Efficient Prediction of pass@k Scaling in Large Language Models

Joshua Kazdan[1†], **Colin Sullivan**[1†*], **Rylan Schaeffer**[1‡], **Youssef Allouah**[1‡], **Kyssen Yu**[2], **Noam Levi**[3], **Sanmi Koyejo**[1]
[1]Stanford University [2]University of Toronto [3] EPFL
[†]Equally contributing first author. [‡] Equally contributing second author.

Abstract

Assessing the capabilities and risks of frontier AI systems is a critical area of research; recent work has shown that repeated sampling can dramatically increase both. For instance, repeated sampling allows models to solve difficult math and coding problems, but it may also increase their risk of being jailbroken. Such results raise a crucial question: how can one accurately predict a model's behavior when scaled to a massive number of attempts, given a vastly smaller sampling budget? This question is directly relevant to model providers, who serve hundreds of millions of users daily, and to governmental regulators, who seek to prevent harms. To answer this question, we make three contributions. First, we find that standard methods for fitting these laws suffer from statistical shortcomings that hinder predictions, especially in data-limited scenarios. Second, we remedy these shortcomings by introducing a robust estimation framework, which uses a beta-binomial distribution to generate more accurate predictions from limited data. Third, we propose a dynamic sampling strategy that allocates a greater budget to harder problems. Combined, these innovations enable more reliable prediction of rare risks and capabilities at a fraction of the computational cost.

1 Introduction

Prompt-based attacks against frontier (multimodal) AI systems often fail when attempted only once [2, 25, 12, 15]. Likewise, many hard math [9] and software engineering [14] tasks are too difficult for models to solve reliably on the first attempt. Through repeated attempts, however, the success rate of these models can climb rapidly to near-100% [3, 13, 17]. Consequently, predicting changes in capabilities and/or risks when a user is allowed many attempts to accomplish a task has become an important problem for companies, researchers, and governmental regulators alike. The relevance of this problem is only underscored by the massive scale at which these frontier AI systems are deployed, with some experiencing billions of daily interactions. However, making such predictions is challenging because sampling from language models at such scale can be prohibitively expensive. How can one predict the behavior of frontier AI systems in this repeated attempts regime using only a limited number of samples?

In this work, we approach this problem through estimation of the widely used $\mathrm{pass@}k$ metric [16, 7], which measures the expected pass rate given k attempts at solving each problem, where a problem is solved if any attempt is successful. Unfortunately, direct estimation at high k is often difficult. While prior work has shown that $\mathrm{pass@}k$ can follow predictable power laws across a range of domains including jailbreaking, mathematical problem-solving, and code generation [13, 3, 8], we find that

*Correspondence to: jkazdan@stanford.edu, cpsullivan@cs.stanford.edu, sanmi@cs.stanford.edu.

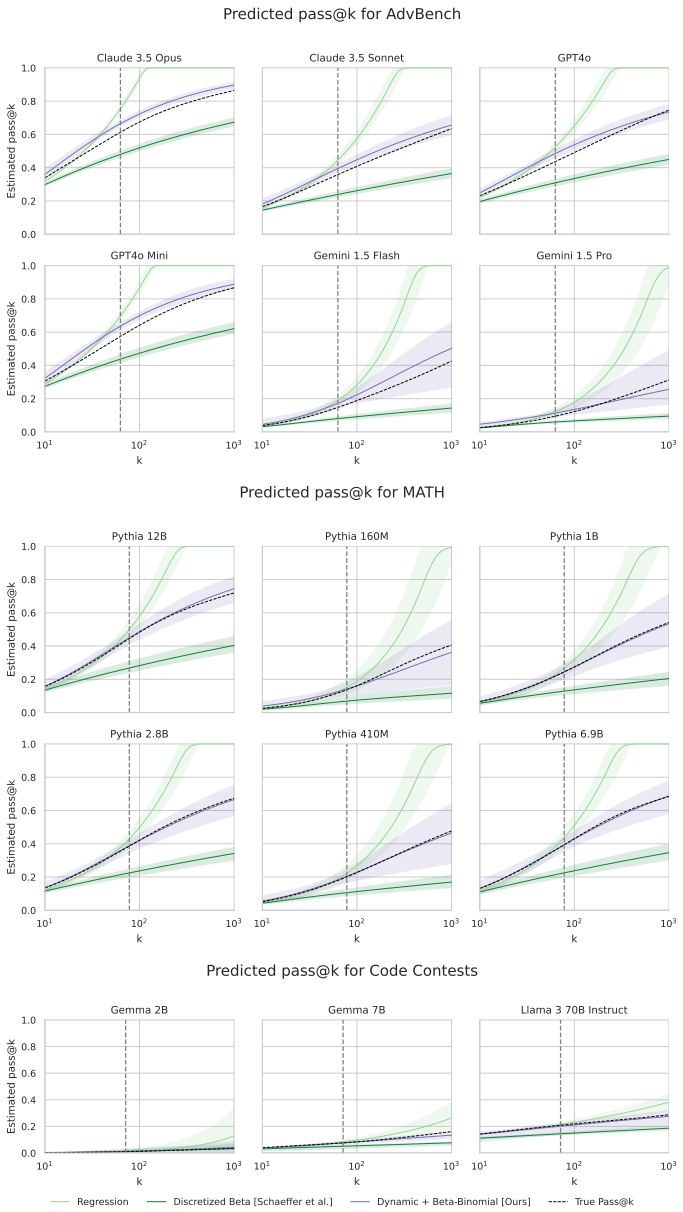

Figure 1: **Comparing Forecasting Methods for** $\text{pass}_{\mathcal{D}}@k$ **Across Different Datasets.** The ground truth is computed based on $10\,000$ actual samples per problem. All predictive models are trained on data from a budget of $10\,000$ total samples. **The gray region** shows k for which $\text{pass}@k$ can be directly estimated given the available budget, while **the white region** shows k for which the $\text{pass}@k$ must be extrapolated given the budget. Our estimator tracks the ground truth far better than prior methods. Error bars represent a bootstrap-averaged 95% percentile range of estimates.

standard methods for fitting these laws [7, 3, 13] suffer from statistical shortcomings that hinder predictive accuracy, especially in data-limited scenarios.

We argue that the shortcomings of prior prediction methods stem from statistical approximations that do not hold in sample-limited regimes. By carefully modeling the data-generating process and developing faithful estimators, we demonstrate that predictions can be substantially improved.

1.1 Contributions

To address the challenge of efficient prediction, this paper makes the following contributions:

1. **Rigorous critique of prior prediction methods.** We discuss statistical flaws that have led to poor prediction accuracy in approaches such as log-log-linear regression and existing distributional fitting techniques.
2. **Robust estimation framework for prediction.** We remedy the shortfalls of previous methods by employing a more suitable distributional model—the beta-binomial—and deriving an improved predictor for $\mathrm{pass@}k$ that faithfully accounts for the data generating process in order to deliver more accurate predictions.
3. **Efficient dynamic sampling strategy.** We show empirically that by allocating our fixed compute budget adaptively to focus on more difficult problems, we achieve more accurate predictions than the standard approach of uniform sampling.

The insights from this work are important for both AI safety and capabilities research. For AI safety, reliable forecasts for the scaling of vulnerability rates is crucial for assessing the societal risk posed by models deployed to millions of users. For capabilities, such predictions are vital for efficiently applying methods like Reinforcement Learning from Verified Rewards, where training on difficult problems requires correctly sizing batches to ensure non-zero success rates. Thus, efficient prediction for $\mathrm{pass@}k$ scaling represents a critical step toward developing aligned and powerful AI systems.

Code for reproducing our experiments is available on GitHub.[2]

2 Problem Statement: Efficient Prediction of Rare Model Behaviors from Repeated Sampling

We consider the performance of AI systems on some problem, defined as a set of prompts with verifiable binary outcomes: each attempt either produces the (un)desirable outcome for that prompt, or does not. For example, we may want our AI system to solve a Millennium Problem, or to not launch a cyberattack on a nation's infrastructure. Our goal is to predict the success rate of an AI system, given many repeated attempts at the problem. To quantitatively measure the system's behavior, we use the widely-adopted $\mathrm{pass@}k$ metric [16, 7]: For a single prompt, indexed by i, from a distribution of prompts $\mathcal{D}$, let $\mathrm{pass}_i@1$ be the model's true probability of success in one attempt. The probability of achieving at least one success in k attempts is then $\mathrm{pass}_i@k$:

$$\mathrm{pass}_i@k = 1 - (1 - \mathrm{pass}_i@1)^k. \tag{1}$$

For the entire dataset $\mathcal{D}$ of m problems, the overall pass rate $\mathrm{pass}_{\mathcal{D}}@k$ is the expected fraction of problems solved within k attempts:

$$\mathrm{pass}_{\mathcal{D}}@k = \mathbb{E}_{i\sim\mathcal{D}}[\mathrm{pass}_i@k]. \tag{2}$$

Our goal is to predict performance given many attempts using data from an economically feasible, small-scale experiment. This leads to our formal research question:

> *Given a compute budget of B samples to be distributed across a dataset $\mathcal{D}$ containing m problems, how should one allocate samples to build a model that predicts $\mathrm{pass}_{\mathcal{D}}@k$ for $k \gg B/m$?*

In this work, we use a small budget (e.g., $B/m \in [10^0, 10^2]$) to predict performance for $\mathrm{pass@}k$ at large scale (e.g., $k \in [10^1, 10^3]$). We evaluate predictions by comparing them to a ground truth estimate of $\mathrm{pass@}k$ computed using a withheld dataset of $10\,000$ samples per problem. To evaluate performance, we compute mean squared error (MSE) relative to the ground truth $\mathrm{pass@}k$ values.

The product of our contributions is an estimator that provides consistently more accurate predictions than existing methods (see Figure 1).

[2] `https://github.com/jlkazdan/monkeys_revisited`

3 Critiquing past methods of predicting $\mathrm{pass@}k$

We now examine past methods of predicting $\mathrm{pass@}k$ scaling and identify their shortcomings.

3.1 Combinatorial Estimation

Directly measuring $\text{pass}_{\mathcal{D}}@k$ for a large k is often computationally expensive. While unbiased estimators exist, such as that of [7], they are only defined when the number of samples taken for each problem is greater than or equal to the number of attempts k. Given b_i samples on problem i with s_i successes, this estimator is:

$$\widehat{\text{pass}_i@k} = 1 - \frac{\binom{b_i - s_i}{k}}{\binom{b_i}{k}}. \tag{3}$$

In this paper, we focus on the regime where $B/m < k < B$. As the size and quantity of benchmarks continues to grow, we may often find ourselves in such constrained contexts. Here, given that $k > B/m$, we cannot allocate the required minimum of k samples for each of m problems. This means the standard unbiased estimator (Equation 3) cannot be directly applied, so we must instead rely on extrapolation and predictive modeling.

3.2 Linear Regression

The first and most common extrapolation of $\mathrm{pass@}k$ uses linear regression [3, 13]. Specifically, given b samples per problem, one first estimates $\text{pass}_{\mathcal{D}}@k$ for k between 1 and b and then fits a least squares regression of the form:

$$-\log(\text{pass}_{\mathcal{D}}@k) \sim a\log(k) + c. \tag{4}$$

Fixing $C = e^{-c}$ corresponds to the power law:

$$\text{pass}_{\mathcal{D}}@k \sim C \cdot k^{-a}. \tag{5}$$

Explicitly, the regression loss takes the form:

$$\frac{1}{b}\sum_{k=1}^{b}\left(-\log\left(\widehat{\text{pass}_{\mathcal{D}}@k}\right) - a\log(k) - c\right)^2. \tag{6}$$

There are several problems with this approach, leading to poor estimates of $\mathrm{pass@}k$ for high k values (see Figure 1):

1. Estimates of $\text{pass}_{\mathcal{D}}@k$ are not independent for different k when computed using the same dataset of samples.
2. Estimates of $\text{pass}_{\mathcal{D}}@k$ are not homoskedastic, i.e. they have different variances for each value of k.
3. $\mathrm{pass@}k$ may not follow a power law for some datasets.
4. Power laws typically apply only for large values of k. Therefore, if the computation budget for sampling is limited, non-leading terms can dominate, resulting in poor fits of the data.

As a concrete example of the fourth point, suppose that

$$1 - \text{pass@}k = \frac{A}{k^{\alpha}} + \frac{B}{k^{\beta}} \tag{7}$$

where $A \gg B$ but $\alpha > \beta$. For small values of k, the first term of Equation 7 dominates. However, for large values of k, the second term, which supplies the true asymptotic power law, dominates. If we lack a sufficient budget to observe samples for large k, then least squares will incorrectly fit to the first term. We quantify statements 1 and 2 more precisely with proofs in Appendix D.

Our work directly remedies these issues by moving away from regression on aggregate statistics, instead modeling the underlying distribution of problem difficulties.

3.3 Discretized-Beta Distributional Fitting

[26] use a variant of empirical Bayes to estimate $\text{pass}@k$ for high k. To describe their method, we first introduce some notation. As before, let $\mathcal{D}$ denote a data set of questions. Define $\mathcal{U}$ to be the distribution of per-problem success probabilities $\text{pass}_i@1$ for $i \in \mathcal{D}$:

$$\text{pass}_i@1 \sim \mathcal{U}, \quad i \in \mathcal{D}. \tag{8}$$

For the i-th question in our dataset, we observe b samples, of which we say that s_i are successful. [26] fit scaled beta distributions to $\widehat{\text{pass}_i@1} = \frac{s_i}{b}$ and leveraged this distribution to estimate $\text{pass}@k$:

Step 1: Fit the scale θ. Recall the probability density function of a scaled beta distribution:

$$\text{Beta}(p; \alpha, \beta, \theta) = \frac{1}{\text{Be}(\alpha, \beta)} \left(\frac{p}{\theta}\right)^{\alpha-1} \left(1 - \frac{p}{\theta}\right)^{\beta-1} \frac{1}{\theta}, \tag{9}$$

[26] provide the following estimate for the scale parameter θ:

$$\hat{\theta} = \frac{b+1}{b} \text{max}_{i \in \mathcal{D}} \left(\widehat{\text{pass}_i@1}\right). \tag{10}$$

They use this estimator because it resembles the uniformly minimum variance unbiased estimator (UMVUE) for the parameter B of a uniform distribution $\text{Uniform}(0, B)$ [20]. Unfortunately, the scaled beta distribution is not an exponential family distribution. In particular, the UMVUE for θ in a scaled beta distribution is unknown. As such, this is not a principled estimator for θ. We provide details for how to estimate θ using a stabilized MLE in Appendix E, but we find empirically that using the scale parameter does not improve predictions.

