Conformal Prediction Sets for Deep Generative Models via Reduction to Conformal Regression

Hooman Shahrokhi[*1] **Devjeet Roy**[*1] **Yan Yan**[1] **Venera Arnaoudova**[1] **Jana Doppa**[1]

[1]School of EECS , Washington State University , Pullman, WA, USA

Abstract

We consider the problem of generating valid and small prediction sets by sampling outputs (e.g., software code and natural language text) from a black-box deep generative model for a given input (e.g., textual prompt). The validity of a prediction set is determined by a user-defined binary admissibility function depending on the target application. For example, requiring at least one program in the set to pass all test cases in code generation application. To address this problem, we develop a simple and effective conformal inference algorithm referred to as *Generative Prediction Sets (GPS)*. Given a set of calibration examples and black-box access to a deep generative model, GPS can generate prediction sets with provable guarantees. The key insight behind GPS is to exploit the inherent structure within the distribution over the minimum number of samples needed to obtain an admissible output to develop a simple conformal regression approach over the minimum number of samples. Experiments on multiple datasets for code and math word problems using different large language models demonstrate the efficacy of GPS over state-of-the-art methods.

1 INTRODUCTION

Conditional deep generative models (DGMs) have recently become the dominant paradigm in a wide range of machine learning problems arising in domains including natural language processing, computer vision, and software engineering. Notable examples of DGMs include large language models (LLMs) (Brown et al. [2020]), diffusion models (Yang et al. [2023]), and vision transformers (Khan et al. [2022]). While these DGMs have displayed increasingly powerful predictive performance across a variety of generative tasks, they are far from infallible. For example, LLMs are known to produce hallucinations (Rawte et al. [2023]). Such erroneous outputs pose challenges in their use in safety-critical domains including medicine, law, education, and finance. For example, in many modern IDEs, users are able to query a LLM to generate programs Asare et al. [2023]. It would be desirable to attach to this set of suggested outputs a measure of uncertainty. This motivates the need for theoretically-sound uncertainty quantification for deep generative models to deploy them safely.

Conformal prediction (CP) (Vovk et al. [2005], Lei and Wasserman [2014]) is a general uncertainty quantification framework to produce prediction sets with theoretical guarantees for any black-box predictive model. Given some calibration data and a user-specified significance level α, CP generates sets which are guaranteed to contain the correct output with probability $1 - \alpha$ (referred to as *coverage*) through a calibration step. However, in the absence of an ordering over outputs, CP requires an enumeration of the entire output space, making it difficult to apply to deep generative models with combinatorial output spaces.

There is little work on this challenging problem setting. Conformal Language Modeling (CLM) Quach et al. [2024], constructs prediction sets iteratively by examining each sampled output. It calibrates three parameters $(\lambda_1, \lambda_2, \lambda_3)$: λ_1 determines if enough high-quality samples have been collected while λ_2 and λ_3 filter the collected samples based on quality criteria (e.g. total log probability of the samples). CLM uses Learn-Then-Test (Angelopoulos et al. [2021]) to find valid configurations of these parameters that provide coverage at the user-specified confidence level α with a high probability. However, it is not always possible to find valid configurations for tight confidence levels. Any calibration algorithm operating with finite sampling is fundamentally constrained by the base model's capacity to generate correct outputs within this finite horizon. To maintain distribution-free coverage when users request prediction sets at α levels exceeding the base model's abstention rate, we must output

[*]These authors contributed equally to the work.

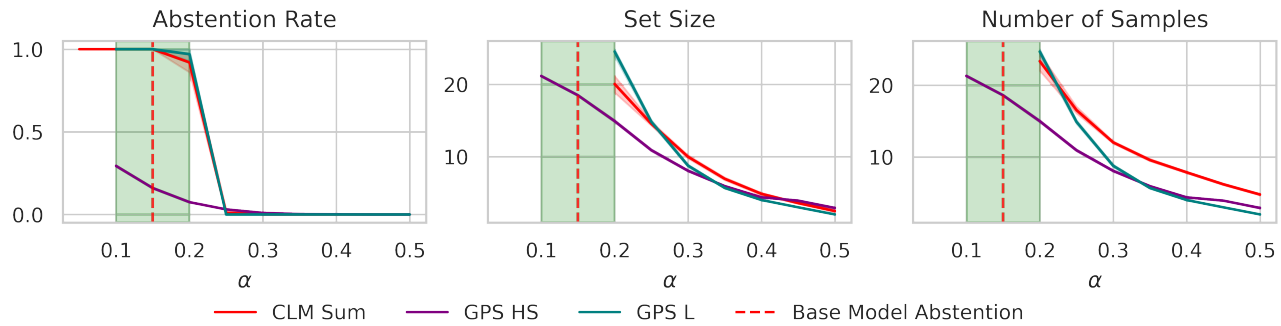


Figure 1: Abstention rate, set sizes, and number of samples collected from the GPT-4o-mini on the MATH benchmark for CLM and two variants of GPS. Since CLM provides PAC-style guarantees, we adjust the α level of GPS to equate the coverage guarantee. GPS HS produces valid prediction sets over tight α levels where CLM does not (green shaded area).

the entire space $\mathcal{Y}$ for at least some inputs (referred to as *abstention*). We can see this phenomenon in Figure 1.

On the MATH benchmark, GPT-4o-mini produces a correct output within the first 25 samples approximately 85% of the time (marginally on the joint distribution). The dashed red line shows this base abstention rate. In practice, CLM's performance is even more constrained – it abstains for 100% of our data until $\alpha = 0.2$, and does not achieve a 0% abstention rate until $\alpha = 0.25$. This gap emerges because CLM's filtering parameters can incorrectly reject correct solutions with non-zero probability, effectively inflating the model's abstention rate. Moreover, CLM's abstention is binary: for a given calibration dataset, CLM either produces a valid configuration, in which case it *never* abstains, or fails to produce a valid configuration, in which it *always* abstains. As a consequence of these limitations, 1) we need to collect more samples to produce valid prediction sets, and 2) if α is close enough to the given DGM's abstention rate, the calibration algorithm must abstain with non-zero probability.

In this work, we present a simple conformal calibration algorithm, referred to as *Generative Prediction Sets (GPS)*, to address these limitations. GPS reduces the problem of constructing prediction sets for DGMs to a conformal regression problem: we learn an auxiliary predictor to estimate how many samples the model needs to produce a correct output, and employ conformal prediction (CP) tools Romano et al. [2019] to obtain valid prediction intervals around these estimates. This interval determines the number of samples we collect at test time to construct prediction sets in the original output space. A basic instantiation of GPS (`GPS L` in Figure 1) uses only the prompt's log probability as input features and can achieve similar abstention rates, set sizes, and sample efficiency as CLM without examining the sampled outputs by calibrating just a single parameter using vanilla CP. Thus, it is both simpler to implement in practice and more computationally efficient. If we incorporate richer signals from the underlying DGM (e.g., hidden state activations for the input prompt), the resulting variant (`GPS HS` in Figure 1) can achieve dramatically lower abstention rates and valid prediction sets at α levels closer to the base model's true abstention rate, while maintaining set sizes competitive with CLM at higher α. In our empirical evaluations, we show that these benefits hold across a wide range of tasks including math, code and natural language tasks, and across a diverse set of base LLMs. We will see later that lower abstention rates are due to GPS' ability to *selectively* abstain on specific inputs by specifying a stopping rule that exceeds our fixed sampling budget.

Contributions. The key contribution is the development and evaluation of the generative prediction sets framework.

- Develop a provable CP method to produce valid prediction sets from a given deep generative model.
- GPS is the first CP method for deep generative models that prescribes a stopping rule for sampling without requiring access to samples from the model at test time.
- Empirical evaluation on multiple benchmark datasets with diverse LLMs over text and code.

2 BACKGROUND AND PROBLEM SETUP

This section provides background and describes the formal problem setup. We use capital letters (X, Y) for random variables and their lowercase (x, y) for specific values.

2.1 CONFORMAL PREDICTION BACKGROUND

CP is a general framework for constructing prediction sets with valid coverage guarantees Vovk et al. [2005], Lei and Wasserman [2014]. Split CP is the most widely used variant of CP which works as follows. Given a calibration dataset $\{(X_i, Y_i)\}_{i=1}^n$ (e.g., image and class label pairs), a predictive model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ (e.g., neural network classifier), a significance level $\alpha \in (0, 1)$, and a new test input (X_{n+1}, Y_{n+1}) such that $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are *exchangeable*,

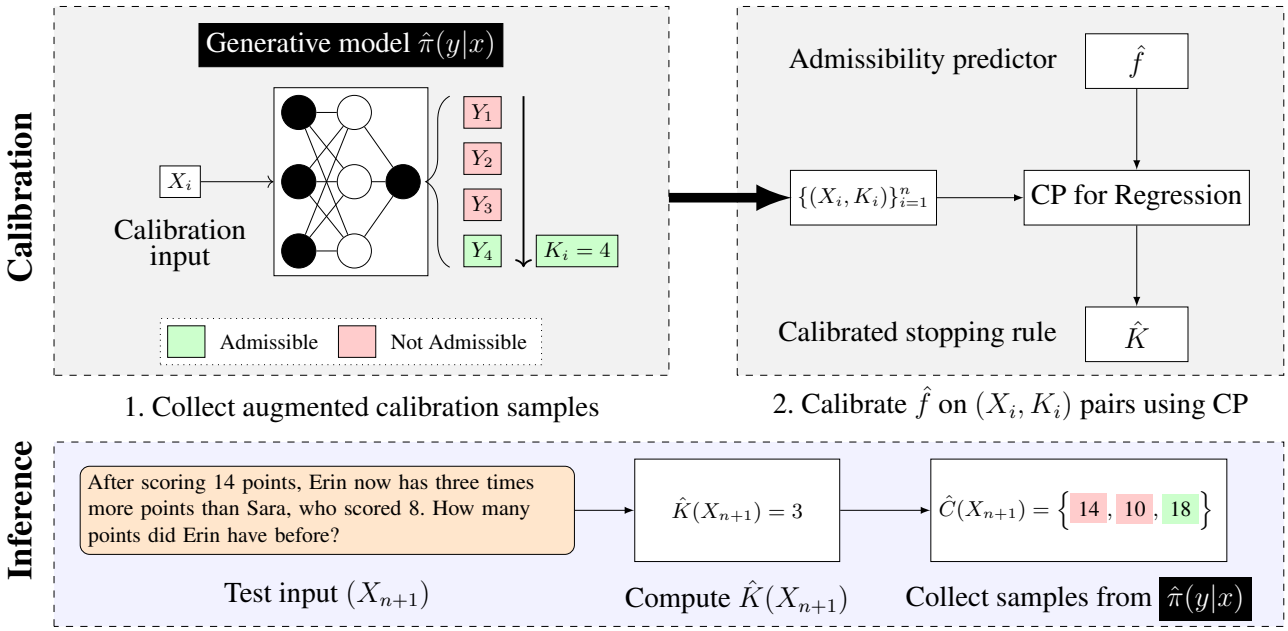


Figure 2: High-level illustration of the GPS framework: Calibration approach (top) and Inference procedure (bottom).

CP constructs a prediction set $\hat{C}(X_{n+1}) \subset \mathcal{Y}$, such that the set contains Y_{n+1} with $1-\alpha$ probability. A key component of CP is the *non-conformity score*, $\mathcal{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a heuristic measure of the degree to which a classifier's prediction conforms to a given input. For classification tasks, the output distribution of a classifier can be used as $\mathcal{S}(x, y)$ = $1 - \hat{f}(y|x), y = 1, \cdots, |\mathcal{Y}|$. Let $S_i = \mathcal{S}(X_i, Y_i)$ denote the non-conformity score of the i-th calibration example and $\tau = Q_{1-\alpha}(\{S_i\}_{i=1}^n)$ is the conformal α-quantile (i.e., the $\lceil (1-\alpha)(n+1)/n \rceil$ quantile) of the empirical distribution of the scores. CP defines the prediction set as:

$$\hat{C}(X_{n+1}) = \{y \in \mathcal{Y} : \mathcal{S}(X_{n+1}, y) \leq \tau\} \tag{1}$$

Let P denote the joint distribution of all $n+1$ samples. CP then provides the following coverage guarantee Lei and Wasserman [2014]:

$$P\{Y_{n+1} \in \hat{C}(X_{n+1})\} \geq 1 - \alpha \tag{2}$$

If scores are almost surely distinct (e.g., scores are continuous), we can obtain an upper bound Romano et al. [2019]:

$$P\{Y_{n+1} \in \hat{C}(X_{n+1})\} \leq 1 - \alpha + \frac{1}{n+1} \tag{3}$$

The guarantees above are *distribution-free*, i.e., they hold for any data distribution P for any $\alpha \in (0, 1)$. They are also *marginal*, in the sense that the probability statements above are over the randomness of all $n+1$ inputs (both the calibration data and the test input), and not conditional on either the calibration data or the test input X_{n+1}.

2.2 CP FOR OPEN-ENDED GENERATIVE TASKS

The non-conformity score $\mathcal{S}(x, y)$ measures the distance between a base predictor's outputs $\hat{f}(x)$ and true labels y. For a calibration example (x, y), this score represents the minimum "radius" around $\hat{f}(x)$ that encompasses the true label y. This interpretation is straightforward in regression for the typical score $\mathcal{S}(x, y) = |y - \hat{f}(x)|$. At test time, computing the appropriate quantile τ of these distances allows us to construct sets containing all y within τ-distance of $\hat{f}(x)$. This approach relies critically on our ability to enumerate all possible outputs within a given radius of $\hat{f}(x)$. For tasks with bounded or ordered output spaces, such as multiple-choice questions using LLMs, this enumeration is feasible and the CP procedure can be directly applied (Kumar et al. [2023]). However, open-ended generative tasks involve combinatorial output spaces that have neither of these properties. In such cases, we can only access a restricted subset of the output space $\mathcal{Y}' \subset \mathcal{Y}$ (such as M samples from a given generative model), which introduces two key challenges:

Sample coverage: We cannot guarantee that other valid outputs $y \in \mathcal{Y}$ don't exist outside our restricted sample space $\mathcal{Y}'$. The vanilla CP process fails to account for the sampling procedure during calibration, as it computes scores $\{S_i\}$ without considering whether true labels appear in the accessible subspace of $\mathcal{Y}$ or not. To address this challenge, we can modify our score to explicitly depend on the samples $(y'_1, \ldots, y'_\tau)$ for some stopping rule τ. One example is:

$$\tilde{\mathcal{S}}(x, (y'_1, \ldots, y'_\tau), y) = \begin{cases} \mathcal{S}(x, y), & \text{if } y \in (y'_1, ..., y'_\tau) \\ \infty \end{cases} \tag{4}$$

In practice, we typically require τ to be bounded. Importantly, for any generative model under finite sampling, there exists a minimum achievable error rate α^*. For any target $\alpha < \alpha^*$, the only way to maintain valid coverage is to output the entire space $\mathcal{Y}$ on some inputs – effectively abstaining on this part of the data distribution.

