninck et al., 2026). January runs have eight responses per item; December runs have eight or 16. We weight all 40 items equally.

GPQA Diamond. We use GPQA Diamond (Rein et al., 2023), with 198 questions and 128 responses per question from four public Hugging Face datasets by `drproduck`: `r1-qwen7b`, `r1-qwen14b`, `qwen3-8b`, and `qwen3-14b`. We retain all released binary scores, including parsing failures. Exact checkpoint revisions and full decoding settings are unavailable, so we retain the released model aliases.

B.3 Item-level accuracy profiles

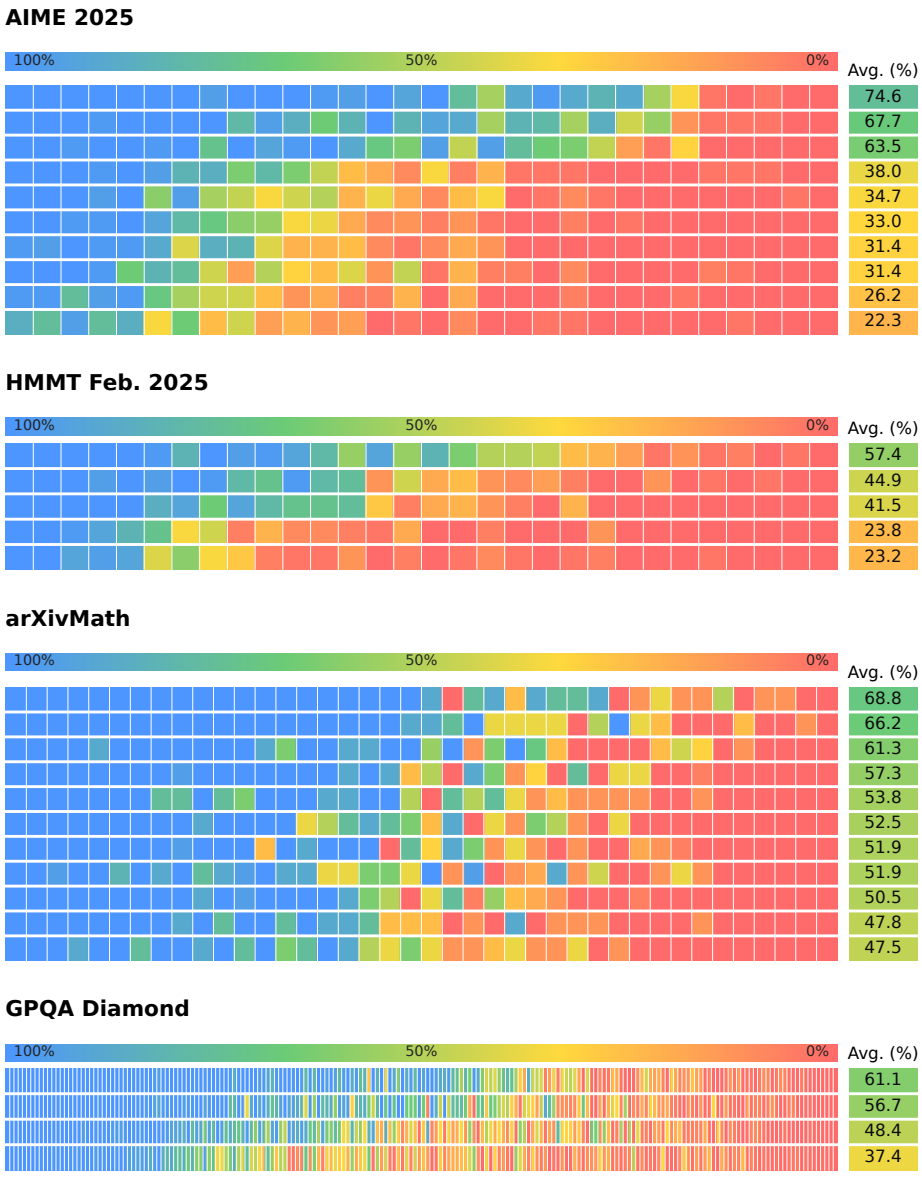


Figure 4: Empirical item-level accuracy across models and benchmarks.

C EXPERIMENTAL SETUP

Simulation and uncertainty. Responses are independent Bernoulli draws. Targets receive equal weight, and reported Monte Carlo standard errors are conditional on the empirical probabilities and reference pools.

Oracle. Oracle minimizes MSE over positive integer response counts assigned to items when $m \geq n$ and historical strata when $m < n$. In the latter case, each response is Bernoulli with the stratum's true mean probability. Final estimates use fresh responses under this fixed allocation. Oracle is not a universal lower bound on MAE or RMSE for adaptive policies.

Budgets. For Figure 3, we use $m/n \in \{1/2, 2/3, 1, 4/3, 3/2, 5/3\}$, rounded to integer budgets. Plotted ratios use the rounded budgets.

Ranking. We use 10,000 replicates at 30 budgets spanning $m = 30$–650 per target. Each replicate scores 1 for the correct strict ranking, $1/2$ for ties without inversions, and 0 otherwise. The 95% recovery thresholds are point estimates obtained by linear interpolation between adjacent budgets.