Step 2: Fit α and β by discretizing. [26] first divide the interval $(0, 1)$ into log-scale bins with endpoints $0 = e_0, e_1, ..., e_\ell = 1$, where the bin widths decrease ($e_i - e_{i-1} > e_{i+1} - e_i$). They then numerically compute the probability mass in each bin and fit α and β by maximizing the multinomial likelihood over the number of problems whose estimated success rate falls into each bin. Specifically, if we assign the estimated probability:

$$A_i(\alpha, \beta, \theta) := \int_{e_i}^{e_{i+1}} \text{Beta}(p; \alpha, \beta, \theta) dp, \tag{11}$$

then [26] fit α and β by optimizing

$$\arg\min_{\alpha,\beta} \; -\log\left(\prod_{i=1}^{\ell} A_i(\alpha, \beta, \theta)^{\sum_{j=1}^{m} \mathbf{1}\{\widehat{\text{pass@1}} \in [e_i, e_{i+1})\}}\right) \tag{12}$$

$$= \arg\min_{\alpha,\beta} \; -\sum_{i=1}^{\ell} \left(\sum_{j=1}^{m} \mathbf{1}\{\widehat{\text{pass@1}} \in [e_i, e_{i+1})\}\right) \log\big(A_i(\alpha, \beta, \theta)\big) \tag{13}$$

This complex discretized beta estimator was used to support the common case when $s_i = 0$. The estimate $\widehat{\text{pass}_i@1}$ is also 0, so the scaled beta density is not supported.

Step 3: Predict $\text{pass}@k$ [26] use the fit distribution to approximate the asymptotic slope of the $\text{pass}@k$ scaling curve and do not attempt to extrapolate $\text{pass}@k$ beyond the provided number of trials. To extend this approach to the high-k regime, we take the expectation over the success probability $\text{pass}_i@1 \sim \text{Beta}(\hat{\alpha}, \hat{\beta}, \hat{\theta})$:

$$\widehat{\text{pass}_{\mathcal{D}}@k} = \mathbb{E}_{\text{pass}_i@1 \sim \text{Beta}(\hat{\alpha}, \hat{\beta}, \hat{\theta})} \left[1 - (1 - \text{pass}_i@1)^k\right]. \tag{14}$$

Analysis of the Discretized-Beta Estimator Because the bins are wider for smaller values, this fitting method consistently produces **downward-biased** estimates of the distribution $\mathcal{U}$. We demonstrate this phenomenon in Figure 2 where the discretized beta distribution is fit on problem success probabilities drawn from a uniform distribution. The fit is visibly skewed, incorrectly up-weighting the left tail of the distribution.

4 Better Estimation of pass@k

In this section, we develop a novel predictor of $\text{pass}_{\mathcal{D}}@k$, inspired by [21] that improves predictive accuracy for large k. As shown in Figure 5, our method provides equivalent or better estimates across all models, values of k, and sampling budgets tested. We no longer assume a fixed sampling budget per question: b_i denotes the budget for the i-th question. Our improvements involve two steps:

1. We develop an alternative distributional fitting method for the problem-difficulty distribution $\mathcal{U}$.
2. We propose a simple dynamic sampling strategy to allocate the sample budget more efficiently.

4.1 Fitting the Problem-Difficulty Distribution $\mathcal{U}$

Denote the underlying distribution of per-problem success probabilities as $\text{pass}_i@1 \sim \mathcal{U}$, where $\mathcal{U}$ is unknown. The number of successes s_i on the i-th problem out of b_i attempts is then binomially distributed: $s_i \sim \text{Binomial}(b_i, \text{pass}_i@1)$. Instead of the biased discretization approach, we model $\mathcal{U}$ as a beta distribution. This allows us to fit a beta-binomial distribution directly to the observed counts of successes and trials (s_i, b_i). The likelihood for the beta-binomial is given by:

$$\Pr\left[s = s_i \mid b = b_i; \alpha, \beta\right] = \binom{b_i}{s_i} \frac{\text{Be}(s_i + \alpha, b_i - s_i + \beta)}{\text{Be}(\alpha, \beta)}, \tag{15}$$

where $\text{Be}(\cdot, \cdot)$ is the beta function. As shown in Figure 2, the discretized estimator badly fits a uniform distribution because it incorrectly puts excessive weight on the left tail. Maximizing the beta-binomial likelihood leads to a superior fit which ultimately results in better predictions of $\text{pass}@k$.

Next, we obtain a maximum likelihood estimate for $\mathcal{U}$:

$$\hat{\alpha}, \hat{\beta} = \arg\max_{\alpha,\beta>0} \prod_{i=1}^{m} \Pr\left[s = s_i \mid b = b_i; \alpha, \beta\right]. \tag{16}$$

Finally, we retrieve an estimate for $\text{pass}@k$:

$$\widehat{\text{pass}_{\mathcal{D}}@k} = \mathbb{E}_{\text{pass}_i@1 \sim \text{Beta}(\hat{\alpha},\hat{\beta})}\left[1 - (1 - \text{pass}_i@1)^k\right]. \tag{17}$$

We see in Figure 2 that our approximate Beta-Bernoulli distribution better fits problem success probabilities sampled from a uniform distribution.

4.2 More Efficient Sampling Strategies

It was demonstrated by [26] that in the high-k regime, $\text{pass}@k$ scaling is governed almost exclusively by the shape of the difficulty distribution near 0. Distinguishing between an easy problem ($\text{pass}_i@1 = 0.25$) and a very easy problem ($\text{pass}_i@1 = 0.75$) provides little to no information. Therefore, we propose to concentrate our budget, sampling responses from the hardest problem with the least number of attempts so far (see Algorithm 2).

This adaptive approach is not directly applicable to the regression-based estimator, which requires uniform sample count across problems to compute intermediate $\text{pass}_{\mathcal{D}}@k$ values. It is also inconsistent with the discretized estimator from [26] since direct estimates $\widehat{\text{pass}_i@1} = \frac{s_i}{b_i}$ have different precision with this dynamic sampling method. However, our distributional fitting method remains valid, as the beta-binomial likelihood (Equation 15) can handle variable numbers of trials (b_i) for each problem. We outline our complete approach in Algorithm 1.

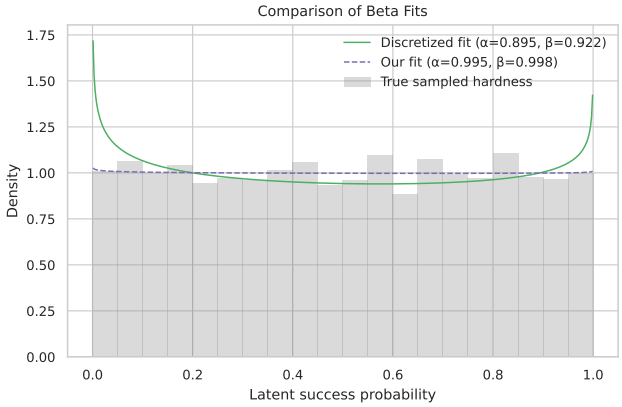

Figure 2: **Comparing Hardness Distribution Fit for Discretized Beta vs. Beta-Bernoulli**. $m = 10\,000$ problem success probabilities are sampled: $\text{pass}_i@1 \sim \text{Uniform}([0, 1])$. $b = 100$ success/failure samples are drawn for each problem, $s_i \sim \text{Bin}(b, \text{pass}_i@1)$.

Algorithm 1 Dynamic Sampling + Beta-Binomial Fit for Efficient $\text{pass}_{\mathcal{D}}@k$ Estimation

Require: Dataset $\mathcal{D}$ with m problems, total sample budget B, and number of repeated attempts k.
  Initialize $\texttt{successes}_i \leftarrow 0$ and $\texttt{attempts}_i \leftarrow 0$ for all $i \in \{1, \ldots, m\}$
  for $t \in \{1, \ldots, B\}$ **do**
    $i_t \leftarrow \texttt{SelectHardestProblem}(\texttt{successes}, \texttt{attempts})$
    $\texttt{attempts}_{i_t} \leftarrow \texttt{attempts}_{i_t} + 1$
    $\texttt{successes}_{i_t} \leftarrow \texttt{successes}_{i_t} + \mathbb{1}\{\texttt{AttemptProblem}(i_t)\}$
  end for
  $\hat{\alpha}, \hat{\beta} \leftarrow \arg\max_{\alpha,\beta>0} \prod_{i=1}^{m} \Pr[s = \texttt{successes}_i \mid b = \texttt{attempts}_i; \alpha, \beta]$ Equation 16
  $\widehat{\text{pass}_i@k} \leftarrow \mathbb{E}_{\text{pass}_i@1 \sim \text{Beta}(\hat{\alpha},\hat{\beta})}\left[1 - (1 - \text{pass}_i@1)^k\right]$ Equation 17
  return $\widehat{\text{pass}_i@k}$

On the sample-efficiency of our estimator. The decision to select problems dynamically based on estimated problem difficulty is motivated by intuition from the theorems in [26]. To provide theoretical motivation for our approach, we introduce a natural frequentist estimator, defined below. Given oracle access to $\text{pass}_i@1$ and control over the number of samples for each problem b_i, we show that the variance can be minimized by prioritizing "harder" problems with low $\text{pass}_i@1$.

Lemma 1. *Consider the following estimator of* $\text{pass}@k$

$$\widehat{\text{pass}_i@k}_{freq} := 1 - \frac{1}{n}\sum_{i=1}^{n}(1 - s_i/b_i)^k.$$

In the asymptotic regime as $n \to +\infty$*, the sampling budget* b^* *that minimizes the variance* $\text{Var}(\widehat{\text{pass}_i@k}_{freq})$ *is:*

$$b_i^* \propto \sqrt{(\text{pass}_i@1)(1 - \text{pass}_i@1)^{2k-1}}.$$

A proof of Lemma 1 is provided in Appendix G. For high k, the optimal budget prioritizes lower $\text{pass}_i@1$. We conjecture that such adaptive strategies can also reduce variance in the context of our multi-stage Bayesian approach, but we leave such analysis for future work. Beyond this, we show in Figure 3 that the distribution of the difficulties of problems selected by our dynamic strategy aligns much more closely with the derived optimal allocation from Lemma 1 than that of a uniform strategy.

However, in the sample-count regimes and distributions in our datasets, it is difficult to empirically isolate the benefits of the sampling method alone. Therefore, we provide some additional empirical support for dynamic sampling on synthetic data in Appendix G. We find that when there are many easy problems and a small number of hard outliers, or a uniform distribution of difficulties, the dynamic sampling method outperforms uniform sampling by large margins. On all distributions tested, dynamic sampling performs better than or comparably to uniform sampling.

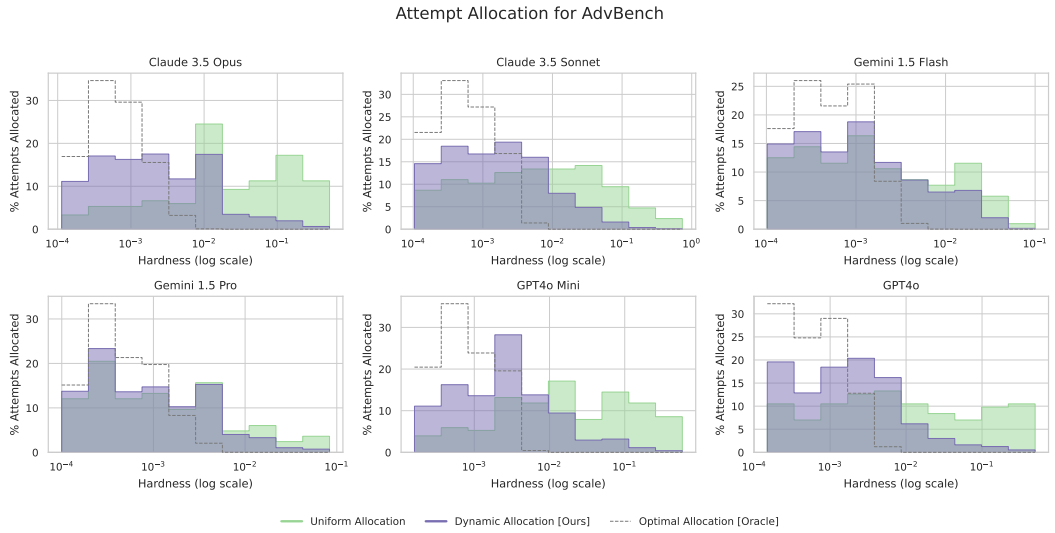

Figure 3: **Budget Allocation by Hardness Relative to the Optimal Allocation from Lemma 1** Contrasted distributions of problem success probabilities for the problems selected by dynamic and uniform sampling strategies on AdvBench. Note that these probabilities are not immediately available to our estimator but rather approximated given a limited amount of samples for each problem. The dotted line represents the distribution of problem success probabilities under the optimal sampling allocation provided in Lemma 1, assuming oracle access to the problem success probabilities. The dynamic strategy more closely aligns with this optimal rate.