Semantic equivalence: Generative model outputs often have multiple semantically equivalent "correct" solutions. For example, different valid sorting algorithms could solve the same programming task. This can lead to situations where our ground truth label, y, might not be accessible in $\mathcal{Y}'$, but there might still be a $y' \in \mathcal{Y}'$ that is semantically equivalent to y. To handle this challenge, Quach et al. [2024] introduce a binary admissibility function $\mathcal{A}(x, y)$ that evaluates correctness independent of specific reference labels which we adopt in our work also.

2.3 PROBLEM SETUP AND CLOSEST WORK

Problem setup. We are now ready to formally describe our problem setting. We are given a conditional generative model $\hat{\pi}(\cdot|X)$ over the space $\mathcal{X} \times \mathcal{Y}$, and calibration data $\mathcal{D}_X$=$\{(X_i)\}_{i=1}^n$ drawn independently from an unknown distribution P_X. We assume that the output space $\mathcal{Y}$ is non-enumerable and unordered. Let $\mathcal{A} : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ denote an admission function that measures the admissibility of a solution $Y \sim \hat{\pi}(Y|X)$. For example, in a code generation task, $\mathcal{A}$ could be a function that checks if a generated program passes all test cases. Given a test input X_{n+1}, our goal is to generate a prediction set $\hat{C}(X_{n+1}) \subset \mathcal{Y}$ such that the set contains at least one admissible output Y with high probability. Formally, for a specified significance level $\alpha \in (0, 1)$, we want to provide the following guarantee:

$$P\{\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1\} \geq 1 - \alpha \quad (5)$$

Our goal is to achieve this guarantee in a finite-sample, distribution-free setting, while minimizing the cost (e.g., number of samples or API usage cost) to generate $\hat{C}(X_{n+1})$.

Conformal language modeling. CLM Quach et al. [2024] is the closest prior work to ours. It constructs prediction sets with a different type of guarantee than vanilla CP. Specifically, given parameters $\alpha, \delta \in (0, 1)$ and test input X_{n+1}, CLM constructs a prediction set $\hat{C}(X_{n+1})$ that satisfies:

$$P(P(\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1|\mathcal{D}_{\text{cal}}) \geq 1 - \alpha) \geq 1 - \delta \quad (6)$$

Here, the inner probability is over draws of X_{n+1} and the outer probability is over draws of the calibration dataset $\mathcal{D}_{\text{cal}}$. This nested probability structure makes direct comparisons between CLM and CP-based methods challenging, as we discuss in Section B. CLM consists of three key components: a set confidence estimator $\mathcal{F}$, a sample quality estimator $\mathcal{Q}$, and a sample similarity function $\mathcal{S}$, each parameterized by a threshold λ. The algorithm iteratively builds prediction sets by generating samples from the generative model and adding them to the set if they meet both quality and diversity thresholds (using $\mathcal{Q}$ and $\mathcal{S}$ respectively). This process continues until the set quality threshold is met, as determined by $\mathcal{F}$. The thresholds $(\lambda_1, \lambda_2, \lambda_3)$ that control risk for each component are determined using the Learn-Then-Test framework Angelopoulos et al. [2021].

3 GENERATIVE PREDICTION SETS

This section describes our *Generative Prediction Sets* (GPS) framework to construct prediction sets from a given DGM.

3.1 OVERVIEW AND ADVANTAGES OF GPS

Overview of GPS. GPS constructs adaptive prediction sets with valid coverage guarantees as illustrated in Figure 2 (see Algorithm 1 for a complete pseudocode). The core idea behind GPS is to reformulate the prediction set construction problem into estimating the number of samples needed to achieve admissible outputs, operating in two key phases.

Algorithm 1 Generative Prediction Sets (GPS) Framework

Input: generative model $\hat{\pi}(\cdot|X)$, calibration data $\mathcal{D}_X = \{X_i\}_{i=1}^n$, admissibility function $\mathcal{A}$, admissibility estimator $\hat{f}$, sampling budget M, test input X_{n+1}
Output: prediction set $\mathcal{C}(X_{n+1})$

1: *// Construct augmented calibration data*
2: Initialize $\mathcal{D}_{\text{cal}} = \varnothing$
3: **for** each $X_i \in \mathcal{D}_X$ **do**
4: $\quad\hat{K}_i \leftarrow$ minimum number of samples to obtain admissible output within the maximum budget M
5: $\quad$If no admissible output, then $\hat{K}_i \leftarrow M + 1$
6: $\quad\mathcal{D}_{\text{cal}} \leftarrow \mathcal{D}_{\text{cal}} \cup \{(X_i, \hat{K}_i)\}$
7: **end for**
8: *// Calibration on minimum number of samples*
9: Execute CP for regression using $\mathcal{D}_{\text{cal}}$ and estimator $\hat{f}$
10: *// Conformal inference on minimum number of samples*
11: Use conformal regressor on X_{n+1} to get $\hat{K}(X_{n+1})$ as per Eq. 10 (i.e., upper-bound of prediction interval)
12: *// Generate $\hat{K}(X_{n+1})$ samples to create prediction set*
13: $\mathcal{C}(X_{n+1}) \leftarrow \{Y_j \sim \hat{\pi}(\cdot|X_{n+1})\}_{j=1}^{\hat{K}(X_{n+1})}$
14: **return** Prediction set $\mathcal{C}(X_{n+1})$

First, GPS creates an augmented calibration dataset by sampling multiple outputs from the given generative model for each calibration input X_i in $\mathcal{D}_X$ (steps 1-7 in Algorithm 1). For each X_i, we sample from $\hat{\pi}(Y|X)$ using an arbitrary, independent sampling procedure to obtain a sequence of samples $\{Y_{ij}\}_{j=1}^M$ up to a maximum budget M. We define K_i to be the number of samples needed to obtain an admissi-

ble output from $\hat{\pi}$ for input X_i. This gives us an augmented calibration dataset, $\{(X_i, K_i)\}_{i=1}^n$.

Second, GPS develops a conformal procedure around this augmented data (steps 8-7 in Algorithm 1). Notice that if we construct a prediction interval around K_i, we will achieve our desired coverage guarantee due to the following equivalence of events:

$$\{K_i \leq \hat{K}_i\} \Leftrightarrow \left\{ \sum_{j=1}^{\hat{K}_i} \mathcal{A}(X_i, Y_{ij}) > 0 \right\} \tag{7}$$

At test time, given a new input X_{n+1}, we predict K_{n+1} (using the upper bound of the conformal interval) and sample from $\hat{\pi}(Y|X)$ exactly K_{n+1} times to construct the prediction set $\mathcal{C}(X_{n+1})$, achieving the guarantee stated in Eq. (5).

Key advantages of GPS. GPS has two qualitative benefits over CLM. First, GPS reduces the problem of generating prediction sets for DGMs to a vanilla CP problem. This reduction allows us to bring to bear the full body of machinery developed in vanilla CP to this problem setting. For example, the approximate conditional coverage method of Gibbs et al. [2023] can be applied off-the-shelf to GPS. Achieving such conditional coverage guarantees with CLM would require a non-trivial extension of the LTT framework. Second, since GPS calibrates only a single parameter, it has substantially lower computational complexity than CLM. Lastly, GPS works in batch mode; it specifies how many samples are to be collected from the DGM apriori (see Figure 2). In contrast, CLM works sequentially, one sample at a time. Thus, GPS can easily be used in modern batch inference pipelines, where it is desirable to produce all samples at once in a single batch to maximize hardware utilization, or in cost-sensitive applications, where users might want to make a trade-off between abstention rate and sampling cost.

3.2 CONFORMAL CALIBRATION ALGORITHM

A key challenge in designing calibration procedures for sampling-based regimes is handling cases where $\hat{\pi}$ fails to produce an admissible solution within M samples. Prior work in CP for generative models has addressed this by either setting non-conformity scores to ∞ (Su et al. [2024]) or assuming such failures don't occur (Wang et al. [2024]). However, if more than $\lceil (1-\alpha)(n+1) \rceil$ calibration examples lack admissible solutions, the conformal quantile $Q_{1-\alpha}(\{S_i\}_{i=1}^n)$ becomes ∞, yielding trivial prediction sets ($\mathcal{Y}$). GPS addresses this challenge via a different approach.

For calibration examples where no admissible solution is found within the sampling budget, we set $K_i = M + 1$, defining $K_i = \inf\{j : \mathcal{A}(X_i, Y_{ij}) = 1\} \cup \{M + 1\}$. This sentinel value $M + 1$ indicates the absence of an admissible solution in M samples. While the true K_i may be unknown in such cases, we know it exceeds M. This definition ensures finite K_i values, enabling the use of any standard regression-based non-conformity scores to calibrate $\{(X_i, K_i)\}_{i=1}^n$. This choice also allows us to handle abstention in a principled way.

To develop this calibration procedure, we first characterize the distribution of K_i. For our problem setting, the samples are collected using an I.I.D. sampling process, and thus, K_i follows a geometric distribution conditional on X_i, with its success probability an unknown function of X_i, i.e. K_i given $X_i = x$ follows Geom($f(x)$). While f is unknown, we can train an estimator $\hat{f}$ to estimate the success probability conditional on X_i, on a separate training split. We then use $\hat{f}$ to produce an estimate of the $1 - \alpha$ conditional quantile for K_{n+1}, being careful to account for finite sampling:

$$\hat{q}_\alpha(x) := \left\lceil \frac{\ln \alpha}{\ln\{1 - \hat{f}(x)\}} \right\rceil \tag{8}$$

While this quantile estimate is simply an estimate and will not yield a $1 - \alpha$ coverage in finite samples, we can conformalize it to obtain the valid, finite-sample coverage guarantees. For this, we use the Conformalized Quantile Regression (CQR) score from Romano et al. [2019]:

$$S(x, k) = \max\{0, k - \hat{q}_\alpha(x)\}$$

The score function is asymmetric, since we are building a one-sided interval of the form $[0, \hat{K}]$. Next, using we can construct a one-sided interval for K_{n+1} by applying the conformal adjustment term:

$$\hat{K}(X_{n+1}) = \hat{q}_\alpha(X_{n+1}) + Q_{1-\alpha}(\{S_i\}_{i=1}^n) \tag{9}$$

Recall that $\hat{Q}_{1-\alpha}$ is the conformal quantile over calibration scores. Now, because of our definition of K_i, $\hat{K}$ will always produce finite values, although they might exceed M. If this does happen, we must allow the model to abstain; it must output infinite sets. We can do this using the below scheme:

$$\hat{K}(X_{n+1}) = \begin{cases} \hat{K}(X_{n+1}) & \text{, if } \hat{K}(X_{n+1}) \leq M \\ \infty & \text{otherwise} \end{cases} \tag{10}$$

With the above $\hat{K}$, we have the below guarantee due to CP:

$$P\{K_{n+1} \leq \hat{K}(X_{n+1})\} \geq 1 - \alpha \tag{11}$$

This can then be translated back to the original output space, $\mathcal{Y}$, resulting in the following coverage guarantee for GPS:

Proposition 3.1 *Let $\{Y_j \sim \hat{\pi}(\cdot|X = X_{n+1})\}_{j=1}^{\hat{K}(X_{n+1})}$ be the prediction set generated according to Algorithm 1. Then we have the following coverage guarantee:*

$$P\{\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1\} \geq 1 - \alpha$$

This statement is a direct result of the equivalence of events stated in Eq. 7. We include a proof in the Appendix.

To understand $\hat{K}$'s abstention behavior, consider two extremes. When $\hat{f}$ accurately predicts $\hat{\pi}$'s abstentions, many calibration scores will be small, resulting in a modest conformal adjustment in Eq. 9. Thus, abstention decisions are primarily driven by $\hat{f}$'s input-specific predictions. Conversely, with a poor estimator $\hat{f}$ that outputs a constant probability (yielding $\hat{q}_\alpha = 1$), calibration examples where $\hat{\pi}$ abstains will have scores of M. With enough such examples, this forces $\hat{K} = M + 1$ universally, leading to constant abstention. Our calibration procedure thus adapts based on both $\hat{f}$'s quality and $\hat{\pi}$'s abstention rate.

We have now obtained a stopping rule for sampling, $\hat{K}(X_{n+1})$, without having to generate any samples from $\hat{\pi}(\cdot|X)$. At this point, the user has the option to abstain, if $\hat{K}$ is either ∞ or deemed too high for their usage scenario. To construct the prediction set from combinatorial space $\mathcal{Y}$, we take $\hat{K}(X_{n+1})$ samples from the generative model:

$$\hat{C}(X_{n+1}) = \{Y_j \sim \hat{\pi}(\cdot|X = X_{n+1})\}_{j=1}^{\hat{K}(X_{n+1})} \quad (12)$$

$\hat{C}(X_{n+1})$ contains only distinct outputs. Due to potential duplicate samples from the generative model, the cardinality of $\hat{C}(X_{n+1})$ may be less than or equal to $\hat{K}(X_{n+1})$.

Construction of admissibility estimator $\hat{f}$. While GPS guarantees valid prediction sets with the desired coverage level, the admissibility predictor $\hat{f}$ determines three key performance characteristics: empirical set sizes, sampling efficiency, and adaptivity of the prediction sets.

The construction of $\hat{f}$ requires collecting multiple samples from the underlying deep generative model using a separate training set. For each training example X_i, we obtain samples $\{Y_{ij}\}_{j=1}^{M_{\text{train}}}$, where M_{train} need not equal the calibration parameter M. The probability of success ($f(X_i)$) can be estimated, for example, using a MAP estimate:

$$p_i = \frac{\sum_{j=1}^{M_{\text{train}}} \mathcal{A}(X_i, Y_{ij}) + 1}{M_{\text{train}} + 2}, \quad (13)$$

yielding training data pairs (X_i, p_i).

Training $\hat{f}$ to predict success probabilities requires input signals derived solely from X_i. The primary options are log probabilities and latent space representations from the generative model. When log probabilities are inaccessible (as in API-based models), one can utilize either a separate embedding model or hidden states from a smaller surrogate model. In Section 4, we demonstrate this approach using both input log probabilities and hidden states from a smaller model (Phi 2) to train $\hat{f}$ for a larger models (GPT4o-mini). A constant predictor serves as a final alternative when no input signals are available—while preserving marginal validity, this yields fixed-size prediction sets.

4 EXPERIMENTAL RESULTS

In this section, we describe the experimental evaluation of our proposed GPS framework, comparison with state-of-the-art methods, and discuss results along different dimensions.

4.1 EXPERIMENTAL SETUP

Benchmark datasets. We employ data from three diverse tasks. *1) Code generation tasks:* we employ the MBPPAustin et al. [2021] and DS-1000 Lai et al. [2023] datasets to evaluate performance on code generation tasks. We use execution based functional correctness as our binary admission function. *2) Math problem-solving tasks:* We employ the GSM-8K Cobbe et al. [2021] and Math Hendrycks et al. [2021] datasets on math problem-solving tasks. *3) Natural language understanding tasks:* We employ a natural language question-answering dataset, Trivia-QA Joshi et al. [2017]. For both math problems and natural language understanding, we use exact match as the admissibility function. Since our baseline CLM wasn't able to provide any valid configurations DS-1000 for all $\alpha \in [0.1, 0.5]$, we defer results for DS-1000 to the Appendix.