Order perturbations. For Table 3, independent Gaussian noise on historical ranks is calibrated to expected Kendall correlations of 0.8, 0.6, 0.4, and 0.2 with the original order. Correlations with target accuracies use Kendall τ_b to account for ties.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 FIXED-TOLERANCE COMPARISONS ACROSS BENCHMARKS

The tables report MAE and RMSE at fixed $\ell = 1, 2$, using the protocol of Figure 3 and Oracle from Appendix C.

Table 6: Percentage MAE and RMSE (with standard error) on AIME 2025.

m	m/n	*Uniform*		*Stratified*		*Stratified + Estimate*		*Stratified + Order*		*Itemwise + Order*		Oracle
		$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	
MAE												
15	1/2	7.84 (0.06)	9.07 (0.07)	6.14 (0.05)	6.11 (0.05)	6.10 (0.05)	6.18 (0.05)	5.73 (0.04)	5.84 (0.04)	5.66 (0.04)	**5.58** (0.04)	5.43 (0.04)
20	2/3	6.25 (0.05)	7.60 (0.06)	5.25 (0.04)	5.27 (0.04)	5.03 (0.04)	5.48 (0.04)	**4.81** (0.04)	4.96 (0.04)	4.87 (0.04)	**4.81** (0.04)	4.69 (0.04)
30	1	4.08 (0.03)	5.75 (0.04)	4.08 (0.03)	4.08 (0.03)	4.08 (0.03)	4.04 (0.03)	4.08 (0.03)	**3.85** (0.03)	4.08 (0.03)	3.91 (0.03)	4.08 (0.03)
40	4/3	3.70 (0.03)	4.47 (0.03)	3.63 (0.03)	3.57 (0.03)	3.34 (0.03)	3.41 (0.03)	3.30 (0.03)	**3.28** (0.02)	3.35 (0.03)	3.37 (0.03)	3.12 (0.02)
45	3/2	3.51 (0.03)	3.97 (0.03)	3.38 (0.03)	3.38 (0.03)	3.15 (0.02)	3.18 (0.02)	**3.10** (0.02)	**3.10** (0.02)	3.16 (0.02)	3.17 (0.02)	2.86 (0.02)
50	5/3	3.30 (0.02)	3.60 (0.03)	3.21 (0.02)	3.19 (0.02)	3.01 (0.02)	2.99 (0.02)	2.97 (0.02)	**2.93** (0.02)	3.04 (0.02)	3.01 (0.02)	2.64 (0.02)
RMSE												
15	1/2	9.84 (0.07)	11.30 (0.08)	7.68 (0.05)	7.70 (0.06)	7.65 (0.05)	7.76 (0.05)	7.24 (0.05)	7.36 (0.05)	7.18 (0.06)	**7.04** (0.05)	7.02 (0.06)
20	2/3	7.80 (0.05)	9.51 (0.07)	6.53 (0.05)	6.55 (0.05)	6.36 (0.05)	6.93 (0.05)	6.11 (0.05)	6.31 (0.05)	6.15 (0.05)	**6.08** (0.05)	6.05 (0.05)
30	1	5.11 (0.04)	7.17 (0.05)	5.11 (0.04)	5.11 (0.04)	5.11 (0.04)	5.18 (0.04)	5.11 (0.04)	**4.92** (0.04)	5.11 (0.04)	4.96 (0.04)	5.11 (0.04)
40	4/3	4.65 (0.03)	5.59 (0.04)	4.54 (0.03)	4.50 (0.03)	4.20 (0.03)	4.31 (0.03)	4.16 (0.03)	**4.14** (0.03)	4.21 (0.03)	4.25 (0.03)	3.93 (0.03)
45	3/2	4.40 (0.03)	4.99 (0.04)	4.25 (0.03)	4.26 (0.03)	3.95 (0.03)	4.01 (0.03)	3.91 (0.03)	**3.90** (0.03)	3.99 (0.03)	4.00 (0.03)	3.60 (0.03)
50	5/3	4.14 (0.03)	4.50 (0.03)	4.01 (0.03)	4.00 (0.03)	3.78 (0.03)	3.76 (0.03)	3.74 (0.03)	**3.69** (0.03)	3.82 (0.03)	3.78 (0.03)	3.33 (0.02)

Table 7: Percentage MAE and RMSE (with standard error) on HMMT Feb. 2025.