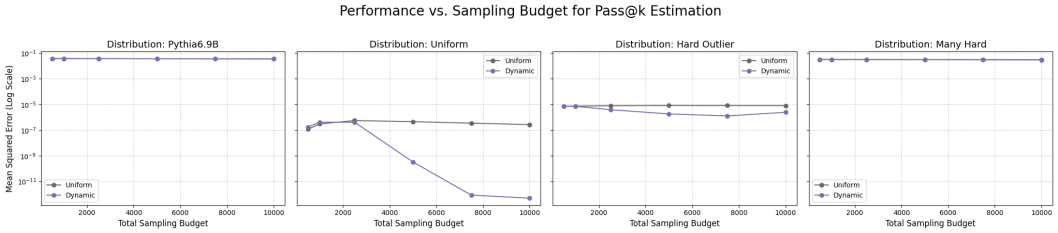

Figure 4: **Evaluating Performance Scaling for Uniform vs. Dynamic Allocation Strategies** Dynamic sampling is most useful when there are a handful of very difficult problems, but many easy problems. These distributions allow it to concentrate a large proportion of the budget on difficult problems. The "Hard Outlier" distribution has a single very difficult problem with success probability $1e-4$, and all other problems with difficulties in the range of 0.1-0.3.

Validity under adaptive sampling. We assume that per-problem attempt outcomes represent fresh Bernoulli draws with fixed per-problem probabilities of success $0 < \text{pass}_i@1 < 1$. This assumption notably *excludes history-dependent methods of repeated sampling*, such as self-reflection [28], code repair guided by execution feedback [19], and adaptive jailbreak prompting [5].

From here, convergence to the pseudo-true $\text{pass}@k$

$$\text{pass}_{\mathcal{D}}@k^* = \mathbb{E}_{\text{pass}_i@1 \sim \text{Beta}(\alpha^*,\beta^*)}\left[1-(1-\text{pass}_i@1)^k\right].$$

with likelihood-maximizing α^*, β^*

$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta} \prod_{i=1}^{m} \frac{\text{pass}_i@1^{\alpha-1}(1-\text{pass}_i@1)^{\beta-1}}{\text{Be}(\alpha,\beta)}.$$

follows from two key properties of our acquisition function: `SelectHardestProblem` depends only previous outcomes and attempts each problem infinitely often as $B \to \infty$, since the minimum success count increases after every problem obtains another success, which occurs almost surely.

By the first property, the acquisition rule contributes a parameter-independent factor to the likelihood of the observed trajectory [18, Eq. (4.7)]. Integrating the Bernoulli outcome factors over independent

$\text{pass}_i@1 \sim \text{Beta}(\alpha, \beta)$ gives

$$L(\alpha, \beta) \propto \prod_{i=1}^{m} \frac{\text{Be}(\alpha + s_i, \beta + b_i - s_i)}{\text{Be}(\alpha, \beta)}.$$

The binomial coefficients in Equation 16 are also parameter-independent, so the two objectives have the same maximizer.

By the second property, every problem is sampled infinitely often as $B \to \infty$ with m fixed, so $s_i/b_i \xrightarrow{\text{a.s.}} \text{pass}_i@1$ by the strong law. For convergence, we take both the fitted and oracle parameters to be global maximizers over the same fixed compact convex domain $\Theta \subset (0, \infty)^2$. After removing parameter-independent terms, the Beta–Binomial log likelihood converges almost surely and uniformly on Θ to the oracle Beta log likelihood obtained by observing the true success probabilities directly. This oracle objective is strictly concave and therefore has a unique maximizer, so $(\hat{\alpha}, \hat{\beta}) \xrightarrow{\text{a.s.}} (\alpha^*, \beta^*)$ by the strong-consistency argument of [24, p. 2122]. By continuity, $\widehat{\text{pass}_i@k} \xrightarrow{\text{a.s.}} \text{pass}_{\mathcal{D}}@k^*$ for each fixed k. The limiting prediction need not equal the true $\text{pass}_{\mathcal{D}}@k$, since the fitted Beta distribution may not capture the finite benchmark's distribution of success probabilities.

5 Results

In this section, we evaluate the predictive accuracy of our method. We estimate $\text{pass}_{\mathcal{D}}@k$ for k in the range $[10^1, 10^3]$ on three real-world datasets and three to six different models for each dataset

5.1 Experimental Setup

We source our data from [3] and [13], which contain 10^4 sampled successful or failed attempts for each of $100 \sim 200$ problems selected from Code Contests [22], MATH [11], and AdvBench [31]. We extrapolate from $10^1 \leq B \leq 10^3$ shuffled samples. Ground truth $\text{pass}@k$ values are computed using all 10^4 samples.

- **Uniform Sampling** (Log-Log Regression, Discretized Beta): use the first B/m for each problem.
- **Dynamic Sampling** (+ Beta-Binomial Fit): adaptive sampling as defined in Algorithm 1.

5.2 Discussion

The predictions for AdvBench, MATH, and Code Contests with different sampling budgets are shown in Figure 1. We observe that **existing estimators diverge significantly from the true** $\text{pass}@k$ **value beyond this threshold**. Figure 5 provides a heat map of errors for different sampling budgets and values of k. Note that, as expected, the error decreases as we increase the sampling budget. Existing estimators especially struggle with high values of k. We also provide the MSE for each estimator across different sampling budgets in Appendix H.

Across models and datasets, our proposed method provides predictions that are closest to the ground truth. The predictions from log-log regression are particularly poor, often diverging to predict impossible pass rates greater than 1 (we clip these at 1 for visualization and error computation). The prior distributional fitting method from [26] performs better than unclipped regression but consistently underestimates $\text{pass}@k$ for large k.

The gains from our method come from several sources. Our fitting uses maximum likelihood estimation on an exponential family model, which is known to have properties like asymptotic normality, unbiasedness, and $O(1/\sqrt{n})$-convergence. We avoid the pitfall of fitting models on top of correlated estimates, as in the regression method. Finally, we align our sampling budget more closely with the distribution that theoretically minimizes variance.

6 Conclusion and Future Work

Predicting the capabilities and vulnerabilities of AI models at scale is a critical challenge. We provide two core improvements: (1) selecting a more appropriate model for the underlying problem

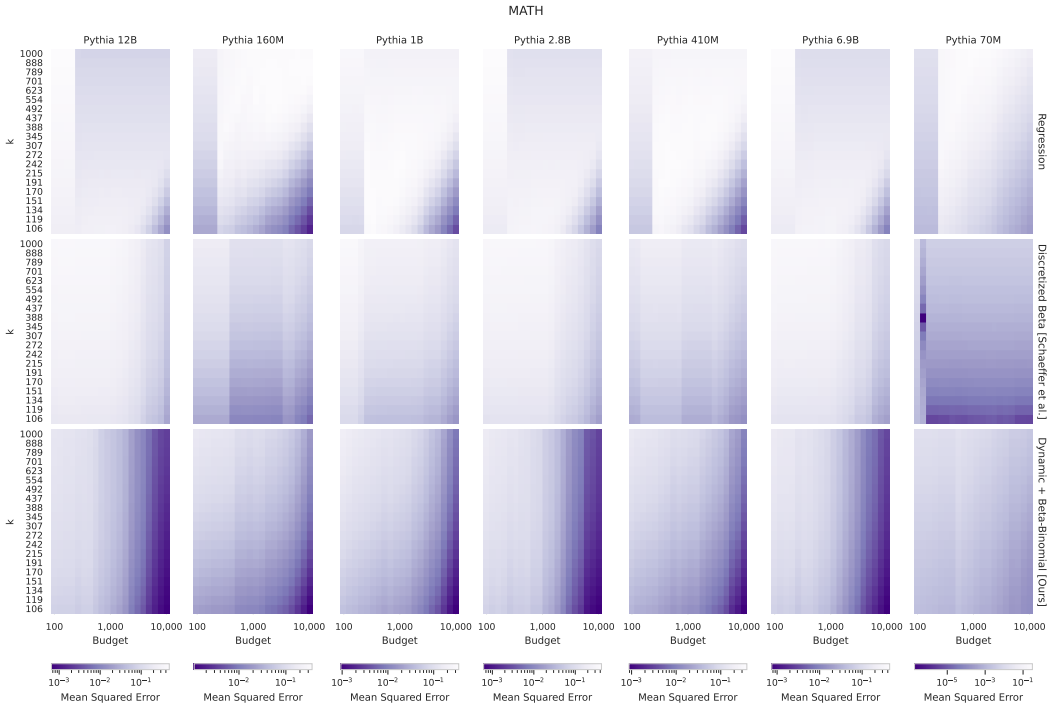

Figure 5: Heatmap depicting how predictions of pass@k change with the sampling budget and k on MATH. Our method minimizes MSE for virtually all values of k and sampling budgets, as evidenced by the darker colors in its heatmap. Figures for MATH and Code Contests are in Appendix H.

difficulties, and (2) utilizing dynamic sampling to concentrate compute on the most difficult problems. We demonstrate the impact of these innovations in Figure 5 on mathematical problem-solving.

We achieved large improvements in predictive accuracy by remedying statistical errors in prior methods and improving sampling techniques, without requiring extra sampling compute. These gains suggest that a closer statistical inspection of other scaling-law fitting methodologies could lead to considerable computational savings and, ultimately, better and safer models.

7 Related Work

Repeated sampling from LMs improves performance on verifiable tasks, and has become the backbone of reinforcement learning from verified rewards [27]. The HumanEval benchmark first defined pass@k and derived an unbiased estimator for this quantity [7]. Follow-up work on code generation studied sampling as a resource that must be allocated across problems; for instance, [10] adaptively prioritizes sampling solvable instances. They estimate pass@k with Monte-Carlo, whereas we predict how pass@k varies with k on a fixed budget.

For unverifiable rewards, repeated sampling must be combined with aggregation techniques such as majority voting [29]. Self-consistency shows that majority voting over multiple chains-of-thought can boost reasoning performance, and adaptive-consistency [1] reduces compute by stopping sampling once a clear majority emerges. [6] study compound AI systems that repeatedly query an LM, aggregate responses, and describe how performance varies with the number of LM calls.

Statistically, estimating pass@k involves binomial success probabilities that can be extremely small on difficult benchmark items, making variance control and uncertainty quantification important. Classical work on binomial proportion intervals compares exact and approximate methods, including the Clopper-Pearson interval and score-based or adjusted Wald intervals [4]. Recent analysis in the rare-event regime emphasizes that both coverage and relative margin of error are important when designing estimators and determining sample sizes [23].

Acknowledgments and Disclosure of Funding

This work is partially supported by NSF 2046795, 2205329, and 2504264, NIH, ARPA-H, the MacArthur Foundation, Good Ventures, Schmidt Sciences, the Hasso Plattner Foundation and Stanford HAI.

References

[1] Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms, 2023.

[2] Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024.

[4] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117, 2001.

[5] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.

[6] Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards the scaling properties of compound ai systems. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 45767–45790. Curran Associates, Inc., 2024.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[8] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA, 2024. Association for Computing Machinery.

[9] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024.

[10] Hojae Han, Yu Jin Kim, Byoungjip Kim, Youngwon Lee, Kyungjae Lee, Kyungmin Lee, Moontae Lee, Kyunghoon Bae, and Seung-won Hwang. On sample-efficient code generation. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 783–791, Singapore, December 2023. Association for Computational Linguistics.

[11] 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.

[12] Nikolaus Howe, Ian McKenzie, Oskar Hollinsworth, Michał Zajac, Tom Tseng, Aaron Tucker, Pierre-Luc Bacon, and Adam Gleave. Scaling trends in language model robustness, 2025.

[13] John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking, 2024.

[14] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Joshua Kazdan, Abhay Puri, Rylan Schaeffer, Lisa Yu, Chris Cundy, Jason Stanley, Sanmi Koyejo, and Krishnamurthy Dvijotham. No, of course i can! deeper fine-tuning attacks that bypass token-level safety mechanisms, 2025.

[16] Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. Spoc: Search-based pseudocode to code. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[17] Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. Robomonkey: Scaling test-time sampling and verification for vision-language-action models, 2025.

[18] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[19] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. CodeRL: Mastering code generation through pretrained models and deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[20] L.E. Lehmann. *Theory of Point Estimation*. A Wiley publication in mathematical statistics. Wiley, 1983.

[21] Noam Levi. A simple model of inference scaling laws, 2024.

[22] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[23] Owen McGrath and Kevin Burke. Binomial confidence intervals for rare events: importance of defining margin of error relative to magnitude of proportion, 2024.

[24] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1994.

[25] Alexander Panfilov, Paul Kassianik, Maksym Andriushchenko, and Jonas Geiping. Capability-based scaling laws for llm red-teaming, 2025.

[26] Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)?, 2025.

[27] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[28] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[30] Michael J. Zellinger and Matt Thomson. Rational tuning of LLM cascades via probabilistic modeling. *Transactions on Machine Learning Research*, 2025.

[31] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Limitations

While we predict pass@k on a scale of pass@10 000 using dozens to hundreds of samples, frontier labs face the challenge of predicting pass@k for k several orders of magnitude higher. These frontier labs also have the capability to generate far more samples to aid their predictions than we can. Although we hope that the lessons of how to cross orders of magnitude when predicting pass@k transfer, power law fitting might be more viable when the number of samples available is higher. In the same vein, because we are limited in the number of samples that we can draw, we are unable to estimate the precision of pass@1 beyond $1e-4$, which creates hard limits for us when testing the accuracy of our models.

B Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted.