Deep generative models. We consider three models across various parameter scales: Phi-2 Li et al. [2023], Llama 3 8B, and GPT-4o mini. For Math and DS-1000, we also add Gemma 2 27B Team et al. [2024], since Phi 2 and Llama 3 8B had low success rates for both of these tasks. Across all models, we use nucleus sampling Holtzman et al. [2019], with a sampling budget of 25. For space reasons, we only present results on GPT-4o mini in the main text, while the results for the other models are provided in the Appendix.

Configuration of GPS. We consider two predictors for the admissibility estimator $\hat{f}$. First, as a baseline, `GPS-L` uses a a linear regressor that directly predicts the probability of succes using the log probability of the input prompt. On the other hand,`GPS-HS` uses a feed-forward neural network that takes latent space representation of the prompt as input to predict the probability of success. Since GPT4o mini's hidden states aren't accessible, we use hidden state activations from Phi-2 as a surrogate. In the Appendix, we provide results for models with accessible hidden states.

CLM baseline. We employ CLM by adapting its code[1] as our baseline for our experiments, using normalized log probabilities as the quality score ($\mathcal{Q}$), and ROUGE-L as the similarity score ($\mathcal{S}$), and $\delta = 0.05$. For the set quality score, $\mathcal{F}$, we consider the four variants reported in the original paper by Quach et al. [2024]. These variants differ only in their choice of $\mathcal{F}$; `CLM First-K` uses the size of the prediction set, while `CLM Sum` and `CLM Max` use the sum and maximum of the quality scores of each sample in the set as the score. `CLM First-K` (nr) is the same as `CLM First-K`, but without any rejection rule (no rejection).

[1] https://github.com/Varal7/conformal-language-modeling

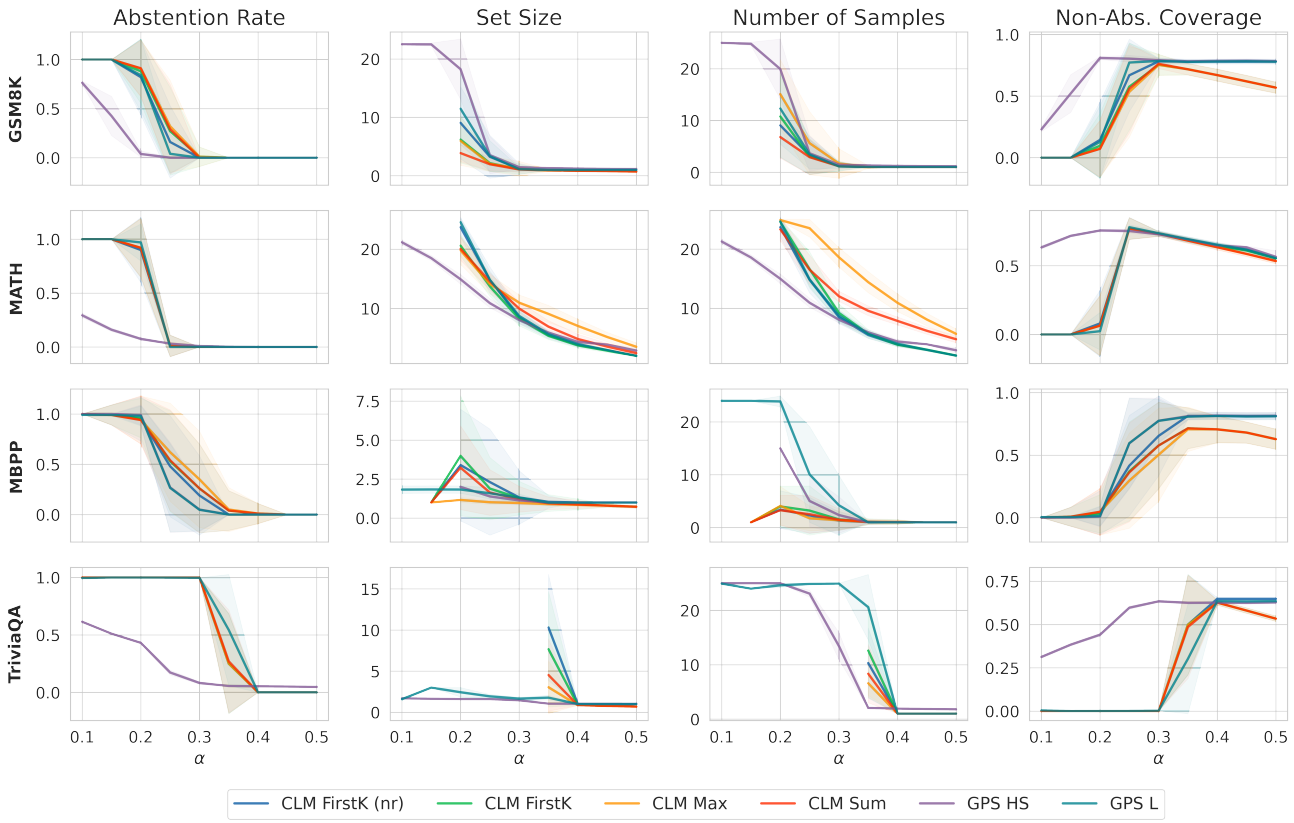


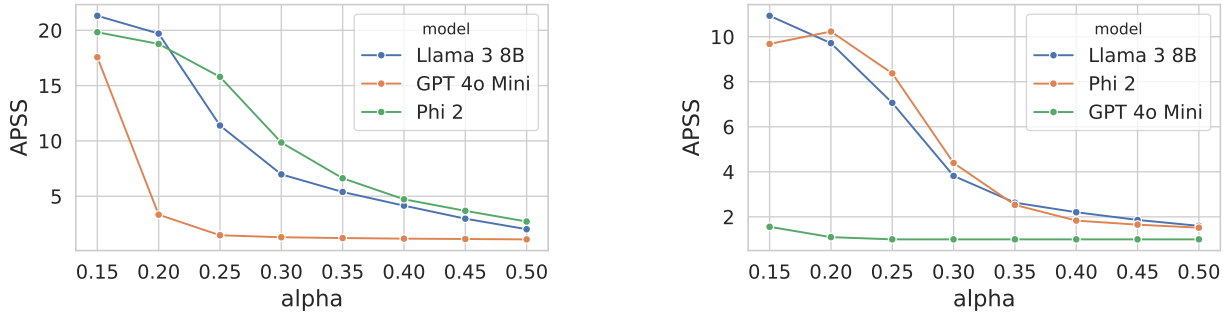
Figure 3: Results comparing GPS and CLM on different datasets (one row each for GSM8K, MATH, MBPP, and TriviaQA) using GPT-4o-mini LLM, across 100 trials, with shaded regions indicating standard deviation. Both `GPS` variants achieve lower abstention rates, and higher non-abstention coverage, while maintaining competitive set sizes with respect to CLM.

Evaluation methodology. We evaluate GPS and CLM across α ranging from 0.1 to 0.5 across 100 trials in increments of 0.05. This covers the spectrum of α values that are useful in practice. CLM's guarantees are qualitatively different from that of vanilla CP, so for a fixed α, the sets generated by CLM and `GPS` cannot be directly compared. However, we can use the fact that calibration conditional coverage of conformal predictors follows Beta distribution depending only on α and n, to find α_δ, the confidence level at which CP will achieve the same guarantee as CLM. Thus, we follow Quach et al. [2024] and set $\delta = 0.05$, and adjust α for GPS to ensure a fair comparison.

Metrics. We consider four metrics to measure on testing samples: **1)** average prediction set size, **2)** average number of samples generated, **3)** abstention rate, and **4)** non-abstention empirical coverage. As mentioned previously, the abstention rate allows us to determine the range of α on which a method can produce valid prediction sets. However, for methods such as GPS that can abstain selectively on certain inputs, the abstention rate is not sufficient; for a given α level, a method might simultaneously have a both low abstention and low coverage when it doesn't abstain, i.e. we are not gaining any coverage as a consequence of lowering the abstention rate. Thus, we also present the non-abstention coverage rate: the fraction of data on which a method a) does not abstain, i.e., outputs finite sized sets, and b) contains an admissible solution. Since both CLM and GPS achieve valid empirical coverage, we show those results in Appendix.

4.2 RESULTS AND DISCUSSION

Overall discussion. To summarize, empirical results in Fig 3 show that `GPS` can produce set sizes that are competitive with CLM, but at a wider range of α and with higher non-abstention coverage. We stress that `GPS` achieves such performance a) *without* ever examining the generated samples, b) using a simpler calibration algorithm that is easy to implement in practice, and c) in α-regimes where CLM simply doesn't produce valid configurations. We highlight our key observations from Figure 3 below: **1.** `GPS-L` and `GPS-HS` achieve higher non-abstention coverage than CLM variants, particularly for $\alpha \leq 0.3$. `GPS-HS` shows superior non-abstention coverage across all α except in MBPP, likely due to its small training set (150 examples) yielding a poor quality $\hat{f}$. On GSM8k (659 examples), `GPS HS` consistently outperforms all methods across all α in both



(a) GSM Benchmark

(b) MBPP Benchmark

Figure 4: Results for GPS-HS: average prediction set size (APSS) vs. coverage level α for (a) GSM8k and (b) MBPP benchmark datasets, using varying quality of models (Phi-2, Llama 3 8b, and GPT-4o-mini).

abstention rates and non-abstention coverage.

2. For GSM, MATH, and TriviaQA, GPS HS maintains abstention rate of < 1 and abstention coverage ≥ 0.45 even at $\alpha \approx 0.1$. This demonstrates $\hat{f}$'s effectiveness at selective abstention on difficult inputs, yielding valid configurations with non-trivial coverage. The predictor quality's impact is clear: GPS L uses same calibration but with a weaker $\hat{f}$.

3. GPS L and GPS HS require fewer samples to get prediction sets, especially on MATH. A narrow α range marking the transition between non-zero and zero abstention shows CLM requiring fewer samples than GPS, but with higher abstention rates and lower non-abstention coverage. The sole exception is GPS L on TriviaQA for $\alpha \in [0.35, 0.4]$, where the best CLM method achieves marginally higher non-abstention coverage with fewer samples.

4. Both GPS variants produce comparable set sizes to the best CLM method when abstention rates approach 0 ($\alpha \geq 0.3$). At tighter confidence levels, CLM yields smaller sets but with near-zero non-abstention coverage (e.g., GSM8k and MATH at $\alpha = 0.2$), while GPS L maintains non-abstention coverage ≥ 0.4. Set size differences diminish as abstention rates approach zero.

Set size vs. model quality. Fig 4a and 4b show GPS-HS set sizes across model performance levels for MBPP and GSM at different α (see Appendix for more results). Higher model pass rates (problems with at least one admissible solution) correlate with smaller generated sets, demonstrating GPS's ability to scale with model quality. Fluctuations at higher coverage levels ($\alpha \in \{0.15, 0.2\}$) stem from abstention rates, which are higher for Phi-2 than Llama on both datasets, reducing prediction set sizes at these levels.

5 RELATED WORK

Conformal prediction. In recent years, CP has emerged as an effective framework for uncertainty quantification across both regression Romano et al. [2019], Papadopoulos et al. [2002] and classification tasks Romano et al. [2020], Angelopoulos et al. [2020a], Sadinle et al. [2019]. Much of the recent focus in the CP community has been on reducing set sizes Romano et al. [2020], Angelopoulos et al. [2020b], Sadinle et al. [2019] and constructing CP procedures for specific usage contexts Huang et al. [2024]. Some works also focus on providing conditional coverage guarantees Gibbs et al. [2023], Jung et al. [2022], Tibshirani et al. [2019]. Conformal risk control extends CP to control generalized risk on prediction sets Angelopoulos et al. [2022], Bates et al. [2021]. The Learn-Then-Test Angelopoulos et al. [2021] is a framework for risk controlled prediction sets that does not require monotonic risks. It is utilized by CLM Quach et al. [2024] for generating prediction sets from LLMs.

Uncertainty estimation for LLMs. Uncertainty for LLMs has largely focused on calibrating output probabilities Jiang et al. [2021], Desai and Durrett [2020], Lin et al. [2023], Kuhn et al. [2023], verbalization Kadavath et al. [2022], and Bayesian approaches Malinin and Gales [2020], Ryabinin et al. [2021]. Deep classifiers can exhibit high levels of calibration error Jiang et al. [2021]. CP is recently utilized for generative models, especially LLMs, in settings with bounded output spaces, such as multiple choice question answering Kumar et al. [2023], Zhang et al. [2021], Rouzrokh et al. [2024], Li et al. [2024]. Lastly, another line of work focuses on selecting a subset of the generated output of an LLM with factuality guarantees Mohri and Hashimoto [2024], Cherian et al. [2024].

6 SUMMARY

This paper introduced GPS, a novel conformal prediction algorithm to produce marginally valid prediction sets for deep generative models with unbounded output spaces. Our score formulation allows GPS to selectively abstain on specific examples, allowing it to achieve lower abstention rates and higher non-abstention coverage at α-levels at which prior work always abstains. Our experiments show that GPS indeed achieves higher non-abstention coverage and lower abstention rates, while maintaining parity with CLM with regard to set sizes at higher α levels.

Acknowledgements. The authors gratefully acknowledge the in part support by USDA-NIFA funded AgAID Institute award 2021-67021-35344. The views expressed are those of the authors and do not reflect the official policy or position of the USDA-NIFA.

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020a.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Anastasios Nikolas Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction, 2020b. URL `https://arxiv.org/abs/2009.14193`.

Owura Asare, Meiyappan Nagappan, and N Asokan. Is github's copilot as bad as humans at introducing vulnerabilities in code? *Empirical Software Engineering*, 28(6): 129, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. August 2021. URL `http://arxiv.org/abs/2108.07732`.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. Large language model validity via enhanced conformal prediction methods. *arXiv preprint arXiv:2406.09714*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. October 2021. URL `http://arxiv.org/abs/2110.14168`.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. May 2023. URL `http://arxiv.org/abs/2305.12616`.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pages 18319–18345. PMLR, 2023.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76 (1):71–96, 2014.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. Traq: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3799–3821, 2024.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Pouria Rouzrokh, Shahriar Faghani, Cooper U Gamble, Moein Shariatnia, and Bradley J Erickson. Conflare: Conformal large language model retrieval. *arXiv preprint arXiv:2404.04287*, 2024.