m	m/n	*Uniform*		*Stratified*		*Stratified + Estimate*		*Stratified + Order*		*Itemwise + Order*		Oracle
		$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	
MAE												
15	1/2	8.16 (0.08)	9.09 (0.10)	6.11 (0.06)	6.07 (0.06)	5.32 (0.06)	5.45 (0.06)	5.28 (0.06)	**5.18** (0.06)	5.36 (0.06)	5.33 (0.06)	5.30 (0.06)
20	2/3	6.32 (0.07)	7.56 (0.08)	5.04 (0.05)	4.98 (0.05)	4.64 (0.05)	4.86 (0.05)	4.63 (0.05)	**4.53** (0.05)	4.70 (0.05)	4.70 (0.05)	4.58 (0.05)
30	1	3.91 (0.04)	5.69 (0.06)	3.91 (0.04)	3.91 (0.04)	3.91 (0.04)	3.86 (0.04)	3.91 (0.04)	**3.75** (0.04)	3.91 (0.04)	**3.75** (0.04)	3.91 (0.04)
40	4/3	3.56 (0.04)	4.41 (0.05)	3.46 (0.04)	3.39 (0.04)	3.22 (0.03)	3.28 (0.04)	3.25 (0.03)	3.21 (0.04)	3.18 (0.03)	**3.16** (0.03)	2.98 (0.03)
45	3/2	3.36 (0.04)	3.94 (0.04)	3.29 (0.04)	3.24 (0.03)	3.06 (0.03)	3.08 (0.03)	3.04 (0.03)	3.03 (0.03)	3.06 (0.03)	**3.01** (0.03)	2.77 (0.03)
50	5/3	3.15 (0.03)	3.49 (0.04)	3.09 (0.03)	3.07 (0.03)	2.93 (0.03)	2.90 (0.03)	2.88 (0.03)	2.88 (0.03)	2.92 (0.03)	**2.87** (0.03)	2.54 (0.03)
RMSE												
15	1/2	10.05 (0.10)	11.41 (0.11)	7.50 (0.08)	7.47 (0.08)	6.85 (0.08)	7.03 (0.08)	6.67 (0.08)	**6.53** (0.07)	6.88 (0.08)	6.86 (0.08)	6.80 (0.08)
20	2/3	7.89 (0.08)	9.47 (0.09)	6.34 (0.07)	6.27 (0.06)	5.91 (0.07)	6.28 (0.07)	5.79 (0.06)	**5.66** (0.06)	5.94 (0.06)	6.02 (0.07)	5.86 (0.07)
30	1	4.94 (0.05)	7.17 (0.07)	4.94 (0.05)	4.94 (0.05)	4.94 (0.05)	4.91 (0.05)	4.94 (0.05)	**4.71** (0.05)	4.94 (0.05)	4.81 (0.05)	4.94 (0.05)
40	4/3	4.49 (0.04)	5.55 (0.06)	4.36 (0.04)	4.29 (0.04)	4.07 (0.04)	4.17 (0.04)	4.10 (0.04)	4.09 (0.04)	4.05 (0.04)	**4.02** (0.04)	3.80 (0.04)
45	3/2	4.24 (0.04)	4.95 (0.05)	4.13 (0.04)	4.04 (0.04)	3.89 (0.04)	3.92 (0.04)	3.86 (0.04)	3.85 (0.04)	3.88 (0.04)	**3.82** (0.04)	3.50 (0.04)
50	5/3	3.99 (0.04)	4.41 (0.04)	3.88 (0.04)	3.84 (0.04)	3.71 (0.04)	3.67 (0.04)	3.66 (0.04)	3.65 (0.04)	3.71 (0.04)	**3.64** (0.04)	3.25 (0.03)

Table 8: Percentage MAE and RMSE (with standard error) on arXivMath.

m	m/n	*Uniform*		*Stratified*		*Stratified + Estimate*		*Stratified + Order*		*Itemwise + Order*		Oracle
		$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	
MAE												
20	1/2	7.03 (0.05)	8.09 (0.06)	5.28 (0.04)	5.26 (0.04)	5.31 (0.04)	5.50 (0.04)	**4.72** (0.04)	4.94 (0.04)	4.86 (0.04)	5.10 (0.04)	4.82 (0.04)
27	$\approx 2/3$	5.37 (0.04)	6.63 (0.05)	4.26 (0.03)	4.25 (0.03)	4.07 (0.03)	4.75 (0.04)	**3.81** (0.03)	4.30 (0.03)	3.97 (0.03)	4.36 (0.03)	4.15 (0.03)
40	1	**3.14** (0.02)	4.95 (0.04)	**3.14** (0.02)	**3.14** (0.02)	**3.14** (0.02)	3.57 (0.03)	**3.14** (0.02)	3.25 (0.02)	**3.14** (0.02)	3.25 (0.02)	3.14 (0.02)
53	$\approx 4/3$	2.85 (0.02)	3.87 (0.03)	2.89 (0.02)	2.89 (0.02)	2.65 (0.02)	2.70 (0.02)	2.63 (0.02)	2.62 (0.02)	**2.58** (0.02)	2.60 (0.02)	2.31 (0.02)
60	3/2	2.72 (0.02)	3.38 (0.02)	2.72 (0.02)	2.73 (0.02)	2.48 (0.02)	2.48 (0.02)	2.48 (0.02)	2.45 (0.02)	2.44 (0.02)	**2.43** (0.02)	2.05 (0.02)
67	$\approx 5/3$	2.54 (0.02)	2.91 (0.02)	2.54 (0.02)	2.53 (0.02)	2.35 (0.02)	2.36 (0.02)	2.35 (0.02)	2.35 (0.02)	2.33 (0.02)	**2.30** (0.02)	1.87 (0.01)
RMSE												
20	1/2	8.77 (0.06)	10.17 (0.07)	6.55 (0.04)	6.52 (0.04)	6.82 (0.05)	7.23 (0.05)	**6.07** (0.04)	6.45 (0.05)	6.13 (0.04)	6.48 (0.05)	6.18 (0.05)
27	$\approx 2/3$	6.77 (0.05)	8.32 (0.06)	5.38 (0.04)	5.35 (0.04)	5.21 (0.04)	6.28 (0.05)	**4.85** (0.03)	5.59 (0.04)	5.04 (0.04)	5.57 (0.04)	5.29 (0.04)
40	1	**3.97** (0.03)	6.22 (0.04)	**3.97** (0.03)	**3.97** (0.03)	**3.97** (0.03)	4.59 (0.04)	**3.97** (0.03)	4.13 (0.03)	**3.97** (0.03)	4.17 (0.03)	3.97 (0.03)
53	$\approx 4/3$	3.61 (0.02)	4.84 (0.03)	3.63 (0.02)	3.63 (0.02)	3.34 (0.02)	3.44 (0.03)	3.31 (0.02)	3.33 (0.02)	**3.27** (0.02)	3.30 (0.02)	2.95 (0.02)
60	3/2	3.43 (0.02)	4.24 (0.03)	3.42 (0.02)	3.41 (0.02)	3.12 (0.02)	3.15 (0.02)	3.11 (0.02)	3.11 (0.02)	3.09 (0.02)	**3.07** (0.02)	2.61 (0.02)
67	$\approx 5/3$	3.22 (0.02)	3.66 (0.03)	3.19 (0.02)	3.18 (0.02)	2.97 (0.02)	2.96 (0.02)	2.97 (0.02)	2.94 (0.02)	2.96 (0.02)	**2.90** (0.02)	2.38 (0.02)