C Adaptive Sampling Criterion

Algorithm 2 `SelectHardestProblem`

Require: Dataset $\mathcal{D}$ with m problems and per-problem counts of successful and total attempts: `successes` and `attempts`, respectively.
$s^* \leftarrow \min_i \texttt{successes}_i$
$H \leftarrow \arg\min_{\{i\,:\,\texttt{successes}_i = s^*\}} \texttt{attempts}_i$
$i^* \sim \mathrm{Uniform}(H)$
return i^*

D Pitfalls of Linear Regression

In this section, we precisely quantify the statements made in Section 3.2.

The estimates $\widehat{\mathrm{pass@}k}$ are not independent for different k: Recall that one of the assumptions of the linear regression model is that the observations are independent. The following lemma characterizes this non-independence on a per-problem basis:

Lemma 2. *Recall that s_i is the number of successes observed out of b attempts on the ith problem of $\mathcal{D}$. If $k \geq l$, and $0 < s_i < b$ then there exists an invertible function f such that*

$$\widehat{pass_i@k} = f\left(\widehat{pass_i@l}\right). \tag{18}$$

This invertible function takes the form:

$$f\left(\widehat{pass_i@l}\right) = \widehat{pass_i@l} + s_i \sum_{m=l}^{k-1} \frac{\binom{b-s_i}{m}}{(b-m)\binom{b}{m}}. \tag{19}$$

Proof.

$$\text{Let} \quad g(m) = \frac{\binom{b-s_i}{m}}{\binom{b}{m}}, \quad \text{then} \quad \widehat{\mathrm{pass}_i@m} = 1 - g(m).$$

$$\text{Now,} \quad \frac{g(m+1)}{g(m)} = \frac{\binom{b-s_i}{m+1}}{\binom{b}{m+1}} \cdot \frac{\binom{b}{m}}{\binom{b-s_i}{m}}$$
$$= \frac{\binom{b-s_i}{m+1}}{\binom{b-s_i}{m}} \cdot \frac{\binom{b}{m}}{\binom{b}{m+1}}$$
$$= \frac{b-s_i-m}{m+1} \cdot \frac{m+1}{b-m}$$
$$= \frac{b-s_i-m}{b-m}.$$

$$\Rightarrow 1 - g(m+1) = 1 - \frac{b-s_i-m}{b-m} g(m)$$
$$\Rightarrow 1 - g(m+1) = (1-g(m)) + g(m)\left(1 - \frac{b-s_i-m}{b-m}\right)$$
$$= (1-g(m)) + g(m) \cdot \frac{s_i}{b-m}.$$

$$\Rightarrow \widehat{\text{pass}_i@(m+1)} = \widehat{\text{pass}_i@m} + g(m) \cdot \frac{s_i}{b-m}$$
$$\Rightarrow \widehat{\text{pass}_i@k} = \widehat{\text{pass}_i@l} + s_i \sum_{m=l}^{k-1} \frac{1}{b-m} g(m) \quad \text{as desired.}$$

□

This lemma implies that given $\text{pass}_i@k$ for any k, $\text{pass}_i@j$ for $j \neq k$ is uniquely determined.

The estimates of $\widehat{\text{pass@}k}$ have different variances for different values of k: A second assumption of the linear regression model is that the noise in the model is homoscedastic, i.e. the noise is the same for all k. This is again not the case for the estimators $\widehat{\text{pass@}k}$. The following lemma gives one instance in which these estimators are not homoscedastic:

Lemma 3. *Suppose that we have n samples from a language model on problem i, and the language model has true probability p of getting problem i correct. Then*

$$\textit{Var}\left(\widehat{\text{pass}_i@n}\right) = (1-p)^n - (1-p)^{2n}, \tag{20}$$

and

$$\textit{Var}\left(\widehat{\text{pass}_i@1}\right) = p(1-p)/n. \tag{21}$$

Proof. Let $c \sim \text{Binomial}(n, p)$ be the number of correct completions obtained from n i.i.d. samples of a fixed problem i. For each $k \in \{0, 1, \ldots, n\}$ define the empirical *pass@k* estimator

$$\widehat{\text{pass}_i@k} = f_k(c), \text{ where } f_k(c) \;=\; 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}$$

Our goal is to show that the variances of $\widehat{\text{pass}_i@k}$ are not constant in k. We begin with the variance in its raw definition:

$$\text{Var}\big[f_k(c)\big] \;=\; \underbrace{\mathbb{E}\big[f_k(c)^2\big]}_{(a)} \;-\; \underbrace{\Big(\mathbb{E}\big[f_k(c)\big]\Big)^2}_{(b)}. \tag{$\star$}$$

Both expectations can be written as finite sums over the binomial probability-mass function:

$$(a) \;=\; \sum_{c=0}^{n} \Big(1 - \tfrac{\binom{n-c}{k}}{\binom{n}{k}}\Big)^2 \binom{n}{c} p^c (1-p)^{n-c}, \quad (b) \;=\; \Big(\sum_{c=0}^{n} \Big(1 - \tfrac{\binom{n-c}{k}}{\binom{n}{k}}\Big) \binom{n}{c} p^c (1-p)^{n-c}\Big)^2.$$

We now specialize to two extreme choices of k.

Case $k = n$

Because $\binom{n-c}{n} = 1$ if $c = 0$ and 0 otherwise,

$$f_n(c) = 1 - \binom{n-c}{n} = \mathbf{1}_{\{c \geq 1\}} \in \{0, 1\}, \quad \text{hence } f_n(c)^2 = f_n(c).$$

Next we compute the first and second moments.

$$\begin{aligned}
\mathbb{E}[f_n(c)] = \mathbb{E}[f_n(c)^2] &= \sum_{c=0}^{n} \mathbf{1}_{\{c \geq 1\}} \binom{n}{c} p^c (1-p)^{n-c} \\
&= \sum_{c=1}^{n} \binom{n}{c} p^c (1-p)^{n-c} \\
&= 1 - \binom{n}{0} p^0 (1-p)^n, \quad \text{Since the binomial PMF is normalized} \\
&= 1 - (1-p)^n
\end{aligned}$$

Plugging the two moments into ($\star$),

$$\mathrm{Var}\big[f_n(c)\big] \;=\; \big[1 - (1-p)^n\big] \;-\; \big[1 - (1-p)^n\big]^2 \;=\; (1-p)^n - (1-p)^{2n}.$$

Case $k = 1$

$$f_1(c) = 1 - \frac{n-c}{n} = \frac{c}{n}.$$

Because $\mathbb{E}[c] = np$ and $\mathrm{Var}[c] = np(1-p)$,

$$\begin{aligned}
\mathbb{E}[f_1(c)] &= \frac{1}{n}\,\mathbb{E}[c] = p, \quad \text{and} \\
\mathbb{E}\big[f_1(c)^2\big] &= \frac{1}{n^2}\,\mathbb{E}[c^2] \\
&= \frac{1}{n^2}\big(\mathrm{Var}[c] + \mathbb{E}[c]^2\big) \\
&= \frac{1}{n^2}\big(np(1-p) + n^2p^2\big) \\
&= \frac{p(1-p)}{n} + p^2.
\end{aligned}$$

finally,

$$\mathrm{Var}\big[f_1(c)\big] \;=\; \Big(\frac{p(1-p)}{n} + p^2\Big) - p^2 \;=\; \frac{p(1-p)}{n}.$$

□

E More Flexible Fitting Methods

[26] claimed that a standard beta distribution was not flexible enough to fit the distribution of $\mathrm{pass}_i@1$, leading them to model the distribution of $\mathrm{pass}_i@k$ as a scaled beta-binomial rather than a beta-binomial distribution. The authors developed the discretized fitting method described in Section 3.3 because they could not find a tractable likelihood for the three-parameter beta-binomial distribution.

In this section, we derive a tractable likelihood for the scaled beta-binomial distribution, allowing us to avoid estimating $\hat{\theta}$ from equation 9 using the unprincipled estimator from equation 10. A tractable likelihood also allows us to fit the scaled beta-binomial distribution directly to n, k_i rather than first estimating $\mathrm{pass}_i@k$ and fitting the scaled beta distribution to these estimates.

We first rewrite the expression for the likelihood of the scaled beta-binomial distribution to remove the integral in the following lemma:

Lemma 4. *The likelihood for the scaled beta-binomial distribution is given by*

$$\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\int_0^\theta p^k(1-p)^{n-k}\left(\frac{p}{\theta}\right)^{\alpha-1}\left(1-\frac{p}{\theta}\right)^{\beta-1}\frac{1}{\theta}dp \tag{22}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\sum_{i=0}^{n-k}\binom{n-k}{i}(-1)^i\theta^{k+i}\mathrm{Be}(k+i+\alpha,\beta). \tag{23}$$

The proof can be found in Appendix F.

Although the resulting likelihood no longer contains an integral, it involves an alternating sum of potentially large terms. Define

$$W_i=\binom{n-k}{i}\theta^{k+i}\mathrm{Be}(k+i+\alpha,\beta). \tag{24}$$

In terms of W_i, our optimization objective is

$$-\log\left(\sum_{i=0}^{n-k}(-1)^iW_i\right). \tag{25}$$

To compute this as stably as possible, we use an alternating log-sum-exp function. Letting $W_m=\max\{W_0,...,W_{n-k}\}$, our log likelihood becomes:

$$-\log\left(\sum_{i=0}^{n-k}(-1)^i\exp(\log(W_i)-\log(W_m))\right)-\log(W_m). \tag{26}$$

F Scaled Beta-Binomial Likelihood

$$\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\int_0^\theta p^k(1-p)^{n-k}\left(\frac{p}{\theta}\right)^{\alpha-1}\left(1-\frac{p}{\theta}\right)^{\beta-1}\frac{1}{\theta}dp \tag{27}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\theta^k\int_0^\theta\left(\frac{p}{\theta}\right)^k(1-p)^{n-k}\left(\frac{p}{\theta}\right)^{\alpha-1}\left(1-\frac{p}{\theta}\right)^{\beta-1}\frac{1}{\theta}dp \tag{28}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\theta^k\sum_{i=0}^{n-k}\binom{n-k}{i}\int_0^\theta(-1)^ip^i\left(\frac{p}{\theta}\right)^{k+\alpha-1}\left(1-\frac{p}{\theta}\right)^{\beta-1}\frac{1}{\theta}dp \tag{29}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\sum_{i=0}^{n-k}\binom{n-k}{i}\int_0^\theta\theta^{k+i}(-1)^i\left(\frac{p}{\theta}\right)^{k+i+\alpha-1}\left(1-\frac{p}{\theta}\right)^{\beta-1}\frac{1}{\theta}dp \tag{30}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\sum_{i=0}^{n-k}\binom{n-k}{i}(-1)^i\theta^{k+i}\mathrm{Be}(k+i+\alpha,\beta) \tag{31}$$

$$=\frac{1}{\mathrm{Be}(\alpha,\beta)}\binom{n}{k}\sum_{i=0}^{n-k}\binom{n-k}{i}(-1)^i\theta^{k+i}\mathrm{Be}(k+i+\alpha,\beta) \tag{32}$$

Define

$$W_i=\binom{n-k}{i}\theta^{k+i}\mathrm{Be}(k+i+\alpha,\beta). \tag{33}$$

Our optimization objective is

$$-\log\left(\sum_{i=0}^{n-k}(-1)^iW_i\right). \tag{34}$$

To compute this as stably as possible, we use an alternating log-sum-exp function. Letting $W_m = \max\{W_0, ..., W_{n-k}\}$, our log likelihood becomes:

$$-\log\left(\sum_{i=0}^{n-k}(-1)^i \exp(\log(W_i) - \log(W_m))\right) - \log(W_m). \tag{35}$$

$$\begin{aligned} \text{pass}_i@1 &\sim \text{Beta}(\alpha, \beta, \theta) \\ k_i &\sim \text{Binomial}(n, \text{pass}_i@1) \end{aligned}$$

G Optimal Distribution of Samples

G.1 Proofs

Lemma 5 (Variance in the Asymptotic Regime)**.** *For a sequence of random random variables $\{x_n\}$ such that $x_n = y_n/n$ where $y_n \sim$ Bin(n, p), we have the following:*

$$\sqrt{n}((1-x_n)^k - (1-p)^k) \xrightarrow{d} \mathcal{N}(0, pk^2(1-p)^{2k-1})$$

Proof. By the Central Limit Theorem,

$$\sqrt{n}((1-x_n) - (1-p)) \xrightarrow{d} \mathcal{N}(0, p(1-p)) \tag{36}$$

Let $g : \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$g(t) = t^k$$

Applying the delta method:

$$\begin{aligned} \sqrt{n}((1-x_n)^k - (1-p)^k) &\xrightarrow{d} \mathcal{N}(0, g'(1-p)^2 p(1-p)) && (37) \\ &\xrightarrow{d} \mathcal{N}(0, (k(1-p)^{k-1})^2 p(1-p)) && (38) \\ &\xrightarrow{d} \mathcal{N}(0, pk^2(1-p)^{2k-1}) && (39) \end{aligned}$$

□

Lemma 6 (Variance-Minimizing Budget)**.** *Consider a random variable $X = \sum_{i=1}^m X_i$ where each X_i is an independent random variable with variance $\text{Var}(X_i) = v_i/b_i$.*

Consider the positive scaled simplex $B = \{b : b_i > 0 \ \ \& \ \ \sum_j^m b_j = B\}$. We have the following:

$$b^* = \arg\min_{b \in B} \text{Var}(X; b) \tag{40}$$

$$b_i^* = \frac{\sqrt{v_i}}{\sum_j^m \sqrt{v_j}} \tag{41}$$

Proof. Our objective is this:

$$\min_{b_i > 0} \sum_{i=1}^m v_i/b_i \ \text{ s.t. } \ \sum_{i=1}^m b_j = B$$

This objective is convex as a sum of convex functions, meaning we can use the Lagrange method:

$$\mathcal{L}(b, \lambda) = \sum_{i=1}^m v_i/b_i + \lambda\left(\sum_{i=1}^m b_i - B\right) \tag{42}$$

Applying first order conditions we get the following:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_i} &= -v_i/b_i^2 + \lambda && (43) \\ 0 &= -v_i/b_i^2 + \lambda && (44) \\ b_i &= \sqrt{v_i/\lambda} && (45) \\ b_i &\propto \sqrt{v_i} && (46) \end{aligned}$$

□

Combining Lemma 5 and Lemma 6, we have Lemma 1.