Max Ryabinin, Andrey Malinin, and Mark Gales. Scaling ensemble distribution distillation to many classes with proxy targets. *Advances in Neural Information Processing Systems*, 34:6023–6035, 2021.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, 2005. URL https://link.springer.com/book/10.1007/978-3-031-06649-8.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint arXiv:2407.00499*, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Shiwei Zhang, Xiuzhen Zhang, Jey Han Lau, Jeffrey Chan, and Cecile Paris. Less is more: Rejecting unreliable reviews for product question answering. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pages 567–583. Springer, 2021.

Appendix

Hooman Shahrokhi[*1] **Devjeet Roy**[†1] **Yan Yan**[1] **Venera Arnaoudova**[1] **Jana Doppa**[1]

[1]School of EECS , Washington State University , Pullman, WA, USA

This Supplementary Material should be submitted together with the main paper.

[*]These authors contributed equally to the work.
[†]These authors contributed equally to the work.

A PROOFS

A.1 PROOF FOR PROP 3.1

Recall that $K_i = \inf\{j : \mathcal{A}(X_i, Y_{ij}) = 1\} \cup \{M+1\}$, and $\hat{C}(X_{n+1}) = \{Y_j \sim \hat{\pi}(\cdot|X = X_{n+1})\}_{j=1}^{\hat{K}(X_{n+1})}$. From the standard CP proof, we have that:

$$P\{K_{n+1} \leq \hat{K}(X_{n+1})\} \geq 1 - \alpha$$

Consider the event $\{K_{n+1} \leq \hat{K}(X_{n+1})\}$:

$$\{K_{n+1} \leq \hat{K}(X_{n+1})\} = \{K_{n+1} \leq \hat{K}(X_{n+1})\}\{\hat{K}(X_{n+1}) < \infty\} \cup \{K_{n+1} \leq \hat{K}(X_{n+1})\}\{\hat{K}(X_{n+1} = \infty\} \tag{14}$$

On $\{\hat{K}(X_{n+1}) = \infty\}$, $\hat{C}(X_{n+1}) = \mathcal{Y}$, and $\{\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1\}$.
On $\{\hat{K}(X_{n+1}) < \infty\}$:

$$\begin{aligned}
\{K_{n+1} \leq \hat{K}(X_{n+1})\} &= \Big\{\inf\{j : \mathcal{A}(X_{n+1}, Y_{n+1,j}) = 1\} < \hat{K}(X_{n+1})\Big\} \\
&= \Big\{\exists j < \hat{K}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y_{n+1,j}) = 1\ \Big\} \\
&= \Big\{\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1\Big\}
\end{aligned}$$

Substituting in Eq.14, we obtain:

$$\{K_{n+1} \leq \hat{K}(X_{n+1})\} = \{\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1\}$$

and plugging this into the probability statement gives us the desired coverage guarantee. □

B EQUATING THEORETICAL GUARANTEES BETWEEN CLM AND GPS

The marginal guarantee of conformal prediction Eq. 2 and CLM's guarantee Eq. 6 are qualitatively different. In this paper, whenever we mention α for comparing our method, it refers to α in Eq. 6, and the conformal value is adjusted accordingly. We use the calibration conditional coverage properties of CP to perform this adjustment.

Assuming non-conformity scores are almost surely distinct, the calibration coverage, conditional on the calibration set of a conformal predictor at level α_0 follows a Beta distribution (Angelopoulos et al. [2024]:

$$P\big\{Y_{n+1} \in C(X_{n+1}) \mid \mathcal{D}_{\text{cal}}\big\} \sim \text{Beta}(k(\alpha_0), n+1-k(\alpha_0)) \tag{15}$$

where $k(\alpha_0) := \lceil (1-\alpha_0)(n+1) \rceil$. Even if scores are not distinct, the miscoverage probability stochastically dominates the Beta$(k_\alpha, n+1-k_\alpha)$ distribution (Angelopoulos et al. [2024][Theorem 4.1]).

Following that, for any value of α, this conformal method has the following guarantee:

$$P\Big(P(\exists Y \in \hat{C}(X_{n+1}) : \mathcal{A}(X_{n+1}, Y) = 1 \mid \mathcal{D}_{\text{cal}}) \geq 1 - \alpha\Big) \geq 1 - \delta$$

where δ is the CDF of the Beta distribution with parameters from Theorem 15 at $1-\alpha$:

$$\delta = \text{BetaCDF}_{k(\alpha_0), n+1-k(\alpha_0)}(1-\alpha) \tag{16}$$

Thus, for any α and δ satisfying the LTT guarantee with the form of Eq. 6, we perform a grid search in the α_0 space to find values such that the δ in Eq. 16 is lower than the δ in Eq. 6. Table 5 shows the adjusted confidence levels, α_0, corresponding to the non-adjusted α-values in Figure 3.

α	DS1000 ($n = 200$)	GSM8k ($n = 330$)	MATH ($n = 1250$)	MBPP ($n = 113$)	TriviaQA ($n = 4486$)
0.10	0.0594	0.0662	0.0814	0.0526	0.0900
0.15	0.0990	0.1084	0.1277	0.0877	0.1381
0.20	0.1485	0.1566	0.1757	0.1228	0.1867
0.25	0.1881	0.1987	0.2236	0.1754	0.2352
0.30	0.2376	0.2469	0.2715	0.2105	0.2843
0.35	0.2772	0.2951	0.3194	0.2631	0.3337
0.40	0.3267	0.3433	0.3690	0.2982	0.3832
0.45	0.3762	0.3915	0.4185	0.3508	0.4331
0.50	0.4257	0.4397	0.4680	0.4035	0.4830

Figure 5: Table of α values and corresponding adjusted α_0 for the different tasks used in this paper.

C EXPERIMENTAL DETAILS

C.1 COMPUTE INFRASTRUCTURE AND SOFTWARE

All inference except for GPT-4o-mini is performed on a single node with 3 H100 GPUs (80GB VRAM), 8 CPUs and 400GB RAM. We utilize `vllm`[1] to generate samples, and perform a separate forward pass using Huggingface Transformers[2] to collect logits and hidden states. GPT-4o-mini inference is conducted using the OpenAI client library[3] for python.

C.2 DATASET DETAILS

All the datasets we use are publicly available from the Hugginface Hub. For GSM[4] we use the test split from `main` subset. For Math[5] we use the test split. For TriviaQA[6] we use the validation split in the `rca.nocontext` subset of the data. DS1000[7] has only a single split, and we use the entire dataset. For MBPP[8], we use both the train and test splits from the `sanitized` subset, due to the small number of examples in the data. Details regarding the size of each split and generation settings are shown in Table 6.

Dataset	Train Split Size	Test Split Size	Temperature	Top-p
DS1000	600	400	0.2	0.95
MBPP	150	227	0.2	0.95
TriviaQA	8972	8972	0.2	0.95
GSM8k	659	660	0.2	0.95
Math	2500	2500	0.6	0.95

Figure 6: Dataset splits and generation settings

C.2.1 Prompts

Here we list the prompts used for our experiments. Text enclosed in {...} refers to columns in the original dataset. teal colored text denotes the system prompt used by GPT-4o-mini. red colored text are only added for Gemma 2 27b.

Listing 1: Prompt for MBPP

[1]https://github.com/vllm-project/vllm
[2]https://huggingface.co/docs/transformers/en/index
[3]https://github.com/openai/openai-python
[4]https://huggingface.co/datasets/openai/gsm8k
[5]https://huggingface.co/datasets/EleutherAI/hendrycks_math
[6]https://huggingface.co/datasets/mandarjoshi/trivia_qa
[7]https://huggingface.co/datasets/xlangai/DS-1000
[8]https://huggingface.co/datasets/google-research-datasets/mbpp

```
Answer the following question. In your response, only write the raw code, do not use
    markdown and do not add explanations.

{description}
{test_list[0]}
```

Listing 2: Prompt for DS-1000

```
Answer the following question. In your response, only write the raw code, do not use
    markdown and do not add explanations
{prompt}
```

Listing 3: Prompt for GSM

```
Q: {question}
A: Let's think step by step.
```

Listing 4: Prompt for Trivia

```
Answer these questions:
Q: Which former major league baseball pitcher, known as The Big Unit, now pitches Geico?
A: Randy Johnson
Q: The Philippines were named after which king of Spain?
A: King Philip II
Q: US Vice-President Joe Biden represents which state?
A: DELAWARE
Q: Which, now defunct, political party was founded by Declan Ganley in April 2009?
A: Libertas Ireland
Q: Sept 30, 1966 saw the public unveiling of which popular model of Boeing aircraft?
A: 747
Q: {question}
A:
```

Listing 5: Prompt for Math

```
Answer the following questions. Your final answer should always follow the same format: '
    Final Answer: The final answer is [answer]. I hope it is correct.
Answer the following questions. Use the answer format provided in the examples. All of
    your latex expressions must be wrapped in $...$ (for example, to write 'x=2' as latex,
     write $x=2$).
Problem:
Find the domain of the expression  $\frac{\sqrt{x-2}}{\sqrt{5-x}}$.}

Solution: The expressions inside each square root must be non-negative. Therefore, $x-2 \
    ge 0$, so $x\ge2$, and $5 - x \ge 0$, so $x \le 5$. Also, the denominator cannot be
    equal to zero, so $5-x>0$, which gives $x<5$. Therefore, the domain of the expression
    is $\boxed{[2,5)}$.
Final Answer: The final answer is $[2,5)$. I hope it is correct.

Problem:
If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12,$ then find $\det (\mathbf{A} \mathbf{B
    }).$

Solution: We have that $\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B})
    = (2)(12) = \boxed{24}.$
Final Answer: The final answer is $24$. I hope it is correct.

Problem:
```

```
Terrell usually lifts two 20-pound weights 12 times. If he uses two 15-pound weights
    instead, how many times must Terrell lift them in order to lift the same total weight?

Solution: If Terrell lifts two 20-pound weights 12 times, he lifts a total of $2\cdot 12\
    cdot20=480$ pounds of weight.  If he lifts two 15-pound weights instead for $n$ times,
     he will lift a total of $2\cdot15\cdot n=30n$ pounds of weight.  Equating this to 480
     pounds, we can solve for $n$:
\begin{align*}
30n&=480\
\Rightarrow\qquad n&=480/30=\boxed{16}
\end{align*}
Final Answer: The final answer is $16$. I hope it is correct.

Problem:
If the system of equations

\begin{align*}
6x-4y&=a,\
6y-9x &=b.
\end{align*}has a solution $(x, y)$ where $x$ and $y$ are both nonzero,
find $\frac{a}{b},$ assuming $b$ is nonzero.

Solution: If we multiply the first equation by $-\frac{3}{2}$, we obtain

$$6y-9x=-\frac{3}{2}a.$$Since we also know that $6y-9x=b$, we have

$$-\frac{3}{2}a=b\Rightarrow\frac{a}{b}=\boxed{-\frac{2}{3}}.$$
Final Answer: The final answer is $-\frac{2}{3}$. I hope it is correct.

Problem: {problem}
Answer:
```

C.3 RESULTS

We provide extended evaluation results for GPS and CLM for $\alpha \in [0.05, 0.1, 0.15, .2, .25, .3, .35, .4, .45, .5]$ in the figures below. Additional models include Llama 3 8b, Phi-2, and 2 variants of Gemma 2 27b - the base model and the instruction tuned version. The figures below follow the same format as Figure 3.

For `GPS-L`, we use a linear regression model from `scikit-learn`, that takes as input the length normalized log probability of the input prompt under the base LLM. For `GPS-HS`, we use a multi-layer perceptron implemented in `torch` that takes hidden state activations of the last token in the input prompt from the last layer of either the underlying model in the case of Llama and Phi-2. For Gemma 2 and GPT-4o-mini, we use hidden state activations from the last token and layer of Phi-2. We perform a grid-search to determine the number of layers ([1, 2, 4]), number of hidden units ([256, 512, 1024]) and training epochs ([15, 20, 25]). Details are provided in the supplementary code.

C.3.1 Figures

We provide plots similar to Figure 3 for each model and dataset.

C.3.2 Tables

Here we provide a comprehensive set of tables including empirical coverage rates, in addition to the metrics considered prior. Blank entries (-) indicate complete abstention.

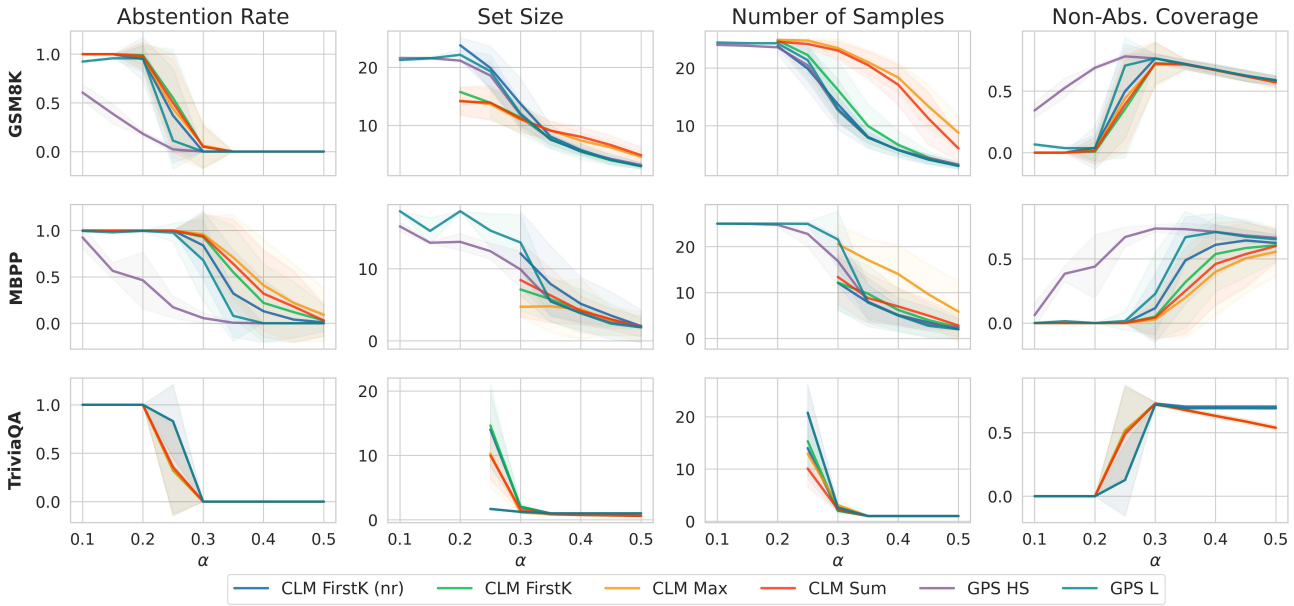


Figure 7: Llama 3 8b results on all datasets.