Table 9: Percentage MAE and RMSE (with standard error) on GPQA Diamond.

m	m/n	*Uniform*		*Stratified*		*Stratified + Estimate*		*Stratified + Order*		*Itemwise + Order*		Oracle
		$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	$\ell=1$	$\ell=2$	
MAE												
99	1/2	3.29 (0.04)	3.67 (0.04)	2.68 (0.03)	**2.66** (0.03)	3.24 (0.04)	3.81 (0.04)	2.68 (0.03)	3.09 (0.04)	2.75 (0.03)	3.19 (0.04)	2.61 (0.03)
132	2/3	2.59 (0.03)	3.12 (0.04)	2.25 (0.03)	2.25 (0.03)	2.51 (0.03)	3.41 (0.04)	**2.19** (0.03)	2.69 (0.03)	2.23 (0.03)	2.87 (0.04)	2.26 (0.03)
198	1	**1.70** (0.02)	2.35 (0.03)	**1.70** (0.02)	1.71 (0.02)	**1.70** (0.02)	2.54 (0.03)	**1.70** (0.02)	2.03 (0.02)	**1.70** (0.02)	2.19 (0.03)	1.70 (0.02)
264	4/3	1.55 (0.02)	1.87 (0.02)	1.52 (0.02)	1.51 (0.02)	1.47 (0.02)	1.84 (0.02)	1.45 (0.02)	1.56 (0.02)	**1.44** (0.02)	1.66 (0.02)	1.34 (0.02)
297	3/2	1.48 (0.02)	1.65 (0.02)	1.44 (0.02)	1.42 (0.02)	1.41 (0.02)	1.56 (0.02)	1.36 (0.02)	1.38 (0.02)	**1.35** (0.02)	1.42 (0.02)	1.24 (0.01)
330	5/3	1.39 (0.02)	1.50 (0.02)	1.35 (0.02)	1.34 (0.02)	1.33 (0.02)	1.36 (0.02)	1.28 (0.02)	1.28 (0.02)	**1.27** (0.02)	1.28 (0.02)	1.15 (0.01)
RMSE												
99	1/2	4.13 (0.05)	4.61 (0.05)	3.37 (0.04)	**3.33** (0.04)	4.09 (0.04)	4.79 (0.05)	3.37 (0.04)	3.88 (0.04)	3.48 (0.04)	4.01 (0.05)	3.33 (0.04)
132	2/3	3.24 (0.04)	3.90 (0.04)	2.82 (0.03)	2.83 (0.03)	3.15 (0.03)	4.33 (0.05)	**2.76** (0.03)	3.39 (0.04)	2.82 (0.03)	3.63 (0.04)	2.87 (0.04)
198	1	**2.13** (0.02)	2.96 (0.03)	**2.13** (0.02)	2.15 (0.02)	**2.13** (0.02)	3.21 (0.03)	**2.13** (0.02)	2.56 (0.03)	**2.13** (0.02)	2.77 (0.03)	2.13 (0.02)
264	4/3	1.95 (0.02)	2.37 (0.03)	1.91 (0.02)	1.90 (0.02)	1.85 (0.02)	2.31 (0.03)	1.83 (0.02)	1.97 (0.02)	**1.82** (0.02)	2.09 (0.02)	1.69 (0.02)
297	3/2	1.86 (0.02)	2.07 (0.02)	1.80 (0.02)	1.79 (0.02)	1.77 (0.02)	1.95 (0.02)	1.71 (0.02)	1.74 (0.02)	**1.70** (0.02)	1.79 (0.02)	1.56 (0.02)
330	5/3	1.74 (0.02)	1.88 (0.02)	1.69 (0.02)	1.69 (0.02)	1.67 (0.02)	1.70 (0.02)	1.61 (0.02)	1.61 (0.02)	**1.60** (0.02)	1.61 (0.02)	1.45 (0.02)

D.2 Limited historical information

In the one-reference condition of Table 2, a non-target reference is sampled independently for each target and batch.