G.2 Synthetic Comparison of Uniform and Dynamic Sampling

We demonstrate the gains possible with dynamic sampling via the following contrived scenario: half of the problems are "easy" ($\text{pass}_i@1 = 0.3$) and half of the problems are "impossible" ($\text{pass}_i@1 = 0$). In this instance, we expect $\text{pass}@k \to 1/2$ as $k \to \infty$. However, without a sufficient allocation of samples to the "impossible" problems, the uniform sampling strategy prevents our estimator from determining whether these problems are impossible or just hard (i.e., still likely to be solved in k attempts). This results in an upwards-biased estimate and relatively slow improvement of MSE as the budget grows.

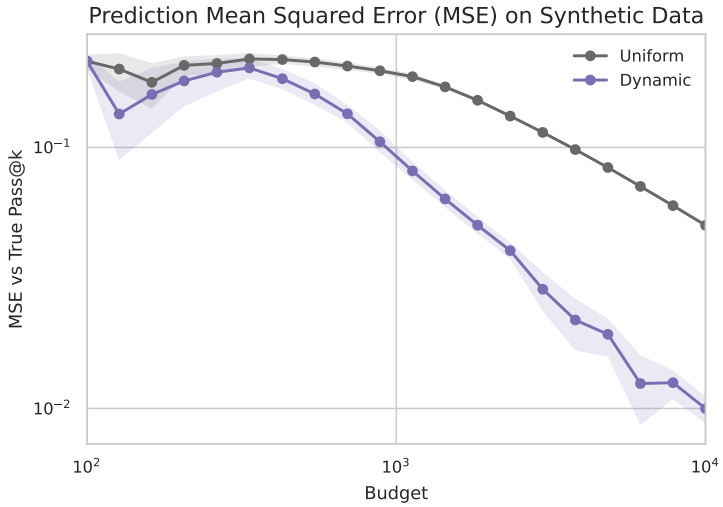

Figure 6: The MSE of our estimator with both dynamic and uniform sampling strategies given the described synthetic problem success probabilities, $n = 100$ problems and $k = 1\,000$. By focusing on the most difficult problems, the dynamic strategy allows our estimator to converge rapidly to the true $\text{pass}@k$ value.

We also provide some insight into the distributions for which dynamic sampling has advantages over uniform. We find that for uniform difficulty distributions or distributions that contain a handful of very hard outlier problems, dynamic sampling provides the most advantage. For distributions with many (or mostly) difficult problems, dynamic sampling holds little to no advantage over uniform sampling, since in these cases, uniform and dynamic sampling distribute the budget very similarly. If only a handful of problems are quickly solved, then dynamic sampling has very little extra samples to allocate to the more difficult problems.

Method	Uniform Sampling	Dynamic Sampling
Beta-Binomial	0.000017 [0.000006, 0.000030]	0.000006 [0.000001, 0.000012]
Discretized Beta	0.030204 [0.023471, 0.037151]	0.002609 [0.001329, 0.004156]

Table 1: The MSE of our estimator and the discretized-Beta estimator with both dynamic and uniform sampling strategies given synthetic problem success probabilities sampled from Uniform($[0, 1]$) for $n = 64$ problems, $k = 1\,000$, and the budget fixed at 256.

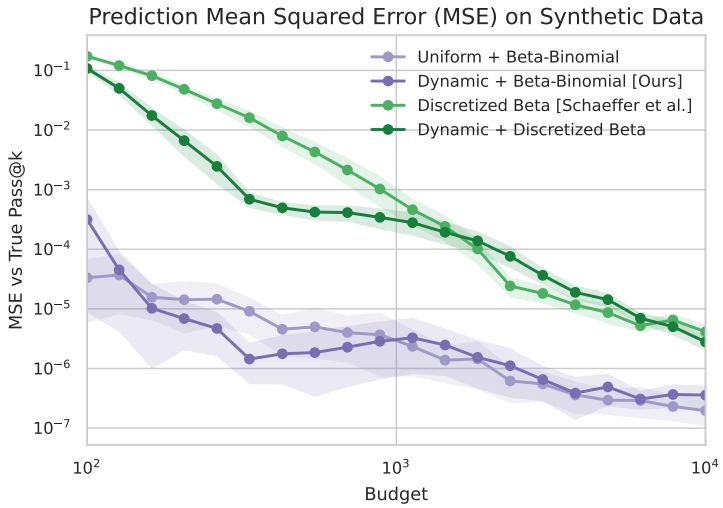

Figure 7: The MSE of our estimator and the discretized-Beta estimator with both dynamic and uniform sampling strategies given synthetic problem success probabilities sampled from Uniform$([0, 1])$ for $n = 64$ problems and $k = 1\,000$. We observe that dynamic sampling leads to modest improvements in both estimators.

H Additional Figures

We provide matching figures from the main paper for the benchmarks that were omitted due to lack of space. Additionally, we include plots that track the scaling of mean squared error (MSE) as budget increases for fixed k, and figures containing the performance of a mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$

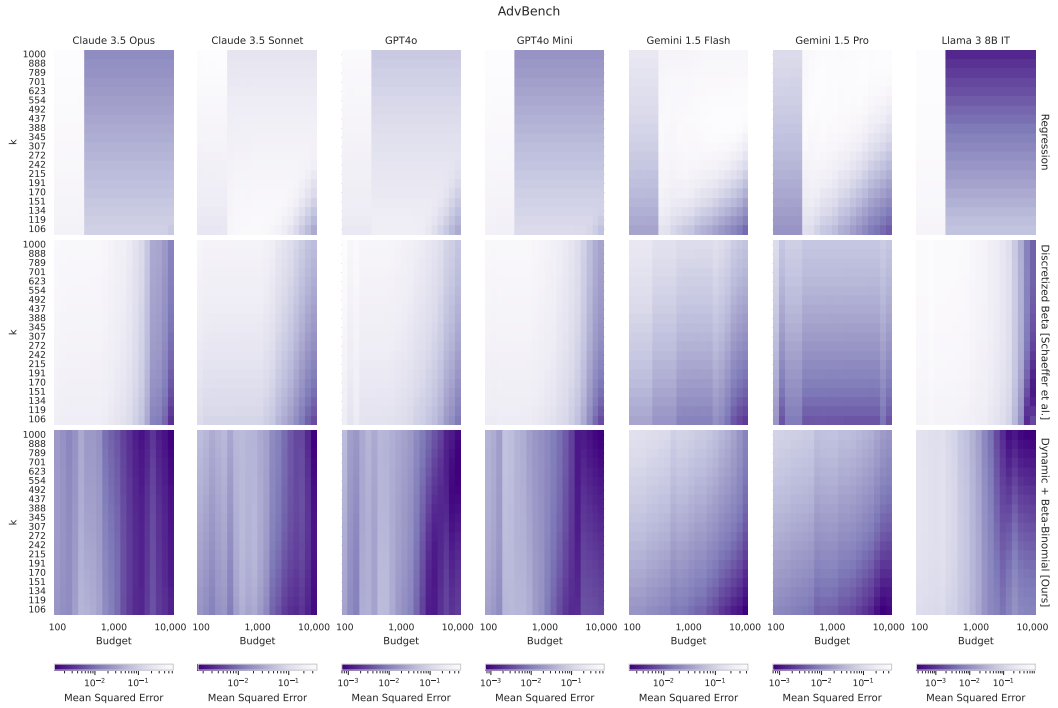

Figure 8: Heatmap depicting how predictions of $\mathrm{pass@}k$ change with the sampling budget and k for the AdvBench benchmark. Darker colors indicate lower mean squared error.

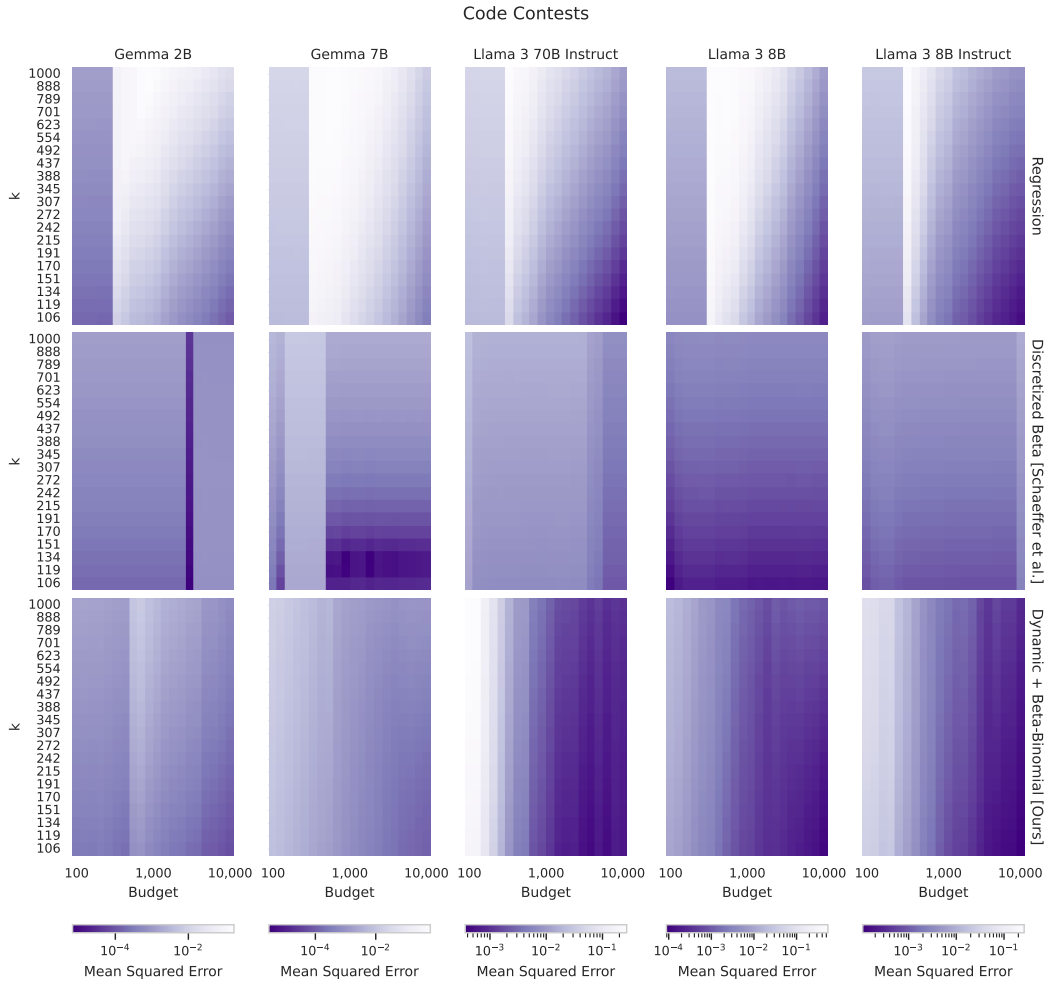

Figure 9: Heatmap depicting how predictions of $\mathrm{pass@}k$ change with the sampling budget and k for the Code Contests benchmark. Darker colors indicate lower mean squared error.

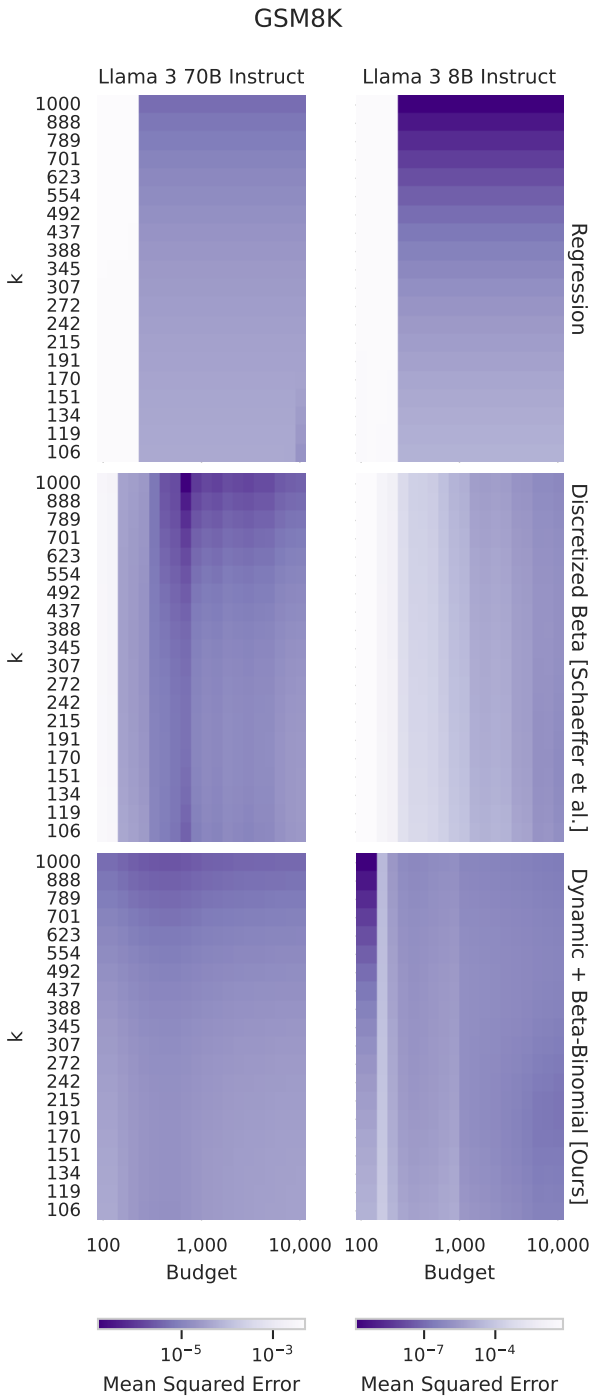

Figure 10: Heatmap depicting how predictions of $\mathrm{pass@}k$ change with the sampling budget and k for the GSM8K benchmark. Darker colors indicate lower mean squared error.