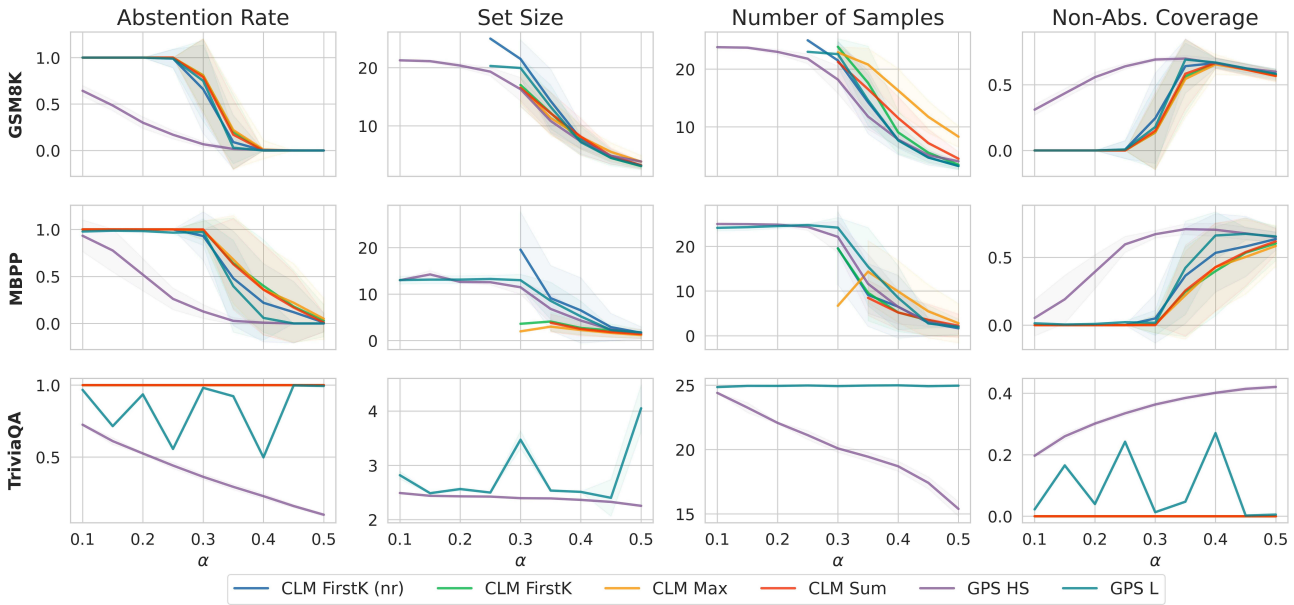


Figure 8: Phi-2 results on all datasets.

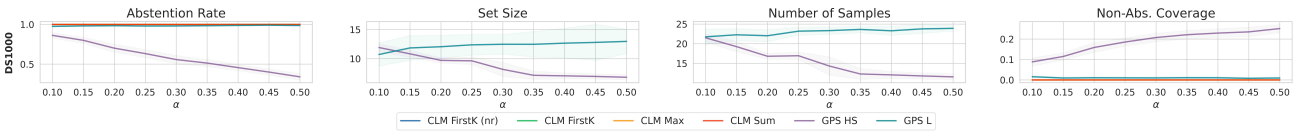


Figure 9: Gemma 2 27b results on DS-1000

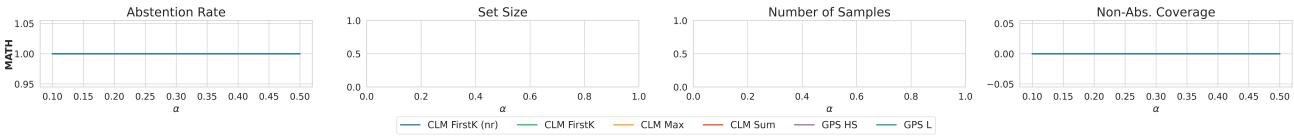


Figure 10: Gemma 2 27b (Instruction Tuned) on DS-1000

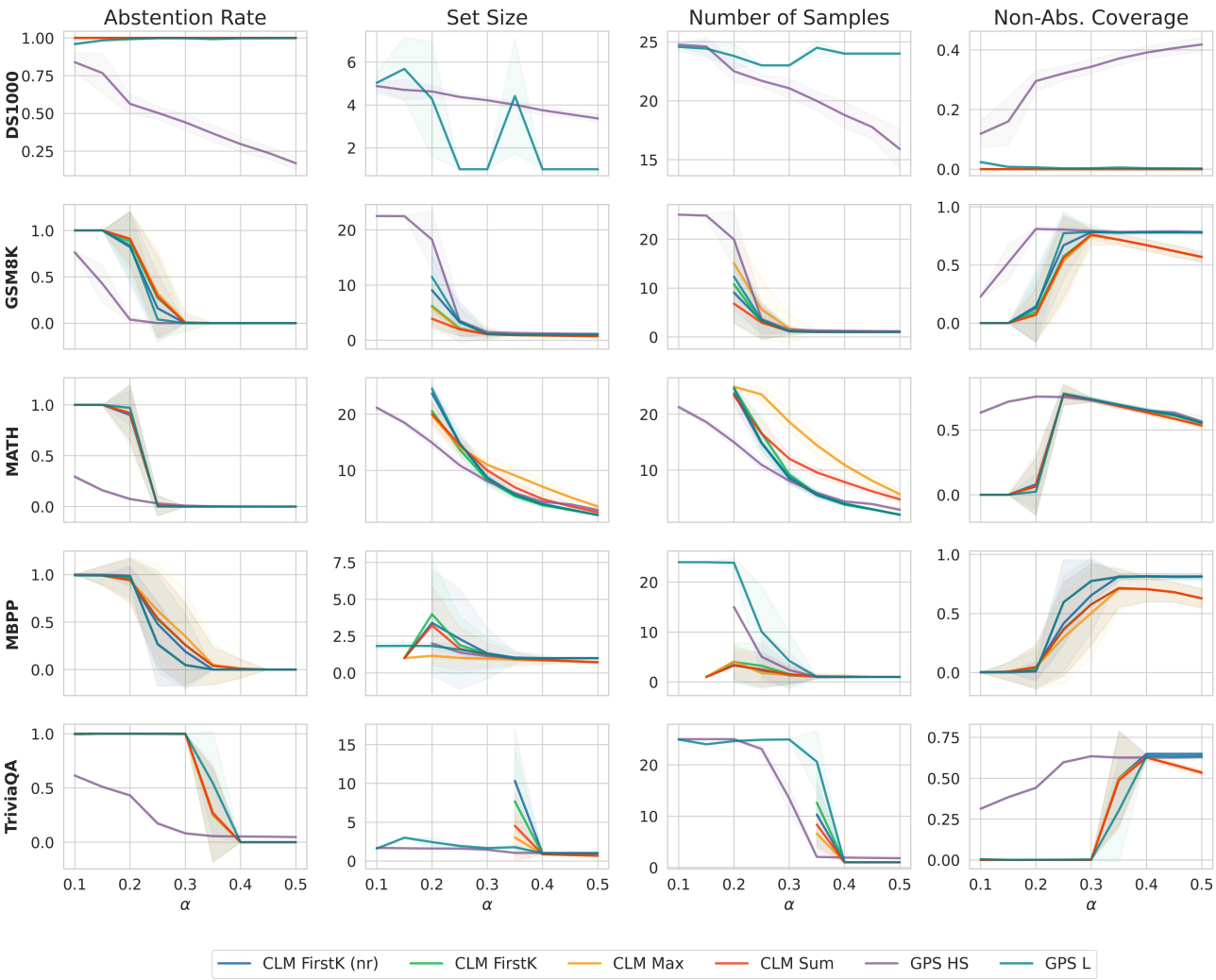


Figure 11: GPT-4o-mini results on all datasets.

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.95±0.02	**0.86±0.04**	**0.17±0.12**	**21.49±1.04**	11.90±0.65
	GPS L	0.99±0.00	0.98±0.01	0.12±0.15	21.75±1.96	**10.72±1.97**
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.91±0.03	**0.80±0.04**	**0.19±0.11**	**19.25±1.27**	**10.83±0.61**
	GPS L	0.99±0.00	0.98±0.01	0.12±0.16	22.27±1.55	11.84±2.10
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.86±0.03	**0.70±0.05**	**0.22±0.08**	**16.80±0.73**	**9.70±0.45**
	GPS L	0.99±0.00	0.98±0.01	0.12±0.15	22.03±1.49	12.05±1.98
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.82±0.04	**0.63±0.05**	**0.23±0.07**	**16.94±0.92**	**9.62±0.63**
	GPS L	0.99±0.00	0.98±0.01	0.12±0.15	23.19±1.02	12.37±1.76
0.300000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.76±0.04	**0.56±0.05**	**0.25±0.06**	**14.31±2.31**	**8.16±1.18**
	GPS L	0.99±0.00	0.98±0.01	0.12±0.15	23.33±0.93	12.47±1.67
0.350000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.73±0.04	**0.51±0.04**	**0.25±0.05**	**12.29±1.34**	**7.14±0.69**
	GPS L	0.99±0.00	0.98±0.01	0.12±0.15	23.64±0.85	12.46±2.18
0.400000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.69±0.04	**0.46±0.04**	**0.26±0.05**	**12.07±0.87**	**7.05±0.51**
	GPS L	1.00±0.00	0.99±0.01	0.12±0.15	23.29±1.13	12.66±2.53
0.450000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.63±0.04	**0.40±0.04**	**0.26±0.04**	**11.79±0.77**	**6.95±0.44**
	GPS L	1.00±0.00	0.99±0.00	0.12±0.15	23.78±1.13	12.79±3.08
0.500000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.59±0.04	**0.34±0.04**	**0.27±0.04**	**11.55±0.77**	**6.81±0.46**
	GPS L	1.00±0.00	0.99±0.01	0.12±0.15	23.91±0.68	12.96±2.09

Figure 12: Results for Gemma 2 27B on DS1000

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.96±0.02	**0.84±0.06**	**0.23±0.16**	24.76±0.15	**4.88±0.18**
	GPS L	0.98±0.01	0.96±0.01	0.17±0.20	**24.57±0.33**	5.03±0.48
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.93±0.05	**0.77±0.13**	**0.26±0.15**	24.60±0.74	**4.70±0.48**
	GPS L	0.99±0.00	0.98±0.01	0.16±0.21	**24.42±0.70**	5.68±1.44
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.86±0.04	**0.56±0.06**	**0.35±0.08**	**22.50±0.89**	4.63±0.28
	GPS L	1.00±0.00	0.99±0.00	0.16±0.21	23.80±1.02	**4.28±2.67**
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.82±0.04	**0.50±0.04**	**0.36±0.07**	**21.69±0.49**	4.37±0.14
	GPS L	1.00±0.00	1.00±0.00	0.15±0.21	23.00±0.00	**1.00±0.00**
0.300000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.78±0.05	**0.44±0.06**	**0.38±0.06**	**21.06±0.85**	4.22±0.16
	GPS L	1.00±0.00	1.00±0.00	0.15±0.22	23.00±0.00	**1.00±0.00**
0.350000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.74±0.04	**0.37±0.05**	**0.40±0.04**	**19.98±0.81**	**4.01±0.17**
	GPS L	1.00±0.00	0.99±0.00	0.16±0.21	24.50±0.59	4.41±2.67
0.400000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.69±0.05	**0.30±0.05**	**0.41±0.04**	**18.79±1.09**	3.75±0.22
	GPS L	1.00±0.00	1.00±0.00	0.15±0.21	24.00±0.00	**1.00±0.00**
0.450000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.64±0.05	**0.24±0.04**	**0.42±0.03**	**17.77±0.98**	3.56±0.15
	GPS L	1.00±0.00	1.00±0.00	0.15±0.21	24.00±0.00	**1.00±0.00**
0.500000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.59±0.05	**0.17±0.04**	**0.43±0.03**	**15.90±1.71**	3.37±0.18
	GPS L	1.00±0.00	1.00±0.00	0.15±0.22	24.00±0.00	**1.00±0.00**

Figure 13: Results for GPT 4o mini on DS1000

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.99±0.00	**0.76±0.02**	**0.43±0.28**	**25.00±0.00**	**22.51±0.09**
	GPS L	1.00±0.00	1.00±0.00	0.28±0.39	-	-
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.95±0.05	**0.43±0.20**	**0.62±0.19**	**24.81±0.32**	**22.50±0.24**
	GPS L	1.00±0.00	1.00±0.00	0.28±0.39	-	-
0.200000	CLM FirstK	0.98±0.06	0.88±0.33	0.10±0.26	10.78±7.60	6.20±3.86
	CLM FirstK (nr)	0.96±0.08	0.82±0.38	0.14±0.31	9.06±6.05	9.06±6.05
	CLM Max	0.98±0.06	0.91±0.29	0.07±0.23	15.09±6.37	5.96±3.98
	CLM Sum	0.98±0.06	0.91±0.29	0.07±0.23	**6.80±3.89**	**3.88±1.40**
	GPS HS	0.85±0.03	**0.04±0.03**	**0.81±0.02**	19.93±5.78	18.24±5.15
	GPS L	0.97±0.07	0.84±0.37	0.36±0.41	12.31±6.69	11.47±6.00
0.250000	CLM FirstK	0.85±0.10	0.27±0.44	0.58±0.35	3.02±3.45	2.16±1.37
	CLM FirstK (nr)	0.83±0.08	0.16±0.37	0.67±0.29	3.25±3.65	3.25±3.65
	CLM Max	0.85±0.10	0.32±0.47	0.53±0.37	5.62±6.08	2.05±1.31
	CLM Sum	0.85±0.10	0.29±0.45	0.56±0.36	**2.91±3.39**	**1.97±1.11**
	GPS HS	0.80±0.03	**0.00±0.00**	**0.80±0.03**	3.66±3.44	3.53±3.16
	GPS L	0.81±0.04	0.04±0.20	0.78±0.13	3.46±2.82	3.36±2.61
0.300000	CLM FirstK	0.76±0.04	0.01±0.10	0.75±0.08	1.18±0.91	1.13±0.83
	CLM FirstK (nr)	0.78±0.02	**0.00±0.00**	0.78±0.02	1.18±0.93	1.18±0.93
	CLM Max	0.76±0.04	0.01±0.10	0.75±0.08	1.71±2.84	1.23±1.24
	CLM Sum	0.76±0.04	**0.00±0.00**	0.76±0.04	1.17±0.59	**1.07±0.46**
	GPS HS	0.79±0.02	**0.00±0.00**	**0.79±0.02**	1.48±0.27	1.46±0.27
	GPS L	0.79±0.02	**0.00±0.00**	0.79±0.02	**1.12±0.35**	1.12±0.35
0.350000	CLM FirstK	0.72±0.05	**0.00±0.00**	0.72±0.05	**1.00±0.00**	**0.91±0.04**
	CLM FirstK (nr)	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
	CLM Max	0.72±0.05	**0.00±0.00**	0.72±0.05	1.03±0.28	0.93±0.18
	CLM Sum	0.72±0.05	**0.00±0.00**	0.72±0.05	**1.00±0.00**	**0.91±0.04**
	GPS HS	0.78±0.02	**0.00±0.00**	**0.78±0.02**	1.33±0.04	1.32±0.04
	GPS L	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
0.400000	CLM FirstK	0.67±0.05	**0.00±0.00**	0.67±0.05	**1.00±0.00**	**0.85±0.05**
	CLM FirstK (nr)	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
	CLM Max	0.67±0.05	**0.00±0.00**	0.67±0.05	**1.00±0.00**	**0.85±0.05**
	CLM Sum	0.67±0.05	**0.00±0.00**	0.67±0.05	**1.00±0.00**	**0.85±0.05**
	GPS HS	0.79±0.01	**0.00±0.00**	**0.79±0.01**	1.25±0.04	1.24±0.04
	GPS L	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
0.450000	CLM FirstK	0.62±0.05	**0.00±0.00**	0.62±0.05	**1.00±0.00**	**0.78±0.05**
	CLM FirstK (nr)	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
	CLM Max	0.62±0.05	**0.00±0.00**	0.62±0.05	**1.00±0.00**	**0.78±0.05**
	CLM Sum	0.62±0.05	**0.00±0.00**	0.62±0.05	**1.00±0.00**	**0.78±0.05**
	GPS HS	0.79±0.02	**0.00±0.00**	**0.79±0.02**	1.19±0.03	1.18±0.03
	GPS L	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
0.500000	CLM FirstK	0.57±0.05	**0.00±0.00**	0.57±0.05	**1.00±0.00**	**0.71±0.05**
	CLM FirstK (nr)	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00
	CLM Max	0.57±0.05	**0.00±0.00**	0.57±0.05	**1.00±0.00**	**0.71±0.05**
	CLM Sum	0.57±0.05	**0.00±0.00**	0.57±0.05	**1.00±0.00**	**0.71±0.05**
	GPS HS	0.78±0.02	**0.00±0.00**	**0.78±0.02**	1.16±0.03	1.16±0.03
	GPS L	0.78±0.02	**0.00±0.00**	0.78±0.02	**1.00±0.00**	1.00±0.00