Treating repeated responses as separate reference rows. For each reference model r and item i, we independently permute a binary column containing $32\theta_{ri}$ ones and $32(1-\theta_{ri})$ zeros. Treating

each row as a separate reference gives 32, 288, and 96 rows in the one-, nine-, and three-lowest-reference conditions of Table 2. IRT assigns each row a distinct subject identifier. All other settings are unchanged.

Table 10 averages three independent rearrangements, each with 1,000 paired target-response replicates per target and condition. These synthetic rows preserve reference means, orders, and strata.

Table 10: Percentage MAE (with standard error) on AIME 2025 with binary references.

References	m	m/n	*Uniform*	*Stratified + Order*	*Itemwise + Order*	AP-W	AP-P	P-IRT	GP-IRT	AIPW
1 reference	15	0.50	7.98 (0.06)	6.13 (0.05)	**6.10** (0.05)	25.11 (0.08)	22.50 (0.02)	9.28 (0.04)	7.60 (0.04)	6.41 (0.05)
	50	1.67	3.30 (0.03)	**3.02** (0.02)	3.05 (0.02)	———	Identical to *Uniform* (when $m > n$)			———
9 references	15	0.50	7.88 (0.06)	5.78 (0.04)	**5.64** (0.04)	11.18 (0.09)	16.04 (0.09)	8.60 (0.04)	7.53 (0.03)	5.84 (0.05)
	20	0.67	6.36 (0.05)	**4.79** (0.04)	4.84 (0.04)	8.18 (0.06)	18.68 (0.09)	6.73 (0.03)	6.30 (0.03)	5.03 (0.04)
	50	1.67	3.34 (0.03)	**3.02** (0.02)	3.08 (0.02)	———	Identical to *Uniform* (when $m > n$)			———
3 lowest	15	0.50	7.91 (0.11)	**6.81** (0.09)	6.94 (0.10)	18.32 (0.10)	51.03 (0.04)	8.20 (0.07)	7.55 (0.07)	7.77 (0.11)
	20	0.67	6.29 (0.09)	**5.46** (0.07)	5.61 (0.08)	18.06 (0.08)	61.12 (0.04)	6.41 (0.06)	6.22 (0.06)	6.12 (0.08)
	50	1.67	3.28 (0.04)	3.04 (0.04)	**3.01** (0.04)	———	Identical to *Uniform* (when $m > n$)			———

GPQA Diamond. We evaluate the strongest GPQA target, qwen3-14b (61.13%), using r1-qwen14b (48.41%) and r1-qwen7b (37.37%), with $h = 20$ and $\ell = 1$. Table 11 reports this single-target comparison.

Table 11: GPQA Diamond mean absolute error with two references.

m	m/n	Uniform	Stratified	Stratified + Estimate	Stratified + Order	Itemwise + Order	AP-W	AP-P	P-IRT	GP-IRT	AIPW
66	0.33	4.18 (0.10)	3.51 (0.08)	4.32 (0.10)	**3.49** (0.09)	3.55 (0.09)	25.40 (0.39)	42.89 (1.36)	5.00 (0.10)	5.44 (0.12)	3.53 (0.08)
99	0.50	3.10 (0.07)	**2.62** (0.06)	3.50 (0.08)	2.68 (0.06)	2.80 (0.07)	26.13 (0.60)	15.77 (0.42)	4.30 (0.09)	5.38 (0.10)	2.65 (0.06)
132	0.67	2.47 (0.06)	2.11 (0.05)	2.35 (0.06)	**1.92** (0.05)	2.08 (0.05)	26.89 (0.64)	17.29 (0.45)	2.61 (0.06)	4.06 (0.08)	2.15 (0.05)

E PROOFS

Under the assumptions of Section 4, all MSEs in (3) are conditional on the preliminary responses.

Proposition 1 (General form)**.** *For arbitrary stratum sizes, let* $p_k := |I_k|^{-1} \sum_{i \in I_k} \theta_i$, $\sigma_k := \sqrt{p_k(1-p_k)}$, *and* $\pi_k^\star := |I_k|\sigma_k / \sum_{r\in[h]} |I_r|\sigma_r$. *Assume* $p_k \in (0,1)$, $\widehat{\pi}_k > 0$, *and that each item in* I_k *receives the positive integer count* $m\widehat{\pi}_k/|I_k|$. *Then we have*

$$\begin{aligned}
&\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}_{\mathrm{strat}}) \\
&\geq \frac{1}{mn}\Big[\sum_{k\in[h]} |I_k|\big(\sigma_k - \tfrac{1}{n}\sum_{r\in[h]} |I_r|\sigma_r\big)^2 - \sum_{k\in[h]}\sum_{i\in I_k}(\theta_i - p_k)^2 - \tfrac{1}{n}\big(\sum_{k\in[h]} |I_k|\sigma_k\big)^2 \sum_{k\in[h]} \frac{(\widehat{\pi}_k - \pi_k^\star)^2}{\widehat{\pi}_k}\Big].
\end{aligned} \tag{13}$$