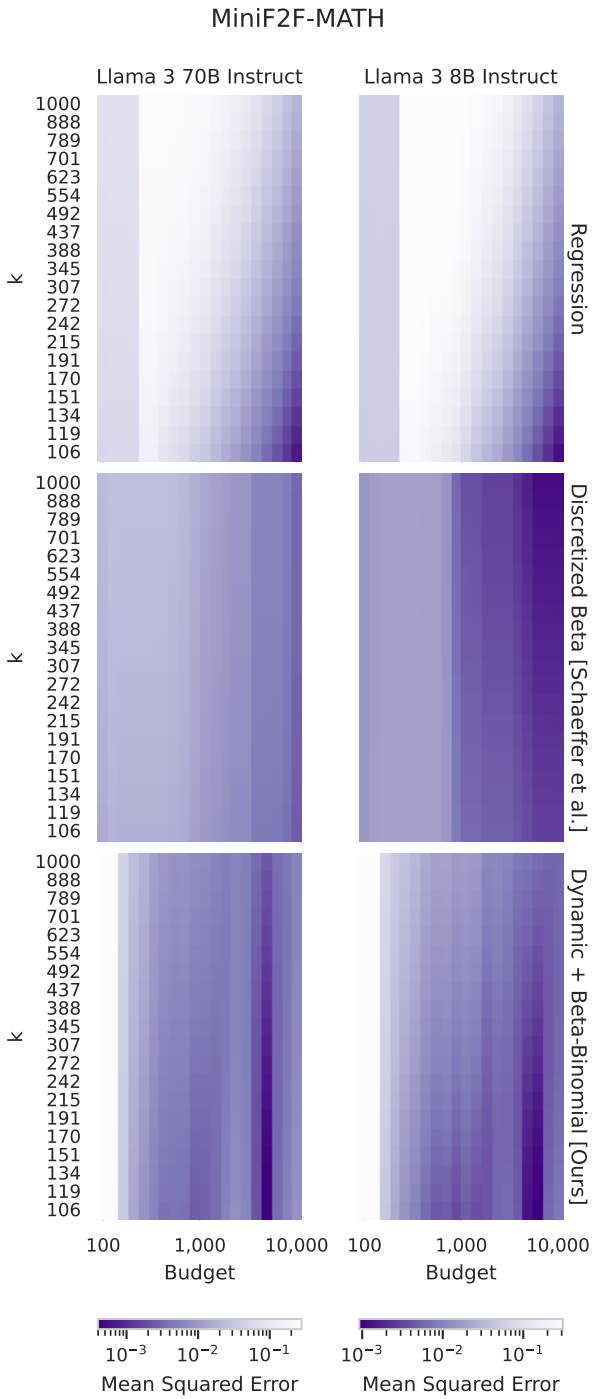

Figure 11: Heatmap depicting how predictions of $\mathrm{pass}@k$ change with the sampling budget and k for the MiniF2F-MATH benchmark. Darker colors indicate lower mean squared error.

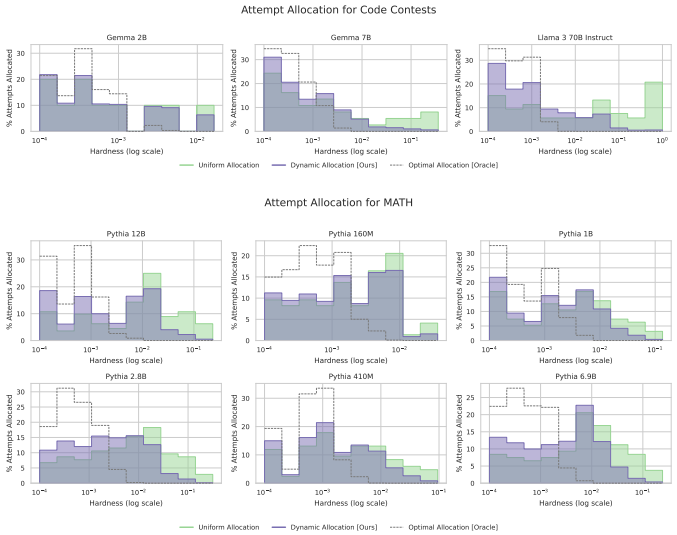

Figure 12: Contrasted distributions of problem success probabilities for the problems selected by dynamic and uniform sampling strategies on Code Contests and MATH.

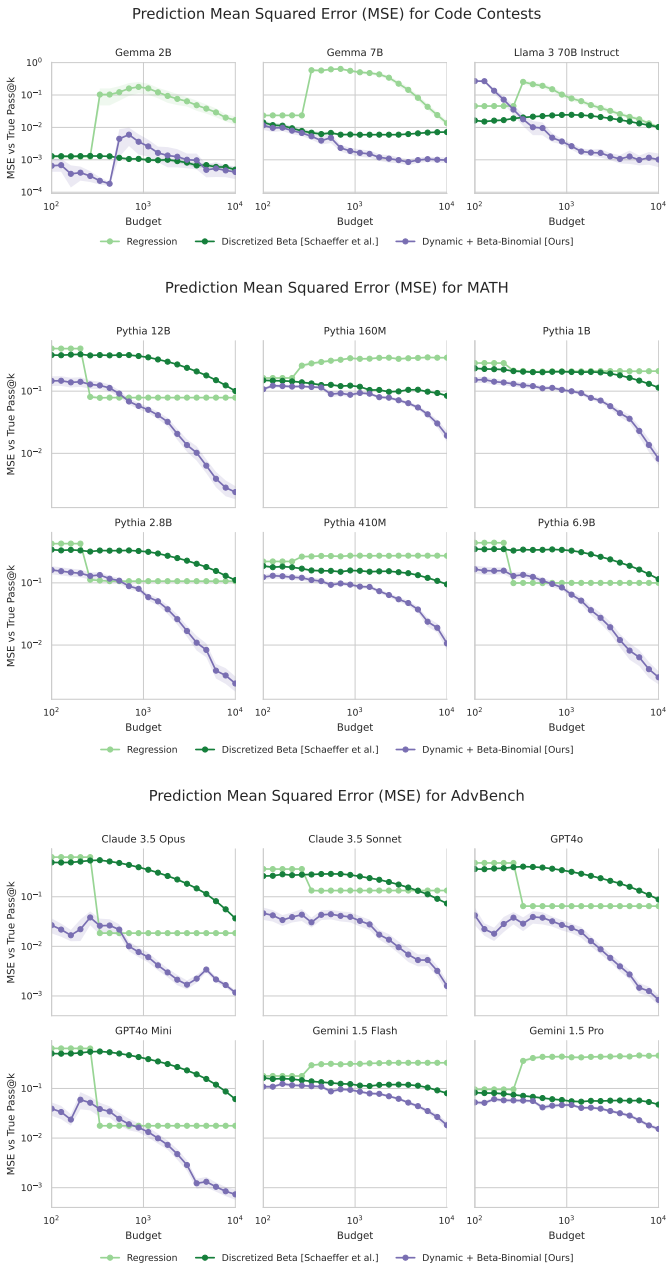

Figure 13: MSE scaling with increasing budget. As expected, more samples generally leads to a reduction in MSE across all approaches. For some models our approach reaches MSE more than 10x lower than its counterparts.

I Datasets

We draw our evaluation data from two recent sources: [3] and [13]. They record, for each of 128 prompt samples, the **number of successful outcomes out of** 10 000 **trials**. These prompts are sampled from three benchmark suites:

- **CodeContests** [22]: A competitive programming benchmark which collects description-to-code tasks from contest platforms such as AtCoder, CodeChef, Codeforces, and HackerEarth. Models are evaluated on precise correctness via test cases. Later refinements (e.g. CodeContests+) improve test case generation and validation to reduce false positives in evaluation.

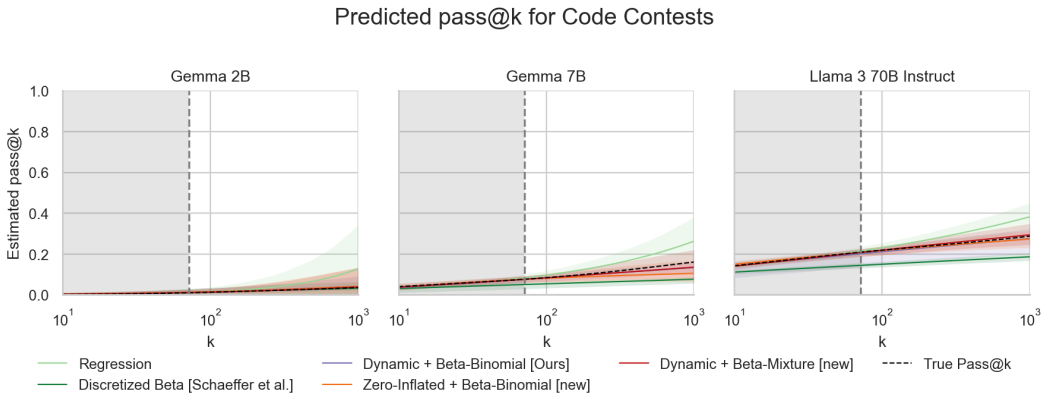

Figure 14: $pass@k$ plots comparing the performance of the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ (dynamic sampling) to other methods on code contest data.

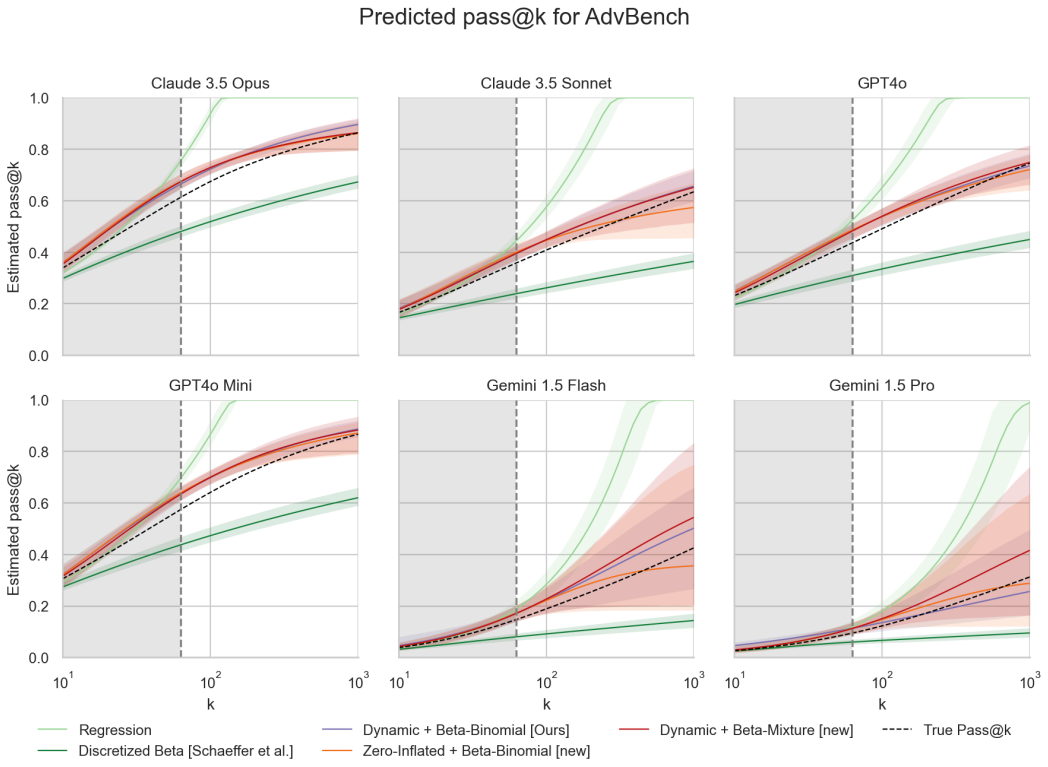

Figure 15: $pass@k$ plots comparing the performance of dynamic sampling applied to the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ [30] to other methods on BON Jailbreaking data.