Figure 14: Results for GPT 4o mini on GSM8K

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.95±0.02	**0.61±0.08**	**0.50±0.23**	**24.02±0.21**	21.61±0.12
	GPS L	0.99±0.00	0.92±0.01	0.32±0.35	24.42±0.16	**21.27±0.25**
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.91±0.03	**0.39±0.10**	**0.62±0.15**	**23.86±0.19**	21.59±0.19
	GPS L	0.99±0.00	0.96±0.01	0.30±0.37	24.30±0.22	**21.55±0.29**
0.200000	CLM FirstK	1.00±0.02	0.99±0.10	0.01±0.08	24.79±0.00	15.76±0.00
	CLM FirstK (nr)	0.99±0.05	0.95±0.22	0.04±0.17	23.80±1.20	23.80±1.20
	CLM Max	1.00±0.03	0.98±0.14	0.02±0.11	24.89±0.04	14.28±1.75
	CLM Sum	1.00±0.03	0.98±0.14	0.02±0.11	24.56±0.34	**14.16±1.85**
	GPS HS	0.87±0.02	**0.18±0.02**	**0.73±0.06**	**23.59±0.13**	21.15±0.10
	GPS L	0.99±0.03	0.96±0.14	0.30±0.38	24.29±0.26	22.16±0.87
0.250000	CLM FirstK	0.90±0.11	0.55±0.50	0.35±0.39	22.24±2.99	13.90±2.51
	CLM FirstK (nr)	0.87±0.10	0.37±0.48	0.50±0.38	**19.89±3.72**	19.89±3.72
	CLM Max	0.88±0.11	0.44±0.50	0.44±0.39	24.76±0.37	**13.65±2.77**
	CLM Sum	0.89±0.11	0.50±0.50	0.39±0.39	24.16±0.76	13.92±2.95
	GPS HS	0.81±0.03	**0.02±0.04**	**0.79±0.02**	20.47±2.80	18.58±2.41
	GPS L	0.82±0.07	0.11±0.29	0.74±0.19	21.36±3.29	19.32±2.81
0.300000	CLM FirstK	0.77±0.06	0.05±0.22	0.72±0.17	16.21±5.03	11.37±2.81
	CLM FirstK (nr)	0.77±0.04	**0.00±0.00**	**0.77±0.04**	13.72±4.34	13.72±4.34
	CLM Max	0.78±0.07	0.06±0.24	0.72±0.18	23.45±1.59	**10.97±2.00**
	CLM Sum	0.77±0.06	0.05±0.22	0.72±0.17	23.03±1.56	11.16±2.42
	GPS HS	0.77±0.03	**0.00±0.00**	0.77±0.03	**12.67±2.27**	11.81±2.01
	GPS L	0.77±0.03	**0.00±0.00**	0.77±0.03	12.95±2.31	12.05±2.04
0.350000	CLM FirstK	0.72±0.04	**0.00±0.00**	0.72±0.04	9.94±3.71	7.68±1.71
	CLM FirstK (nr)	0.72±0.04	**0.00±0.00**	0.72±0.04	8.06±1.97	8.06±1.97
	CLM Max	0.72±0.04	**0.00±0.00**	0.72±0.04	21.03±2.14	9.13±1.62
	CLM Sum	0.72±0.04	**0.00±0.00**	0.72±0.04	20.51±2.37	9.09±1.58
	GPS HS	0.72±0.03	**0.00±0.00**	**0.72±0.03**	**7.91±0.89**	**7.54±0.82**
	GPS L	0.72±0.03	**0.00±0.00**	0.72±0.03	7.95±1.05	7.57±0.97
0.400000	CLM FirstK	0.67±0.05	**0.00±0.00**	0.67±0.05	6.69±2.35	**5.47±1.35**
	CLM FirstK (nr)	0.67±0.04	**0.00±0.00**	0.67±0.04	**5.78±1.12**	5.78±1.12
	CLM Max	0.67±0.04	**0.00±0.00**	0.67±0.04	18.36±2.58	7.40±1.35
	CLM Sum	0.67±0.04	**0.00±0.00**	0.67±0.04	17.11±3.45	8.05±1.66
	GPS HS	0.67±0.03	**0.00±0.00**	0.67±0.03	5.83±0.70	5.60±0.65
	GPS L	0.67±0.03	**0.00±0.00**	**0.67±0.03**	5.78±0.62	5.56±0.58
0.450000	CLM FirstK	0.62±0.04	**0.00±0.00**	0.62±0.04	4.48±1.07	**4.00±0.87**
	CLM FirstK (nr)	0.63±0.05	**0.00±0.00**	0.63±0.05	**4.13±0.89**	4.13±0.89
	CLM Max	0.62±0.04	**0.00±0.00**	0.62±0.04	13.36±3.73	6.21±1.14
	CLM Sum	0.62±0.04	**0.00±0.00**	0.62±0.04	11.28±4.51	6.62±1.58
	GPS HS	0.63±0.04	**0.00±0.00**	**0.63±0.04**	4.40±0.64	4.25±0.60
	GPS L	0.62±0.05	**0.00±0.00**	0.62±0.05	4.17±0.77	4.03±0.73
0.500000	CLM FirstK	0.59±0.04	**0.00±0.00**	0.59±0.04	3.15±0.66	3.03±0.54
	CLM FirstK (nr)	0.59±0.04	**0.00±0.00**	0.59±0.04	3.04±0.55	3.04±0.55
	CLM Max	0.57±0.04	**0.00±0.00**	0.57±0.04	8.78±2.91	4.58±0.90
	CLM Sum	0.57±0.04	**0.00±0.00**	0.57±0.04	6.06±2.17	4.87±1.42
	GPS HS	0.59±0.04	**0.00±0.00**	0.59±0.04	3.27±0.47	3.18±0.45
	GPS L	0.59±0.03	**0.00±0.00**	**0.59±0.03**	**3.03±0.33**	**2.97±0.31**

Figure 15: Results for Llama 3 8B on GSM8K

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.95±0.02	**0.64±0.06**	**0.45±0.21**	**23.81±0.15**	**21.27±0.15**
	GPS L	1.00±0.00	1.00±0.00	0.25±0.35	-	-
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.92±0.02	**0.48±0.06**	**0.54±0.15**	**23.72±0.22**	**21.12±0.25**
	GPS L	1.00±0.00	1.00±0.00	0.25±0.35	-	-
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.86±0.03	**0.30±0.06**	**0.62±0.09**	**22.99±0.43**	**20.37±0.39**
	GPS L	1.00±0.00	1.00±0.00	0.25±0.35	-	-
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.03	0.99±0.10	0.01±0.07	25.00±0.00	25.00±0.00
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.81±0.03	**0.17±0.04**	**0.67±0.05**	**21.81±0.85**	**19.32±0.70**
	GPS L	1.00±0.03	0.99±0.10	0.25±0.35	23.00±0.00	20.30±0.00
0.300000	CLM FirstK	0.95±0.11	0.81±0.39	0.14±0.28	23.87±1.38	17.03±3.59
	CLM FirstK (nr)	0.90±0.13	0.66±0.47	0.24±0.34	21.50±3.24	21.50±3.24
	CLM Max	0.95±0.11	0.81±0.39	0.14±0.28	22.91±0.93	16.60±2.54
	CLM Sum	0.94±0.12	0.79±0.41	0.15±0.29	21.26±2.78	16.50±3.14
	GPS HS	0.76±0.03	**0.07±0.03**	**0.70±0.02**	**18.19±2.37**	**16.30±2.00**
	GPS L	0.93±0.12	0.75±0.43	0.37±0.37	22.60±2.22	19.95±1.80
0.350000	CLM FirstK	0.76±0.12	0.20±0.40	0.56±0.28	17.63±6.33	12.22±3.54
	CLM FirstK (nr)	0.73±0.09	0.09±0.29	0.64±0.20	14.27±5.46	14.27±5.46
	CLM Max	0.76±0.13	0.22±0.41	0.54±0.29	20.80±2.81	11.36±3.03
	CLM Sum	0.75±0.12	0.17±0.38	0.58±0.27	16.53±3.85	12.17±3.61
	GPS HS	0.72±0.03	**0.02±0.01**	**0.70±0.02**	**11.79±2.22**	**10.83±1.92**
	GPS L	0.72±0.06	0.03±0.17	0.70±0.10	14.63±4.42	13.27±3.75
0.400000	CLM FirstK	0.66±0.04	**0.00±0.00**	0.66±0.04	9.04±3.69	7.41±2.49
	CLM FirstK (nr)	0.66±0.04	**0.00±0.00**	0.66±0.04	7.74±2.56	7.74±2.56
	CLM Max	0.66±0.05	0.01±0.10	0.65±0.08	16.30±3.80	8.09±2.10
	CLM Sum	0.66±0.04	**0.00±0.00**	0.66±0.04	11.52±3.72	8.14±3.13
	GPS HS	0.67±0.04	0.01±0.00	0.66±0.04	7.80±1.28	7.30±1.15
	GPS L	0.67±0.03	**0.00±0.00**	**0.67±0.03**	**7.64±1.32**	**7.19±1.19**
0.450000	CLM FirstK	0.62±0.04	**0.00±0.00**	0.62±0.04	5.52±1.85	**4.47±1.34**
	CLM FirstK (nr)	0.62±0.04	**0.00±0.00**	0.62±0.04	4.69±1.32	4.69±1.32
	CLM Max	0.61±0.04	**0.00±0.00**	0.61±0.04	11.73±2.76	5.52±1.56
	CLM Sum	0.62±0.04	**0.00±0.00**	0.62±0.04	7.18±2.40	4.81±1.36
	GPS HS	0.63±0.03	0.00±0.00	0.62±0.03	5.08±0.65	4.85±0.60
	GPS L	0.62±0.03	**0.00±0.00**	**0.62±0.03**	**4.67±0.79**	4.50±0.73
0.500000	CLM FirstK	0.57±0.04	**0.00±0.00**	0.57±0.04	3.43±0.87	**3.06±0.62**
	CLM FirstK (nr)	0.58±0.04	**0.00±0.00**	0.58±0.04	**3.21±0.57**	3.21±0.57
	CLM Max	0.57±0.04	**0.00±0.00**	0.57±0.04	8.29±1.82	3.83±1.07
	CLM Sum	0.57±0.04	**0.00±0.00**	0.57±0.04	4.48±1.20	3.22±0.64
	GPS HS	0.60±0.02	**0.00±0.00**	**0.60±0.02**	4.02±0.22	3.86±0.20
	GPS L	0.58±0.03	**0.00±0.00**	0.58±0.03	**3.21±0.41**	3.13±0.38

Figure 16: Results for Phi 2 on GSM8K

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.150000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.200000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.250000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.300000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.350000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.400000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.450000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
0.500000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.10±0.14**	-	-