Proof. Let $c_i = \sum_{j\in[m_i]} y_{i,j}$. Then $\mathbb{E}[c_i] = m_i\theta_i$ and $\mathrm{Var}(c_i) = m_i\theta_i(1-\theta_i)$. Equal allocation within each stratum gives

$$\mathbb{E}\left[\frac{\sum_{i\in I_k} c_i}{m\widehat{\pi}_k}\right] = \frac{1}{m\widehat{\pi}_k}\sum_{i\in I_k}\frac{m\widehat{\pi}_k}{|I_k|}\theta_i = p_k.$$

Since $\sum_{k\in[h]} |I_k|p_k/n = \theta$, the estimator $\widehat{\theta}_{\mathrm{strat}}$ is unbiased. By independence of the responses within each stratum,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i\in I_k} c_i\right) &= \frac{m\widehat{\pi}_k}{|I_k|}\sum_{i\in I_k}\theta_i(1-\theta_i) \\
&= m\widehat{\pi}_k\left[p_k(1-p_k) - \frac{1}{|I_k|}\sum_{i\in I_k}(\theta_i - p_k)^2\right] \\
&\leq m\widehat{\pi}_k\, p_k(1-p_k).
\end{aligned}$$

Independence across strata gives

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{strat}}) \le \frac{1}{n^2} \sum_{k\in[h]} \frac{|I_k|^2 p_k(1-p_k)}{m\widehat{\pi}_k}. \tag{14}$$

The oracle stratum proportions, obtained by approximating each θ_i in I_k by p_k, are

$$\pi_k^\star = \frac{|I_k|\sigma_k}{\sum_{r\in[h]} |I_r|\sigma_r}. \tag{15}$$

Substituting the oracle proportions in (15) into the bound in (14), we obtain

$$\begin{aligned}
\frac{1}{n^2} \sum_{k\in[h]} \frac{|I_k|^2 p_k(1-p_k)}{m\widehat{\pi}_k} &= \frac{1}{m} \left(\tfrac{1}{n} \sum_{r\in[h]} |I_r|\sigma_r \right)^2 \sum_{k\in[h]} \frac{(\pi_k^\star)^2}{\widehat{\pi}_k} \\
&= \frac{1}{m} \left(\tfrac{1}{n} \sum_{r\in[h]} |I_r|\sigma_r \right)^2 \left[1 + \sum_{k\in[h]} \frac{(\widehat{\pi}_k - \pi_k^\star)^2}{\widehat{\pi}_k} \right],
\end{aligned}$$

where the last equality follows from $\sum_{k\in[h]} \widehat{\pi}_k = \sum_{k\in[h]} \pi_k^\star = 1$.

For the comparison with *Uniform*, let m_i now denote its random evaluation counts and $O_m := \{i \mid m_i > 0\}$ its nonempty evaluated-item set. The rule in Section 3.1 selects items independently of their responses and treats all items symmetrically. Conditional on $|O_m|$, each item therefore has probability $|O_m|/n$ of being observed, so

$$\mathbb{E}[\widehat{\theta}_{\mathrm{unif}} \mid |O_m|] = \frac{1}{|O_m|} \sum_{i\in[n]} \frac{|O_m|}{n} \theta_i = \theta.$$

Cauchy–Schwarz gives

$$|O_m|^2 \le \left(\sum_{i\in O_m} \frac{1}{m_i} \right) \left(\sum_{i\in O_m} m_i \right) = m \sum_{i\in O_m} \frac{1}{m_i}.$$

By the law of total variance,

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) &\ge \mathbb{E}\Big[\mathrm{Var}(\widehat{\theta}_{\mathrm{unif}} \mid m_1, \cdots, m_n) \Big] \\
&= \mathbb{E}\left[\frac{1}{|O_m|^2} \sum_{i\in O_m} \frac{\theta_i(1-\theta_i)}{m_i} \right] \\
&= \frac{1}{n} \sum_{i\in[n]} \theta_i(1-\theta_i)\, \mathbb{E}\left[\frac{1}{|O_m|^2} \sum_{j\in O_m} \frac{1}{m_j} \right] \\
&\ge \frac{1}{mn} \sum_{i\in[n]} \theta_i(1-\theta_i) \\
&= \frac{1}{m} \left[\tfrac{1}{n} \sum_{k\in[h]} |I_k| p_k(1-p_k) - \tfrac{1}{n} \sum_{k\in[h]} \sum_{i\in I_k} (\theta_i - p_k)^2 \right].
\end{aligned}$$

Combining this lower bound with the expansion of (14), the MSE difference is bounded by

$$\begin{aligned}
\frac{1}{m} \Bigg[\tfrac{1}{n} \sum_{k\in[h]} |I_k| p_k(1-p_k) - \left(\tfrac{1}{n} \sum_{r\in[h]} |I_r|\sigma_r \right)^2 - \tfrac{1}{n} \sum_{k\in[h]} \sum_{i\in I_k} (\theta_i - p_k)^2 \\
- \left(\tfrac{1}{n} \sum_{r\in[h]} |I_r|\sigma_r \right)^2 \sum_{k\in[h]} \frac{(\widehat{\pi}_k - \pi_k^\star)^2}{\widehat{\pi}_k} \Bigg].
\end{aligned}$$