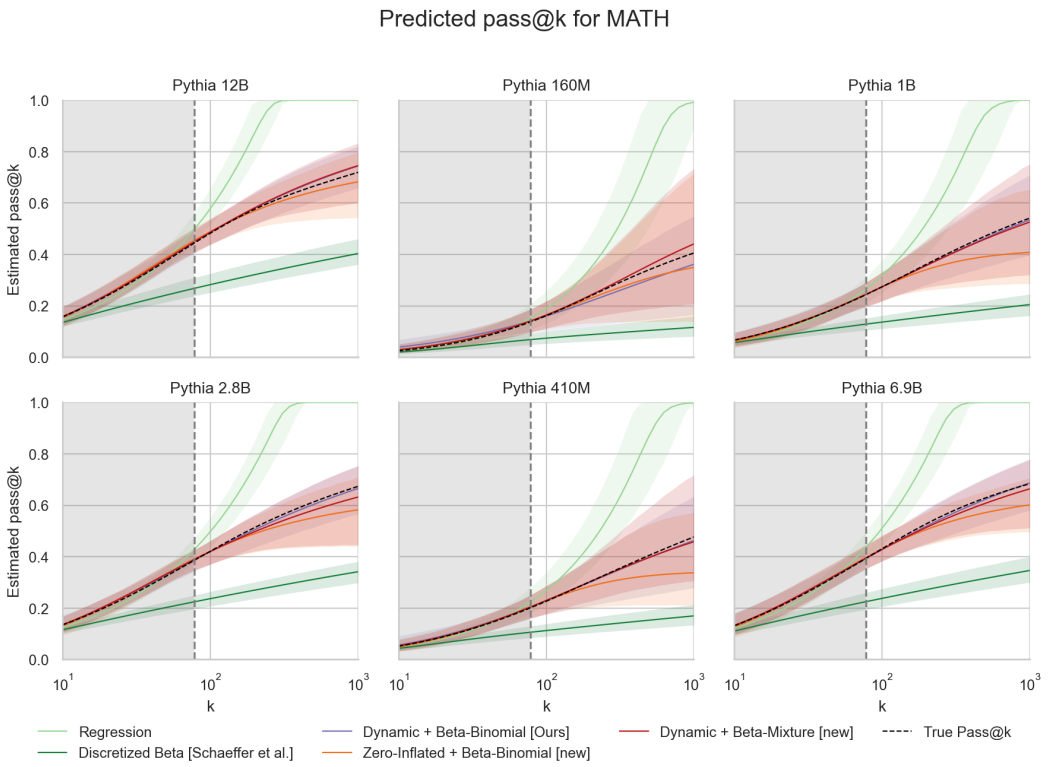

Figure 16: $pass@k$ plots comparing the performance of dynamic sampling applied to the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ [30] to other methods on MATH data.

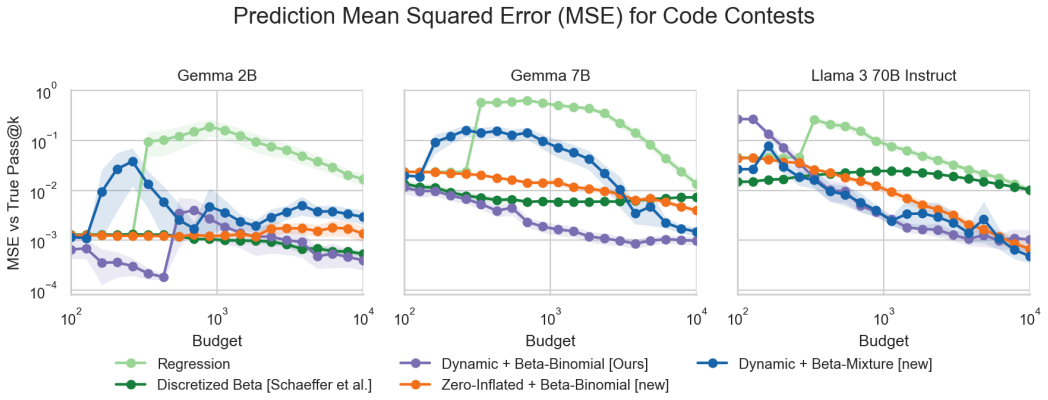

Figure 17: MSE plots comparing the performance of the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ (dynamic sampling) to other methods on code contest data.

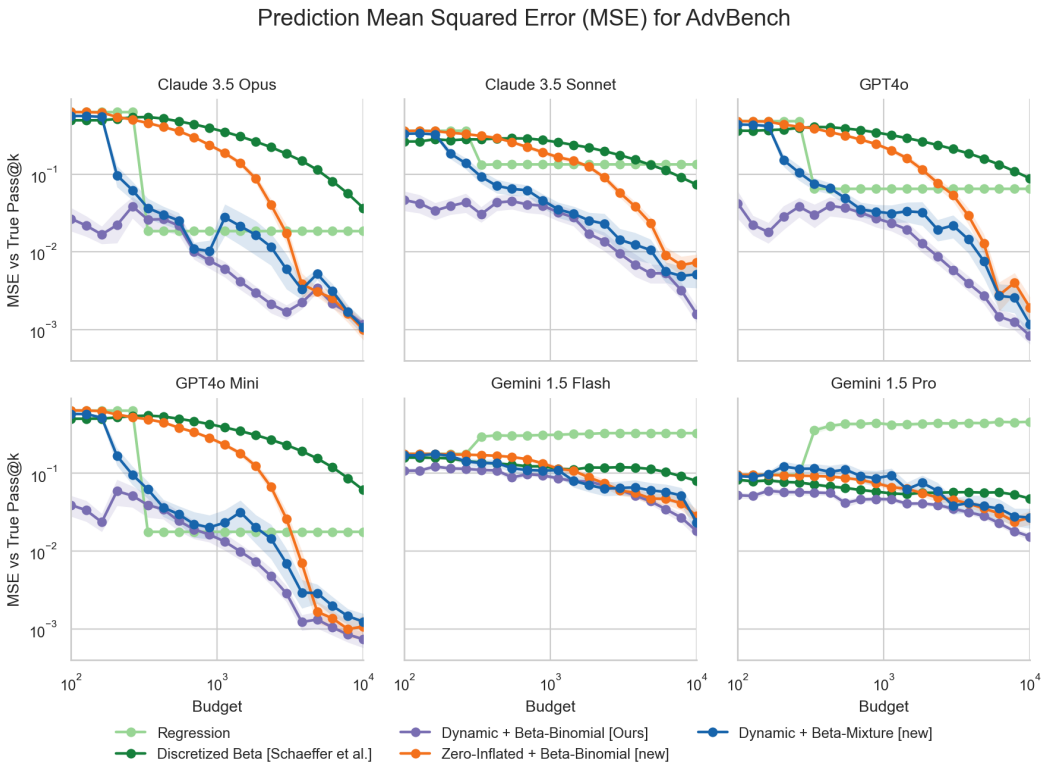

Figure 18: MSE plots comparing the performance of the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ (dynamic sampling) to other methods on jail-breaking data.

- **MATH** [11]: A mathematical reasoning dataset of 12,500 high school competition problems (e.g. AMC, AIME). Each problem comes with a full solution path and final answer. The benchmark evaluates model proficiency in multi-step reasoning across domains such as algebra, number theory, geometry, and combinatorics.
- **AdvBench** [31]: An adversarial NLP benchmark oriented toward security tasks. It emphasizes realistic attacker goals and evaluates models' success or failure under adversarial prompting strategies.

This combination lets us evaluate the efficacy of our estimator on problem success probability distributions extracted from **coding**, **mathematical reasoning**, and **adversarial robustness** domains.

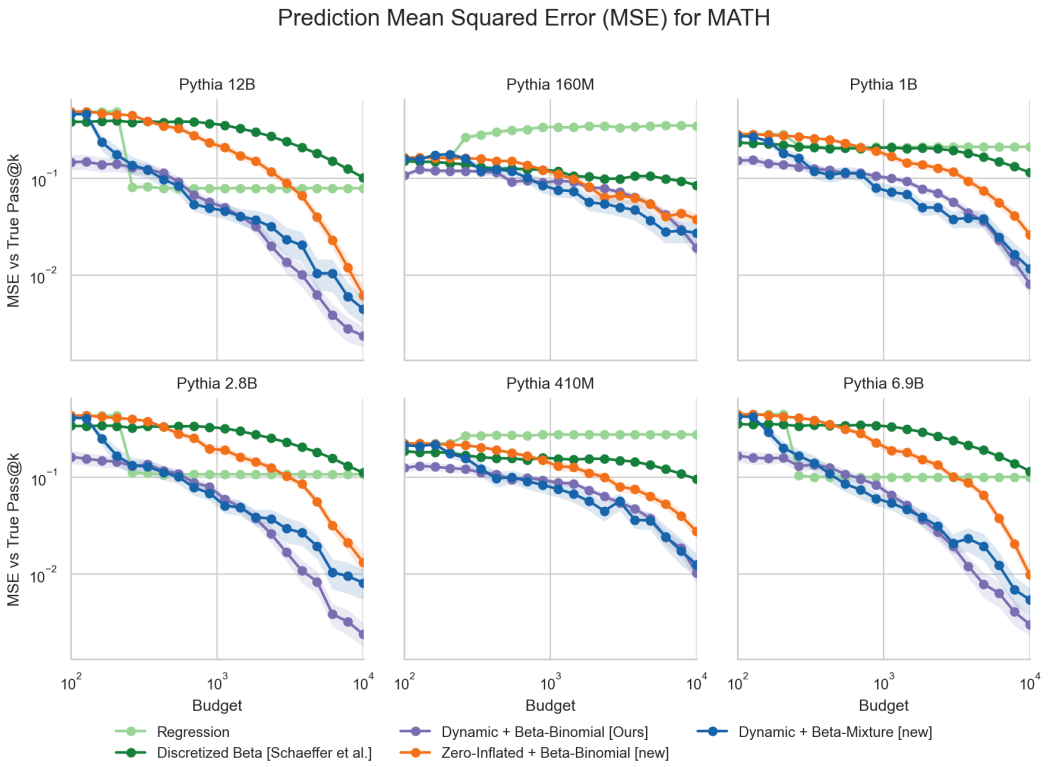

Figure 19: MSE plots comparing the performance of the mixed continuous-discrete model with a discrete lump of probability at $pass@1 = 0$ (dynamic sampling) to other methods on math data.

J Schaeffer et al., Comparison Experiments

We compare the estimators [26] under uniform and dynamic sampling. Each success probability is drawn from a Beta or Kumaraswamy distribution with shape (a, b), then multiplied by s. All 32 settings use $k = 1000$ and 30 seeds; shading shows 95% bootstrap intervals for mean squared error.

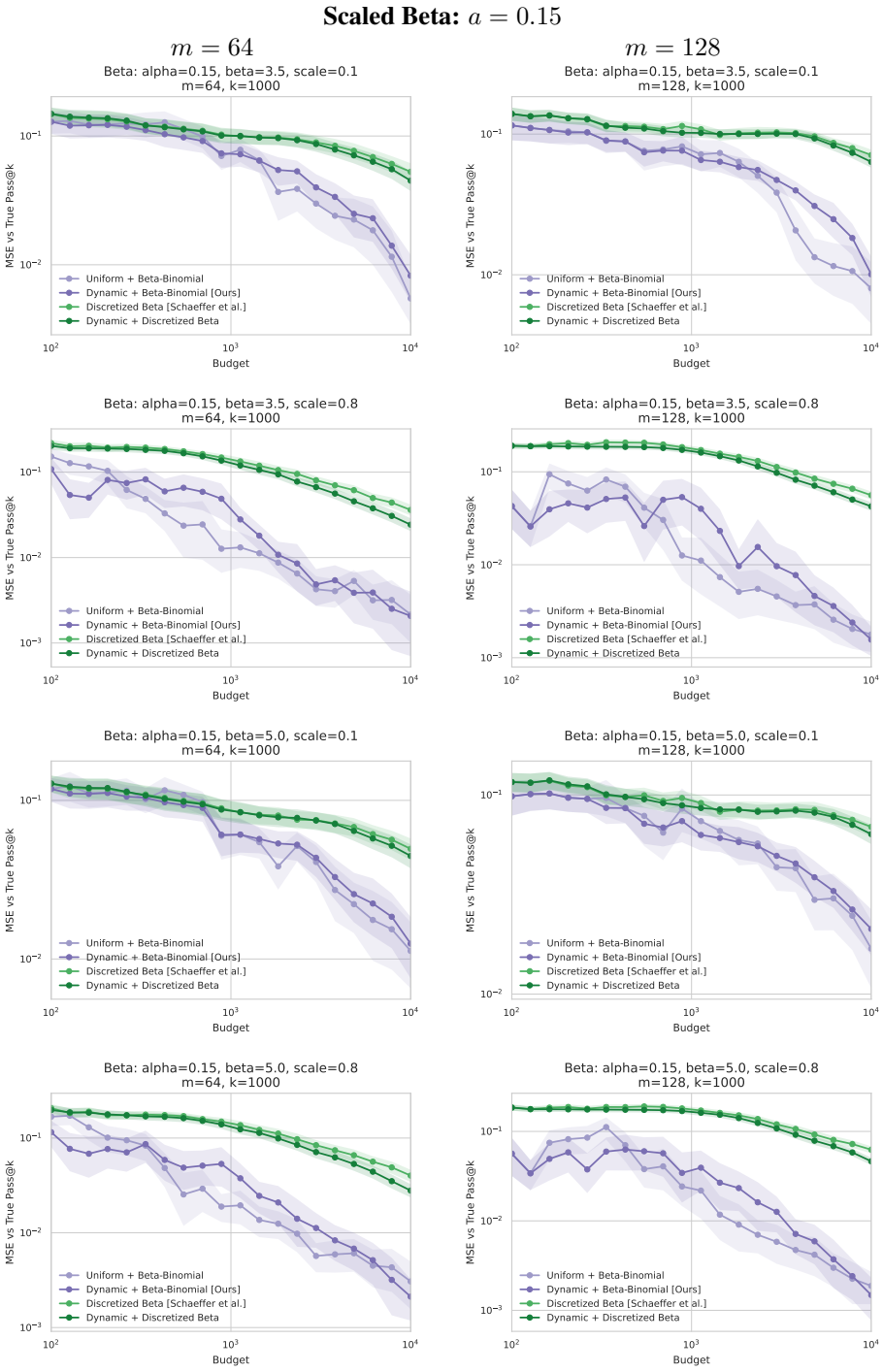

Figure 20: MSE comparisons for scaled Beta success probabilities with $a = 0.15$ at $k = 1000$. Columns: $m = 64$ and $m = 128$ problems. Rows, from top to bottom: $(b, s) = (3.5, 0.1)$, $(3.5, 0.8)$, $(5.0, 0.1)$, and $(5.0, 0.8)$.