Figure 17: Results for Gemma 27B-IT on MATH

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.93±0.01	**0.29±0.02**	**0.69±0.08**	**21.28±0.39**	**21.16±0.39**
	GPS L	1.00±0.00	1.00±0.00	0.27±0.38	-	-
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.88±0.01	**0.16±0.02**	**0.75±0.04**	**18.59±0.43**	**18.49±0.43**
	GPS L	1.00±0.00	1.00±0.00	0.27±0.38	-	-
0.200000	CLM FirstK	0.98±0.05	0.92±0.27	0.06±0.22	24.59±0.55	20.59±1.68
	CLM FirstK (nr)	0.98±0.06	0.90±0.30	0.08±0.24	23.70±1.57	23.70±1.57
	CLM Max	0.98±0.05	0.92±0.27	0.06±0.22	24.91±0.15	19.78±2.25
	CLM Sum	0.98±0.05	0.92±0.27	0.06±0.22	23.33±1.78	20.05±1.81
	GPS HS	0.83±0.02	**0.07±0.01**	**0.77±0.02**	**14.99±0.81**	**14.92±0.80**
	GPS L	0.99±0.04	0.97±0.17	0.29±0.39	24.67±0.49	24.54±0.49
0.250000	CLM FirstK	0.78±0.03	0.01±0.10	0.77±0.08	16.61±2.94	13.73±1.66
	CLM FirstK (nr)	0.78±0.03	0.01±0.10	0.77±0.08	14.81±2.22	14.81±2.22
	CLM Max	0.78±0.03	0.01±0.10	0.77±0.08	23.53±1.49	14.11±1.61
	CLM Sum	0.78±0.03	0.01±0.10	0.77±0.08	16.53±2.45	14.55±1.45
	GPS HS	0.78±0.02	0.03±0.01	0.76±0.01	**10.94±0.77**	**10.90±0.76**
	GPS L	0.78±0.02	**0.00±0.00**	**0.78±0.02**	14.89±1.46	14.83±1.45
0.300000	CLM FirstK	0.73±0.02	**0.00±0.00**	0.73±0.02	9.22±1.51	8.25±1.16
	CLM FirstK (nr)	0.73±0.02	**0.00±0.00**	0.73±0.02	8.52±1.06	8.52±1.06
	CLM Max	0.73±0.02	**0.00±0.00**	0.73±0.02	18.62±1.63	10.98±1.30
	CLM Sum	0.73±0.02	**0.00±0.00**	0.73±0.02	12.03±0.86	9.98±1.15
	GPS HS	0.74±0.02	0.01±0.00	0.73±0.02	**8.07±0.56**	**8.05±0.56**
	GPS L	0.74±0.02	**0.00±0.00**	**0.74±0.02**	8.78±0.92	8.75±0.92
0.350000	CLM FirstK	0.69±0.02	**0.00±0.00**	0.69±0.02	5.69±0.73	**5.38±0.62**
	CLM FirstK (nr)	0.69±0.02	**0.00±0.00**	0.69±0.02	**5.50±0.59**	5.50±0.59
	CLM Max	0.68±0.02	**0.00±0.00**	0.68±0.02	14.42±1.79	9.09±1.44
	CLM Sum	0.68±0.02	**0.00±0.00**	0.68±0.02	9.58±0.82	6.94±1.00
	GPS HS	0.69±0.02	0.00±0.00	0.69±0.02	5.94±0.45	5.93±0.45
	GPS L	0.69±0.02	**0.00±0.00**	**0.69±0.02**	5.69±0.50	5.68±0.50
0.400000	CLM FirstK	0.64±0.02	**0.00±0.00**	0.64±0.02	**3.86±0.34**	**3.78±0.42**
	CLM FirstK (nr)	0.65±0.02	**0.00±0.00**	0.65±0.02	3.86±0.35	3.86±0.35
	CLM Max	0.64±0.02	**0.00±0.00**	0.64±0.02	10.93±1.48	7.06±1.13
	CLM Sum	0.63±0.02	**0.00±0.00**	0.63±0.02	7.87±0.66	4.85±0.82
	GPS HS	0.65±0.02	0.00±0.00	0.65±0.02	4.39±0.33	4.38±0.33
	GPS L	0.65±0.02	**0.00±0.00**	**0.65±0.02**	4.01±0.30	4.00±0.30
0.450000	CLM FirstK	0.61±0.02	**0.00±0.00**	0.61±0.02	3.00±0.24	**2.90±0.31**
	CLM FirstK (nr)	0.61±0.01	**0.00±0.00**	0.61±0.01	**2.98±0.14**	2.98±0.14
	CLM Max	0.59±0.02	**0.00±0.00**	0.59±0.02	8.11±1.07	5.20±0.79
	CLM Sum	0.59±0.02	**0.00±0.00**	0.59±0.02	6.21±0.59	3.58±0.58
	GPS HS	0.63±0.01	**0.00±0.00**	**0.63±0.01**	3.93±0.05	3.92±0.05
	GPS L	0.62±0.01	**0.00±0.00**	0.62±0.01	3.00±0.00	2.99±0.00
0.500000	CLM FirstK	0.55±0.01	**0.00±0.00**	0.55±0.01	2.02±0.19	**2.00±0.01**
	CLM FirstK (nr)	0.55±0.02	**0.00±0.00**	0.55±0.02	2.02±0.14	2.02±0.14
	CLM Max	0.53±0.02	**0.00±0.00**	0.53±0.02	5.68±0.83	3.54±0.66
	CLM Sum	0.53±0.02	**0.00±0.00**	0.53±0.02	4.78±0.60	2.47±0.45
	GPS HS	0.56±0.04	**0.00±0.00**	**0.56±0.04**	2.91±0.45	2.91±0.45
	GPS L	0.55±0.01	**0.00±0.00**	0.55±0.01	**2.01±0.10**	2.01±0.10

Figure 18: Results for GPT 4o mini on MATH

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	1.00±0.00	1.00±0.00	0.28±0.40	-	-
	GPS L	1.00±0.00	**0.99±0.01**	**0.28±0.40**	**24.00±0.00**	**1.82±0.24**
0.150000	CLM FirstK	1.00±0.02	**0.99±0.10**	0.01±0.08	**1.00±0.00**	**1.00±0.00**
	CLM FirstK (nr)	1.00±0.02	**0.99±0.10**	0.01±0.08	**1.00±0.00**	**1.00±0.00**
	CLM Max	1.00±0.02	**0.99±0.10**	0.01±0.08	**1.00±0.00**	**1.00±0.00**
	CLM Sum	1.00±0.02	**0.99±0.10**	0.01±0.08	**1.00±0.00**	**1.00±0.00**
	GPS HS	1.00±0.00	1.00±0.00	0.28±0.40	-	-
	GPS L	1.00±0.00	0.99±0.01	**0.28±0.39**	24.00±0.00	1.83±0.24
0.200000	CLM FirstK	0.99±0.04	0.96±0.20	0.03±0.15	4.00±3.43	3.99±3.44
	CLM FirstK (nr)	0.99±0.05	0.95±0.22	0.04±0.17	3.40±3.28	3.40±3.28
	CLM Max	0.99±0.05	**0.94±0.24**	0.05±0.19	4.11±3.42	**1.16±0.15**
	CLM Sum	0.99±0.05	**0.94±0.24**	0.05±0.19	**3.29±2.61**	3.23±2.50
	GPS HS	1.00±0.02	0.99±0.10	0.29±0.40	15.00±0.00	2.00±0.00
	GPS L	0.99±0.02	0.98±0.10	**0.29±0.40**	23.88±1.03	1.83±0.24
0.250000	CLM FirstK	0.90±0.11	0.53±0.50	0.37±0.39	3.22±4.61	1.89±1.98
	CLM FirstK (nr)	0.90±0.10	0.48±0.50	0.42±0.40	2.31±3.38	2.31±3.38
	CLM Max	0.92±0.11	0.62±0.49	0.30±0.38	**1.76±2.31**	**1.01±0.13**
	CLM Sum	0.90±0.11	0.54±0.50	0.36±0.39	2.54±3.55	1.64±1.45
	GPS HS	0.86±0.09	0.27±0.44	0.67±0.32	5.04±4.85	1.39±0.36
	GPS L	0.86±0.08	**0.27±0.44**	**0.67±0.31**	10.07±9.03	1.59±0.36
0.300000	CLM FirstK	0.84±0.11	0.26±0.44	0.58±0.34	1.56±2.07	1.28±1.07
	CLM FirstK (nr)	0.84±0.08	0.19±0.39	0.65±0.32	**1.32±1.71**	1.32±1.71
	CLM Max	0.85±0.11	0.35±0.48	0.50±0.37	1.38±1.79	**0.97±0.10**
	CLM Sum	0.84±0.11	0.26±0.44	0.58±0.35	1.49±2.02	1.21±0.82
	GPS HS	0.83±0.05	0.05±0.22	**0.79±0.15**	2.38±3.44	1.13±0.28
	GPS L	0.82±0.05	**0.05±0.21**	0.79±0.14	4.23±5.63	1.30±0.24
0.350000	CLM FirstK	0.75±0.09	0.04±0.20	0.71±0.16	1.11±0.60	0.97±0.54
	CLM FirstK (nr)	0.81±0.03	**0.00±0.00**	**0.81±0.03**	1.04±0.40	1.04±0.40
	CLM Max	0.76±0.09	0.05±0.22	0.71±0.18	1.06±0.43	**0.89±0.11**
	CLM Sum	0.75±0.09	0.04±0.20	0.71±0.16	1.09±0.49	0.95±0.40
	GPS HS	0.81±0.03	**0.00±0.00**	0.81±0.03	1.00±0.00	1.00±0.00
	GPS L	0.81±0.03	**0.00±0.00**	0.81±0.03	**0.99±0.01**	0.99±0.01
0.400000	CLM FirstK	0.72±0.08	0.01±0.10	0.71±0.11	1.06±0.44	0.88±0.34
	CLM FirstK (nr)	0.81±0.03	**0.00±0.00**	0.81±0.03	1.00±0.00	1.00±0.00
	CLM Max	0.72±0.08	0.01±0.10	0.71±0.11	1.06±0.42	**0.85±0.12**
	CLM Sum	0.72±0.08	0.01±0.10	0.71±0.11	1.05±0.39	0.87±0.27
	GPS HS	0.82±0.03	**0.00±0.00**	**0.82±0.03**	1.00±0.00	1.00±0.00
	GPS L	0.81±0.03	**0.00±0.00**	0.81±0.03	**0.99±0.01**	0.99±0.01
0.450000	CLM FirstK	0.68±0.08	**0.00±0.00**	0.68±0.08	1.00±0.00	**0.79±0.10**
	CLM FirstK (nr)	0.81±0.03	**0.00±0.00**	**0.81±0.03**	1.00±0.00	1.00±0.00
	CLM Max	0.68±0.08	**0.00±0.00**	0.68±0.08	1.00±0.00	**0.79±0.10**
	CLM Sum	0.68±0.08	**0.00±0.00**	0.68±0.08	1.00±0.00	**0.79±0.10**
	GPS HS	0.81±0.03	**0.00±0.00**	0.81±0.03	1.00±0.00	1.00±0.00
	GPS L	0.81±0.03	**0.00±0.00**	0.81±0.03	**0.99±0.01**	0.99±0.01
0.500000	CLM FirstK	0.63±0.08	**0.00±0.00**	0.63±0.08	1.00±0.00	**0.72±0.09**
	CLM FirstK (nr)	0.81±0.03	**0.00±0.00**	0.81±0.03	1.00±0.00	1.00±0.00
	CLM Max	0.63±0.08	**0.00±0.00**	0.63±0.08	1.00±0.00	**0.72±0.09**
	CLM Sum	0.63±0.08	**0.00±0.00**	0.63±0.08	1.00±0.00	**0.72±0.09**
	GPS HS	0.81±0.03	**0.00±0.00**	**0.81±0.03**	1.00±0.00	1.00±0.00
	GPS L	0.81±0.03	**0.00±0.00**	0.81±0.03	**0.99±0.01**	0.99±0.01

Figure 19: Results for GPT 4o mini on MBPP

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.99±0.03	**0.92±0.12**	**0.30±0.34**	**25.00±0.02**	**15.90±0.72**
	GPS L	1.00±0.00	1.00±0.00	0.26±0.37	25.00±0.00	18.00±0.00
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.95±0.03	**0.57±0.09**	**0.52±0.19**	**24.96±0.13**	**13.62±0.18**
	GPS L	1.00±0.00	0.98±0.01	0.27±0.36	25.00±0.00	15.26±1.83
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.90±0.06	**0.46±0.30**	**0.55±0.26**	**24.75±0.44**	**13.74±1.13**
	GPS L	1.00±0.00	1.00±0.00	0.26±0.37	25.00±0.00	18.00±0.00
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.84±0.05	**0.17±0.12**	**0.71±0.08**	**22.75±2.68**	**12.45±1.09**
	GPS L	0.99±0.03	0.98±0.10	0.27±0.36	24.95±0.43	15.31±2.34
0.300000	CLM FirstK	0.98±0.07	0.93±0.26	0.05±0.18	12.16±5.79	7.13±2.71
	CLM FirstK (nr)	0.96±0.10	0.84±0.37	0.12±0.27	**12.12±5.68**	12.12±5.68
	CLM Max	0.99±0.06	0.96±0.20	0.03±0.14	20.54±3.90	**4.72±1.26**
	CLM Sum	0.98±0.07	0.94±0.24	0.04±0.17	13.40±5.62	8.44±4.80
	GPS HS	0.79±0.05	**0.06±0.06**	**0.75±0.03**	16.87±5.45	9.91±2.52
	GPS L	0.91±0.12	0.68±0.44	0.41±0.36	21.56±6.13	13.65±3.81
0.350000	CLM FirstK	0.87±0.15	0.55±0.50	0.32±0.35	9.75±6.27	5.77±3.07
	CLM FirstK (nr)	0.81±0.14	0.32±0.47	0.49±0.34	7.88±5.26	7.88±5.26
	CLM Max	0.91±0.15	0.71±0.45	0.20±0.31	17.08±4.55	**4.79±3.49**
	CLM Sum	0.89±0.15	0.64±0.48	0.25±0.34	8.81±5.25	6.30±3.75
	GPS HS	0.74±0.05	**0.01±0.01**	**0.73±0.05**	8.03±3.73	5.53±1.98
	GPS L	0.75±0.09	0.08±0.27	0.69±0.17	**7.83±3.94**	5.46±2.05
0.400000	CLM FirstK	0.76±0.14	0.22±0.41	0.54±0.29	6.22±4.58	4.15±3.00
	CLM FirstK (nr)	0.74±0.11	0.13±0.34	0.61±0.24	5.14±3.66	5.14±3.66
	CLM Max	0.81±0.17	0.41±0.49	0.40±0.34	14.04±6.23	4.39±4.17
	CLM Sum	0.78±0.16	0.32±0.47	0.46±0.32	6.98±4.03	4.18±1.98
	GPS HS	0.71±0.05	0.00±0.00	**0.71±0.05**	5.12±2.03	3.86±1.25
	GPS L	0.71±0.06	**0.00±0.00**	0.71±0.06	**4.99±2.28**	**3.82±1.38**
0.450000	CLM FirstK	0.70±0.12	0.12±0.33	0.58±0.22	4.02±3.62	2.91±2.22
	CLM FirstK (nr)	0.68±0.09	0.04±0.20	0.64±0.15	3.51±3.28	3.51±3.28
	CLM Max	0.72±0.16	0.22±0.41	0.50±0.27	9.65±6.43	2.76±1.86
	CLM Sum	0.72±0.14	0.18±0.38	0.54±0.26	5.00±3.32	3.00±1.47
	GPS HS	0.68±0.04	**0.00±0.00**	**0.68±0.04**	3.04±0.79	2.54±0.54
	GPS L	0.67±0.05	**0.00±0.00**	0.67±0.05	**2.77±0.99**	**2.39±0.67**
0.500000	CLM FirstK	0.64±0.10	0.03±0.17	0.61±0.13	2.43±2.19	**1.86±1.15**
	CLM FirstK (nr)	0.63±0.08	0.01±0.10	0.62±0.09	2.04±2.25	2.04±2.25
	CLM Max	0.65±0.13	0.09±0.29	0.56±0.19	5.84±6.68	1.91±2.14
	CLM Sum	0.63±0.10	0.03±0.17	0.60±0.13	2.82±2.41	1.95±1.43
	GPS HS	0.66±0.03	**0.00±0.00**	**0.66±0.03**	2.41±0.11	2.10±0.09
	GPS L	0.65±0.03	**0.00±0.00**	0.65±0.03	**2.01±0.14**	1.87±0.10