Finally,

$$\frac{1}{n}\sum_{k\in[h]}|I_k|p_k(1-p_k)-\left(\frac{1}{n}\sum_{r\in[h]}|I_r|\sigma_r\right)^2=\frac{1}{n}\sum_{k\in[h]}|I_k|\left(\sigma_k-\frac{1}{n}\sum_{r\in[h]}|I_r|\sigma_r\right)^2.$$

For $|I_k|=n/h$, substituting $\bar{\sigma}:=\frac{1}{h}\sum_{k\in[h]}\sigma_k$ and $\pi_k^\star=\sigma_k/(h\bar{\sigma})$ gives (9). □

For Lemma 1 and Proposition 2, m_i denotes the number of preliminary responses to item i.

Lemma 1. *Assume $m_i>0$ for all $i\in[n]$ and $2\delta\le\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}$. Then $\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2\le\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2$.*

Proof. For any $u_1\ge\cdots\ge u_n$ and $t\in(0,1]$, the vector $\widetilde{\boldsymbol{\theta}}+t(\boldsymbol{u}-\widetilde{\boldsymbol{\theta}})$ is also nonincreasing. By the definition of $\widetilde{\boldsymbol{\theta}}$,

$$0\le\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}-t(\boldsymbol{u}-\widetilde{\boldsymbol{\theta}})\|_{\boldsymbol{m}}^2-\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}^2.$$

From Bellec (2018), we have

$$\sum_{i\in[n]}m_i(\widehat{\theta}_i-\widetilde{\theta}_i)(u_i-\widetilde{\theta}_i)\le 0. \tag{16}$$

Expanding the squared estimation error gives

$$\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2=\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2-\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}^2+2\sum_{i\in[n]}m_i(\widehat{\theta}_i-\widetilde{\theta}_i)(\theta_i-\widetilde{\theta}_i).$$

Let $\boldsymbol{u}$ attain the minimum defining δ. Substituting $\theta_i-\widetilde{\theta}_i=(\theta_i-u_i)+(u_i-\widetilde{\theta}_i)$ into the cross term and applying (16) and the Cauchy–Schwarz inequality, we obtain

$$\sum_{i\in[n]}m_i(\widehat{\theta}_i-\widetilde{\theta}_i)(\theta_i-\widetilde{\theta}_i)\le\delta\,\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}.$$

Therefore

$$\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2\le\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2-\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}^2+2\delta\,\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}. \tag{17}$$

Under the condition $2\delta\le\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}$, the last two terms sum to a nonpositive quantity. □

Proposition 2. *Assume $m_i>0$ and $\widehat{\theta}_i\in(0,1)$ for all $i\in[n]$ and $2\delta\le\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}$. Then we have*

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{iso}})-\mathrm{MSE}(\widehat{\theta}^\star)\le\underbrace{c\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_{\boldsymbol{m}}^2}_{\text{upper bound on }\mathrm{MSE}(\widehat{\theta}_{\mathrm{raw}})-\mathrm{MSE}(\widehat{\theta}^\star)}-c\big[\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}^2-2\delta\,\|\widehat{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{m}}\big], \tag{10}$$

where $\widehat{\theta}^\star$ uses the oracle allocation and $c>0$ is a constant independent of $\widetilde{\boldsymbol{\theta}}$, given explicitly in the proof below.

Proof. For proportions $\pi_i>0$ summing to one, the estimator $\widehat{\theta}$ in (2), using $m\pi_i$ independent final responses to item i, is unbiased, and

$$\mathrm{MSE}(\widehat{\theta})=\frac{1}{n^2m}\sum_{i\in[n]}\frac{\theta_i(1-\theta_i)}{\pi_i}. \tag{18}$$

Let $\pi_i^\star:=m_i^\star/m$ and $\widehat{\theta}^\star$ denote the oracle proportions and estimator. Items with $\theta_i\in\{0,1\}$ contribute zero to (18), including when $\pi_i^\star=0$.

For either choice of priorities, $q_i=\sqrt{\widehat{\theta}_i(1-\widehat{\theta}_i)}$ or $q_i=\sqrt{\widetilde{\theta}_i(1-\widetilde{\theta}_i)}$, substitute $\pi_i=q_i/\sum_{j\in[n]}q_j$ into (18) to obtain

$$\mathrm{MSE}(\widehat{\theta})=\frac{1}{n^2m}\left(\sum_{i\in[n]}q_i\right)\left(\sum_{i\in[n]}\frac{\theta_i(1-\theta_i)}{q_i}\right). \tag{19}$$