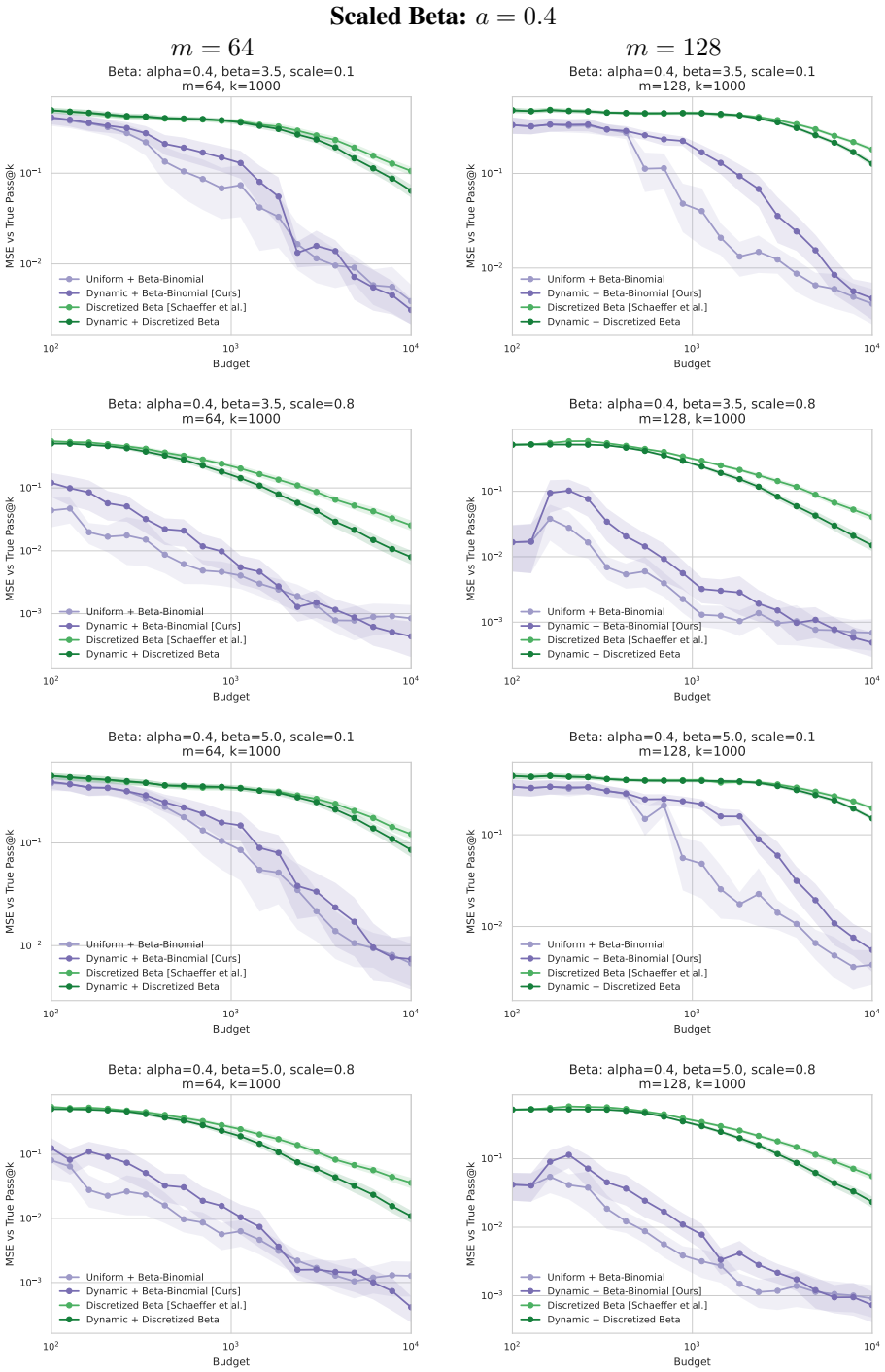

Figure 21: MSE comparisons for scaled Beta success probabilities with $a = 0.4$ at $k = 1000$. Columns: $m = 64$ and $m = 128$ problems. Rows, from top to bottom: $(b, s) = (3.5, 0.1)$, $(3.5, 0.8)$, $(5.0, 0.1)$, and $(5.0, 0.8)$.

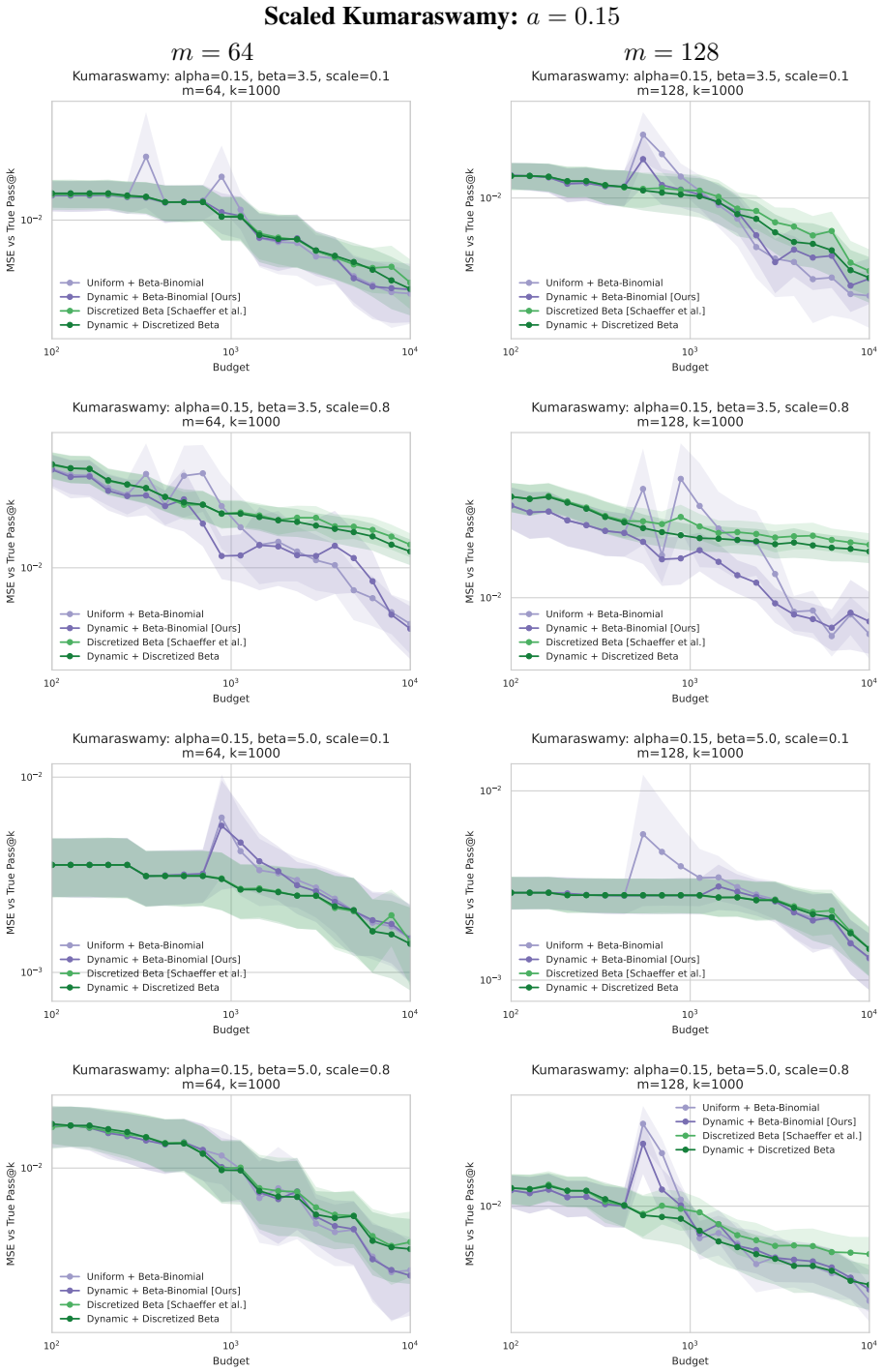

Figure 22: MSE comparisons for scaled Kumaraswamy success probabilities with $a = 0.15$ at $k = 1000$. Columns: $m = 64$ and $m = 128$ problems. Rows, from top to bottom: $(b, s) = (3.5, 0.1)$, $(3.5, 0.8)$, $(5.0, 0.1)$, and $(5.0, 0.8)$.

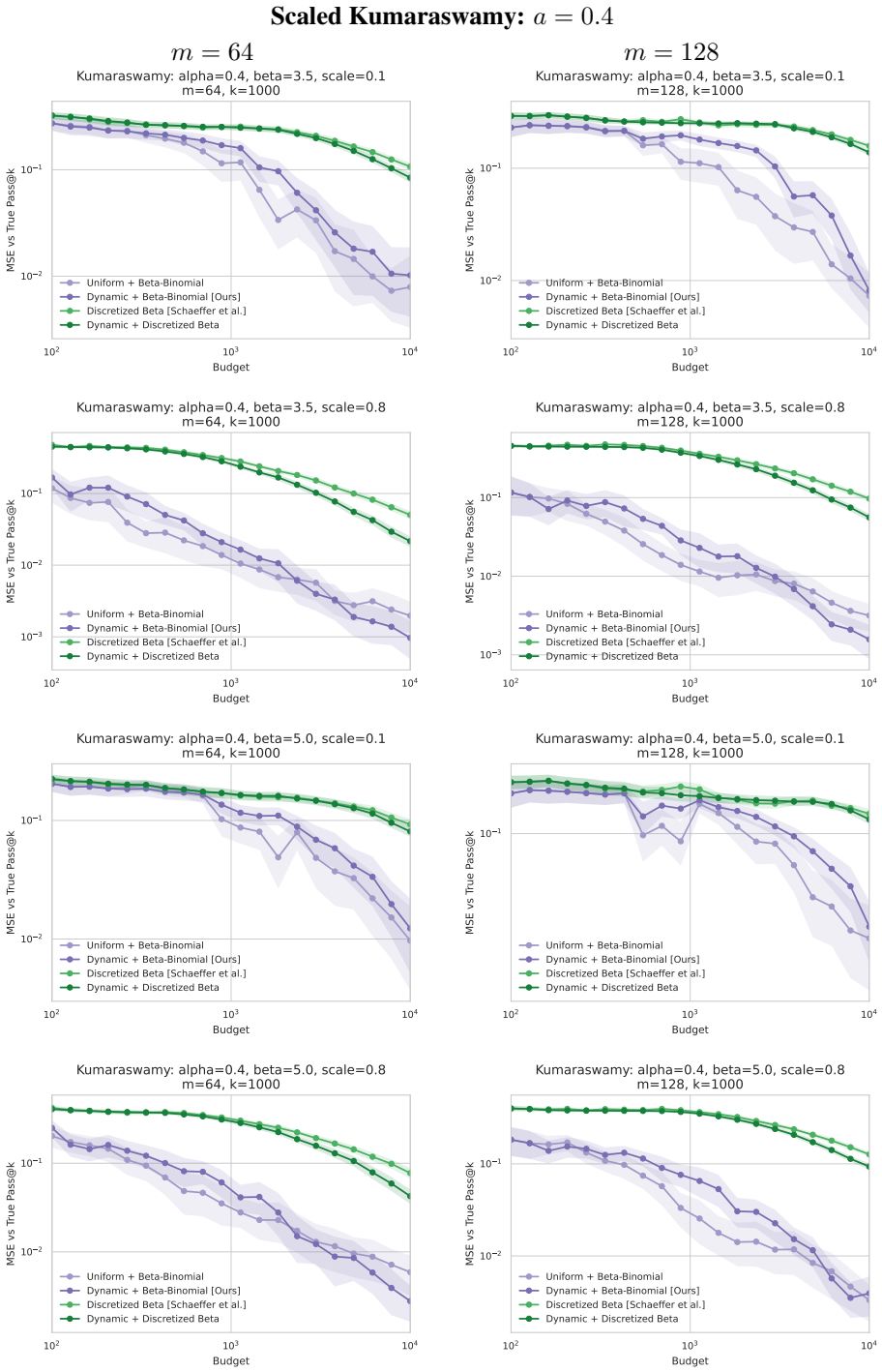

Figure 23: MSE comparisons for scaled Kumaraswamy success probabilities with $a = 0.4$ at $k = 1000$. Columns: $m = 64$ and $m = 128$ problems. Rows, from top to bottom: $(b, s) = (3.5, 0.1)$, $(3.5, 0.8)$, $(5.0, 0.1)$, and $(5.0, 0.8)$.

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: As claimed, we test our proposed $\text{pass@}k$ extrapolation in the sample budget range $[10^1, 10^3]$ and show capacity to predict well across orders of magnitude.

   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: We include a limitations section in the appendix.

   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: Yes. Proofs with assumptions are included in the appendix.

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: Yes, we provide a python package with a one-line script to reproduce results as well as an environment specified by a Docker image.

   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: Yes, see previous.

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: Important details are provided in the paper. The remainder necessary to reproduce results are present in the provided code.

   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: Yes. Error bars are provided for each of the non-synthetic $\mathrm{pass@}k$ predictions.

   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 require 2-4 hours to complete. This is specified in the code. Gathering samples requires much greater time and compute, but we provide this as part of the package.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the code of ethics to confirm.

   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: [Yes]

    Justification: We provide an impacts statement in the appendix and highlight the importance of extrapolated repeated sampling performance throughout the paper.

    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: The paper poses no such risks.

    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: [N/A]

    Justification: We gathered data and designed models ourselves.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: Yes. The code is well-organized, documented, and legible.

    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: We do not use crowdsourcing or perform research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We do not use crowdsourcing or perform research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

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.