Figure 20: Results for Llama 3 8B on MBPP

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.99±0.04	**0.93±0.17**	**0.28±0.35**	25.00±0.00	**12.99±0.13**
	GPS L	0.99±0.01	0.98±0.01	0.26±0.35	**24.14±0.68**	13.02±0.84
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.97±0.05	**0.78±0.23**	**0.38±0.30**	24.97±0.06	14.25±0.83
	GPS L	0.99±0.01	0.99±0.01	0.25±0.35	**24.29±0.70**	**13.12±0.84**
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.91±0.05	**0.52±0.17**	**0.51±0.20**	24.84±0.28	**12.62±0.20**
	GPS L	0.99±0.01	0.98±0.01	0.26±0.35	**24.53±0.54**	13.13±0.83
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.86±0.06	**0.26±0.11**	**0.65±0.09**	**24.34±0.76**	**12.57±0.57**
	GPS L	0.99±0.01	0.97±0.01	0.27±0.35	24.76±0.27	13.26±0.76
0.300000	CLM FirstK	1.00±0.03	0.99±0.10	0.01±0.06	19.53±0.00	3.61±0.00
	CLM FirstK (nr)	0.98±0.08	0.93±0.26	0.05±0.18	19.57±7.68	19.57±7.68
	CLM Max	1.00±0.04	0.99±0.10	0.01±0.06	**6.71±0.00**	**1.98±0.00**
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.80±0.06	**0.13±0.07**	**0.69±0.04**	22.12±3.47	11.47±1.49
	GPS L	0.99±0.05	0.97±0.14	0.26±0.35	24.18±2.25	13.00±1.26
0.350000	CLM FirstK	0.89±0.15	0.64±0.48	0.25±0.33	9.61±6.18	4.10±2.02
	CLM FirstK (nr)	0.84±0.15	0.48±0.50	0.36±0.35	9.13±6.97	9.13±6.97
	CLM Max	0.90±0.15	0.68±0.47	0.22±0.32	14.34±6.87	**3.00±1.22**
	CLM Sum	0.89±0.15	0.63±0.48	0.26±0.33	**8.49±3.91**	3.84±1.89
	GPS HS	0.74±0.06	**0.03±0.04**	**0.71±0.04**	11.63±6.59	6.81±3.07
	GPS L	0.82±0.15	0.40±0.49	0.52±0.32	15.46±8.70	8.55±4.01
0.400000	CLM FirstK	0.80±0.17	0.40±0.49	0.40±0.33	5.22±4.24	2.72±1.45
	CLM FirstK (nr)	0.75±0.14	0.22±0.41	0.53±0.29	6.50±6.99	6.50±6.99
	CLM Max	0.79±0.16	0.36±0.48	0.43±0.32	9.91±6.89	**2.28±1.07**
	CLM Sum	0.79±0.16	0.36±0.48	0.43±0.32	**5.20±3.77**	2.50±1.50
	GPS HS	0.71±0.06	**0.01±0.01**	**0.70±0.05**	6.46±3.81	4.31±1.95
	GPS L	0.72±0.09	0.06±0.23	0.68±0.15	8.49±7.11	5.21±3.28
0.450000	CLM FirstK	0.71±0.15	0.18±0.38	0.53±0.26	3.45±3.32	2.18±1.41
	CLM FirstK (nr)	0.70±0.12	0.12±0.33	0.58±0.22	2.90±2.95	2.90±2.95
	CLM Max	0.72±0.16	0.22±0.41	0.50±0.27	5.46±5.91	**1.65±0.90**
	CLM Sum	0.71±0.14	0.17±0.38	0.54±0.25	3.56±3.50	1.84±1.12
	GPS HS	0.68±0.05	0.00±0.00	**0.68±0.05**	3.07±1.00	2.38±0.64
	GPS L	0.67±0.05	**0.00±0.00**	0.67±0.05	**2.75±1.00**	2.27±0.64
0.500000	CLM FirstK	0.63±0.09	0.03±0.17	0.60±0.13	1.95±1.76	1.43±0.95
	CLM FirstK (nr)	0.65±0.06	0.01±0.10	0.64±0.08	**1.67±1.29**	1.67±1.29
	CLM Max	0.63±0.11	0.05±0.22	0.58±0.15	2.80±4.40	**1.19±0.62**
	CLM Sum	0.63±0.08	0.01±0.10	0.62±0.09	2.24±2.60	1.39±1.01
	GPS HS	0.65±0.04	**0.00±0.00**	0.65±0.04	2.05±0.26	1.69±0.18
	GPS L	0.65±0.03	**0.00±0.00**	**0.65±0.03**	1.99±0.14	1.77±0.10

Figure 21: Results for Phi 2 on MBPP

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.93±0.00	**0.61±0.01**	**0.43±0.17**	25.00±0.00	1.69±0.01
	GPS L	1.00±0.00	0.99±0.00	0.23±0.31	**24.92±0.08**	**1.60±0.14**
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.89±0.00	**0.51±0.01**	**0.48±0.13**	25.00±0.00	**1.64±0.00**
	GPS L	1.00±0.00	1.00±0.00	0.22±0.31	**24.00±0.00**	3.00±0.00
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.87±0.00	**0.43±0.01**	**0.52±0.11**	25.00±0.00	**1.61±0.00**
	GPS L	1.00±0.00	1.00±0.00	0.22±0.32	**24.62±0.36**	2.44±0.56
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.77±0.02	**0.17±0.02**	**0.62±0.03**	**23.09±0.64**	**1.60±0.01**
	GPS L	1.00±0.00	1.00±0.00	0.22±0.31	24.88±0.16	1.94±0.24
0.300000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.72±0.01	**0.08±0.01**	**0.64±0.01**	**13.44±2.50**	**1.45±0.06**
	GPS L	1.00±0.00	1.00±0.00	0.22±0.31	24.94±0.07	1.66±0.16
0.350000	CLM FirstK	0.69±0.08	0.06±0.24	0.61±0.18	5.39±5.71	4.33±4.18
	CLM FirstK (nr)	0.69±0.08	0.06±0.24	0.61±0.18	4.89±4.52	4.89±4.52
	CLM Max	0.69±0.08	0.06±0.24	0.61±0.18	4.48±2.08	1.69±1.87
	CLM Sum	0.69±0.08	0.07±0.25	0.61±0.19	4.19±3.91	2.94±2.62
	GPS HS	0.68±0.01	**0.06±0.00**	**0.63±0.01**	**2.05±0.12**	**1.05±0.01**
	GPS L	0.84±0.17	0.54±0.48	0.42±0.31	20.59±6.00	1.78±0.28
0.400000	CLM FirstK	0.62±0.01	**0.00±0.00**	0.62±0.01	1.00±0.00	**0.85±0.02**
	CLM FirstK (nr)	0.65±0.01	**0.00±0.00**	**0.65±0.01**	1.00±0.00	1.00±0.00
	CLM Max	0.62±0.01	**0.00±0.00**	0.62±0.01	1.00±0.00	**0.85±0.02**
	CLM Sum	0.62±0.01	**0.00±0.00**	0.62±0.01	1.00±0.00	**0.85±0.02**
	GPS HS	0.68±0.01	0.05±0.00	0.63±0.01	1.92±0.04	1.04±0.01
	GPS L	0.64±0.01	**0.00±0.00**	0.64±0.01	**1.00±0.00**	1.00±0.00
0.450000	CLM FirstK	0.58±0.01	**0.00±0.00**	0.58±0.01	1.00±0.00	**0.76±0.02**
	CLM FirstK (nr)	0.65±0.01	**0.00±0.00**	**0.65±0.01**	1.00±0.00	1.00±0.00
	CLM Max	0.58±0.01	**0.00±0.00**	0.58±0.01	1.00±0.00	**0.76±0.02**
	CLM Sum	0.58±0.01	**0.00±0.00**	0.58±0.01	1.00±0.00	**0.76±0.02**
	GPS HS	0.68±0.01	0.05±0.00	0.63±0.01	1.85±0.04	1.04±0.01
	GPS L	0.63±0.01	**0.00±0.00**	0.63±0.01	**1.00±0.00**	1.00±0.00
0.500000	CLM FirstK	0.53±0.01	**0.00±0.00**	0.53±0.01	1.00±0.00	**0.67±0.02**
	CLM FirstK (nr)	0.65±0.01	**0.00±0.00**	**0.65±0.01**	1.00±0.00	1.00±0.00
	CLM Max	0.53±0.01	**0.00±0.00**	0.53±0.01	1.00±0.00	**0.67±0.02**
	CLM Sum	0.53±0.01	**0.00±0.00**	0.53±0.01	1.00±0.00	**0.67±0.02**
	GPS HS	0.68±0.01	0.05±0.00	0.63±0.01	1.79±0.04	1.04±0.01
	GPS L	0.64±0.01	**0.00±0.00**	0.64±0.01	**1.00±0.00**	1.00±0.00

Figure 22: Results for GPT 4o mini on TriviaQA

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
0.150000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
0.200000	CLM FirstK	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Max	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	CLM Sum	1.00±0.00	**1.00±0.00**	0.00±0.00	-	-
	GPS HS	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
	GPS L	1.00±0.00	**1.00±0.00**	**0.25±0.36**	-	-
0.250000	CLM FirstK	0.78±0.07	0.08±0.28	0.68±0.24	7.50±4.84	7.37±4.71
	CLM FirstK (nr)	0.78±0.07	0.08±0.28	0.68±0.24	7.23±4.43	7.23±4.43
	CLM Max	0.78±0.06	**0.08±0.27**	**0.68±0.24**	9.11±2.97	3.06±4.18
	CLM Sum	0.79±0.07	0.09±0.29	0.67±0.25	**5.75±3.07**	4.90±3.18
	GPS HS	0.96±0.09	0.83±0.38	0.34±0.38	20.76±5.25	**1.68±0.07**
	GPS L	0.96±0.09	0.83±0.38	0.34±0.38	20.76±5.25	**1.68±0.07**
0.300000	CLM FirstK	0.72±0.01	**0.00±0.00**	0.72±0.01	**1.32±0.52**	1.26±0.54
	CLM FirstK (nr)	0.72±0.01	**0.00±0.00**	**0.72±0.01**	1.32±0.52	1.32±0.52
	CLM Max	0.72±0.01	**0.00±0.00**	0.72±0.01	1.59±1.10	**1.00±0.12**
	CLM Sum	0.72±0.01	**0.00±0.00**	0.72±0.01	1.39±0.63	1.14±0.36
	GPS HS	0.72±0.01	**0.00±0.00**	0.72±0.01	2.64±0.56	1.23±0.05
	GPS L	0.72±0.01	**0.00±0.00**	0.72±0.01	2.64±0.56	1.23±0.05
0.350000	CLM FirstK	0.67±0.01	**0.00±0.00**	0.67±0.01	**1.00±0.00**	**0.82±0.03**
	CLM FirstK (nr)	0.72±0.01	**0.00±0.00**	**0.72±0.01**	**1.00±0.00**	1.00±0.00
	CLM Max	0.67±0.01	**0.00±0.00**	0.67±0.01	**1.00±0.00**	**0.82±0.03**
	CLM Sum	0.67±0.01	**0.00±0.00**	0.67±0.01	**1.00±0.00**	**0.82±0.03**
	GPS HS	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
	GPS L	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
0.400000	CLM FirstK	0.63±0.01	**0.00±0.00**	0.63±0.01	**1.00±0.00**	**0.73±0.02**
	CLM FirstK (nr)	0.72±0.01	**0.00±0.00**	**0.72±0.01**	**1.00±0.00**	1.00±0.00
	CLM Max	0.63±0.01	**0.00±0.00**	0.63±0.01	**1.00±0.00**	**0.73±0.02**
	CLM Sum	0.63±0.01	**0.00±0.00**	0.63±0.01	**1.00±0.00**	**0.73±0.02**
	GPS HS	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
	GPS L	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
0.450000	CLM FirstK	0.58±0.01	**0.00±0.00**	0.58±0.01	**1.00±0.00**	**0.66±0.02**
	CLM FirstK (nr)	0.72±0.01	**0.00±0.00**	**0.72±0.01**	**1.00±0.00**	1.00±0.00
	CLM Max	0.58±0.01	**0.00±0.00**	0.58±0.01	**1.00±0.00**	**0.66±0.02**
	CLM Sum	0.58±0.01	**0.00±0.00**	0.58±0.01	**1.00±0.00**	**0.66±0.02**
	GPS HS	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
	GPS L	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
0.500000	CLM FirstK	0.53±0.02	**0.00±0.00**	0.53±0.02	**1.00±0.00**	**0.59±0.02**
	CLM FirstK (nr)	0.72±0.01	**0.00±0.00**	**0.72±0.01**	**1.00±0.00**	1.00±0.00
	CLM Max	0.53±0.02	**0.00±0.00**	0.53±0.02	**1.00±0.00**	**0.59±0.02**
	CLM Sum	0.53±0.02	**0.00±0.00**	0.53±0.02	**1.00±0.00**	**0.59±0.02**
	GPS HS	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00
	GPS L	0.69±0.01	**0.00±0.00**	0.69±0.01	**1.00±0.00**	1.00±0.00

Figure 23: Results for Llama 3 8B on TriviaQA

α	Method	Set Coverage	Abstention Rate	Non-Abs. Coverage	Set Size	Number of Samples
0.100000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.92±0.00	**0.72±0.01**	**0.28±0.12**	**24.41±0.01**	**2.49±0.01**
	GPS L	0.99±0.00	0.97±0.00	0.16±0.20	24.86±0.04	2.82±0.07
0.150000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.87±0.01	**0.61±0.02**	**0.32±0.09**	**23.27±0.38**	**2.44±0.02**
	GPS L	0.88±0.00	0.71±0.01	0.26±0.13	24.95±0.01	2.49±0.01
0.200000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.83±0.01	**0.53±0.01**	**0.35±0.07**	**22.08±0.07**	**2.43±0.01**
	GPS L	0.97±0.00	0.94±0.00	0.18±0.19	24.95±0.02	2.56±0.03
0.250000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.78±0.02	**0.44±0.02**	**0.37±0.05**	**21.12±0.34**	**2.42±0.01**
	GPS L	0.80±0.01	0.56±0.01	0.31±0.10	24.98±0.00	2.50±0.00
0.300000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.73±0.02	**0.36±0.02**	**0.39±0.04**	**20.09±0.30**	**2.40±0.01**
	GPS L	0.99±0.00	0.98±0.00	0.16±0.21	24.94±0.03	3.47±0.16
0.350000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.68±0.02	**0.29±0.02**	**0.40±0.03**	**19.45±0.34**	**2.39±0.01**
	GPS L	0.97±0.00	0.92±0.00	0.18±0.19	24.98±0.01	2.54±0.02
0.400000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.63±0.02	**0.23±0.02**	**0.42±0.02**	**18.70±0.39**	**2.36±0.02**
	GPS L	0.77±0.01	0.50±0.01	0.33±0.08	25.00±0.00	2.51±0.01
0.450000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.57±0.01	**0.16±0.01**	**0.42±0.01**	**17.42±0.46**	**2.33±0.01**
	GPS L	1.00±0.00	1.00±0.00	0.15±0.21	24.93±0.09	2.40±0.34
0.500000	CLM FirstK	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM FirstK (nr)	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Max	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	CLM Sum	1.00±0.00	1.00±0.00	0.00±0.00	-	-
	GPS HS	0.52±0.01	**0.10±0.01**	**0.43±0.01**	**15.39±0.57**	**2.26±0.02**
	GPS L	1.00±0.00	0.99±0.00	0.15±0.21	24.97±0.03	4.05±0.43

Figure 24: Results for Phi 2 on TriviaQA