Then

$$\left(\sum_{i\in[n]} q_i\right)\left(\sum_{i\in[n]} \frac{\theta_i(1-\theta_i)}{q_i}\right) - \left(\sum_{i\in[n]} \sqrt{\theta_i(1-\theta_i)}\right)^2$$
$$= \left(\sum_{i\in[n]} q_i\right)\sum_{i\in[n]} \frac{(q_i - \sqrt{\theta_i(1-\theta_i)})^2}{q_i} - \left(\sum_{i\in[n]} (q_i - \sqrt{\theta_i(1-\theta_i)})\right)^2.$$

Since $q_i \le 1/2$, omitting the final nonpositive term and using $q_i \ge \min_{j\in[n]} q_j$ gives

$$\mathrm{MSE}(\widehat{\theta}) \le \mathrm{MSE}(\widehat{\theta}^\star) + \frac{\sum_{i\in[n]}(q_i - \sqrt{\theta_i(1-\theta_i)})^2}{2nm\min_{i\in[n]} q_i}. \tag{20}$$

For the raw priorities, $|1 - \widehat{\theta}_i - \theta_i| \le 1$ gives

$$|q_i - \sqrt{\theta_i(1-\theta_i)}| = \frac{|\widehat{\theta}_i - \theta_i|\,|1-\widehat{\theta}_i-\theta_i|}{q_i + \sqrt{\theta_i(1-\theta_i)}} \le \frac{|\widehat{\theta}_i - \theta_i|}{q_i}.$$

Therefore

$$\sum_{i\in[n]} (q_i - \sqrt{\theta_i(1-\theta_i)})^2 \le \frac{1}{(\min_{i\in[n]} q_i)^2}\sum_{i\in[n]} (\widehat{\theta}_i - \theta_i)^2 \le \frac{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2}{(\min_{i\in[n]} q_i)^2 \min_{i\in[n]} m_i}.$$

Substituting into (20) gives

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{raw}}) - \mathrm{MSE}(\widehat{\theta}^\star) \le c\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2, \tag{21}$$

where

$$c = \frac{1}{2nm\left(\min_{i\in[n]}\sqrt{\widehat{\theta}_i(1-\widehat{\theta}_i)}\right)^3 \min_{i\in[n]} m_i} > 0.$$

The same argument applies to the isotonic estimates. Each fitted value is a weighted average of its raw estimates, and the function $\sqrt{x(1-x)}$ is concave in x. Therefore,

$$\min_{i\in[n]} \sqrt{\widetilde{\theta}_i(1-\widetilde{\theta}_i)} \ge \min_{i\in[n]} \sqrt{\widehat{\theta}_i(1-\widehat{\theta}_i)},$$

so the same constant c applies to both choices of priorities. Then

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{iso}}) - \mathrm{MSE}(\widehat{\theta}^\star) \le c\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_{\boldsymbol{m}}^2.$$

□

For Proposition 3, $\widehat{\theta}_{\mathrm{unif}}$ denotes the estimator under continuous equal allocation $m_i = m/n$.

Proposition 3. *Assume $\frac{1}{n}\sum_{i\in[n]} \sqrt{\theta_i(1-\theta_i)} \ge c$ for some $c > 0$. Then the oracle counts satisfy*

$$\max_{i\in[n]} m_i^\star - \min_{i\in[n]} m_i^\star \le \frac{m}{2cn}. \tag{11}$$

Moreover, there exists an allocation whose final counts differ by at most $m/(2n)$ and give every item at least $(1-c)m/n$ evaluations, and whose estimator $\widehat{\theta}_{\mathrm{mix}}$ satisfies

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}_{\mathrm{mix}}) \ge c\left[\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) - \mathrm{MSE}(\widehat{\theta}^\star)\right]. \tag{12}$$

Proof. By (4),

$$\max_{i\in[n]} m_i^\star - \min_{i\in[n]} m_i^\star = \frac{m}{n}\,\frac{\max_{i\in[n]}\sqrt{\theta_i(1-\theta_i)} - \min_{i\in[n]}\sqrt{\theta_i(1-\theta_i)}}{\frac{1}{n}\sum_{i\in[n]}\sqrt{\theta_i(1-\theta_i)}} \le \frac{m}{2cn},$$

using $0 \le \sqrt{\theta_i(1-\theta_i)} \le 1/2$. This proves (11) and also implies $c \le 1/2$.

Define the counts

$$m_i := (1-c)\frac{m}{n} + cm_i^\star, \tag{22}$$

and let $\widehat{\theta}_{\mathrm{mix}}$ be the estimator under this allocation. Since $0 < c \le 1/2$, these counts are positive and sum to m. Moreover,

$$\max_{i\in[n]} m_i - \min_{i\in[n]} m_i \le c\frac{m}{2cn} = \frac{m}{2n}.$$

Thus the final counts differ by at most $m/(2n)$, and $m_i \ge (1-c)m/n$ for every item.

The proportions satisfy $\boldsymbol{\pi} = (1-c)\frac{1}{n}\mathbf{1} + c\boldsymbol{\pi}^\star$. Since $\theta_i(1-\theta_i)/x$ is convex in $x > 0$, applying (18) gives

$$\mathrm{MSE}(\widehat{\theta}_{\mathrm{mix}}) \le (1-c)\,\mathrm{MSE}(\widehat{\theta}_{\mathrm{unif}}) + c\,\mathrm{MSE}(\widehat{\theta}^\star).$$

□