

PAC REASONING: CONTROLLING THE PERFORMANCE LOSS FOR EFFICIENT REASONING

Hao Zeng^{1*}, Jianguo Huang^{2*}, Bingyi Jing^{3,1}, Hongxin Wei^{1†}, Bo An²

¹Department of Statistics and Data Science, Southern University of Science and Technology, China

²College of Computing and Data Science, Nanyang Technological University, Singapore

³School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Large reasoning models (LRMs) have achieved remarkable progress in complex problem-solving tasks. Despite this success, LRM typically suffer from high computational costs during deployment, highlighting a need for efficient inference. A popular direction of efficiency improvement is to switch the LRM between thinking and nonthinking modes dynamically. However, such approaches often introduce additional reasoning errors and lack statistical guarantees for the performance loss, which are critical for high-stakes applications. In this work, we propose Probably Approximately Correct (PAC) reasoning that controls the performance loss under the user-specified performance loss tolerance. In particular, we construct an upper confidence bound on the performance loss, formulated as a monotone function of the uncertainty score, and subsequently determine a threshold for switching to the nonthinking model. Theoretically, using the threshold to switch between the thinking and nonthinking modes ensures bounded performance loss in a distribution-free manner. Our comprehensive experiments on reasoning benchmarks show that the proposed method can save computational budgets and control the user-specified performance loss.

1 INTRODUCTION

Large reasoning models (LRMs) have demonstrated strong performance in tackling complex problem-solving (DeepSeek-AI et al., 2025; Yang et al., 2025a; NVIDIA et al., 2025). However, this strong performance largely depends on long reasoning chains, which substantially increase the computational cost during inference. This phenomenon, often referred to as overthinking (Yue et al., 2025), is evident in mathematical and logic-intensive tasks. And, in applications requiring real-time interaction or large-scale processing, such as text generation (Zhang et al., 2022) and chatbot (Roller et al., 2020), inference efficiency directly determines usability and user experience. Therefore, it is essential to improve the inference efficiency of LRMs.

To address this, existing works proposed to switch the LRM into a nonthinking mode to avoid overthinking (Cheng et al., 2025; Chung et al., 2025; Fang et al., 2025; Li et al., 2025; Liang et al., 2025; Ma et al., 2025; Paliotta et al., 2025; Pan et al., 2025; Xiao et al., 2025; Yong et al., 2025; Yue et al., 2025). While effective in reducing computational demands, using a nonthinking model often degrades solution quality or introduces additional errors. For instance, in theorem-proving tasks, switching techniques may lead to invalid logical steps, and in mathematical reasoning, it can result in calculation mistakes or overlooked solution paths. Besides, such methods lack a rigorous theoretical guarantee for performance loss. This limitation raises a fundamental issue:

How to improve the efficiency of LRMs, guaranteeing the performance loss?

In this work, we formalize this challenge by introducing the concept of a PAC efficient model, a notion where an LRM provides statistical guarantees that its performance loss stays within a user-specified tolerance, as defined in Definition 1. To meet this requirement, we propose **PAC reason-**

*Equal Contribution.

†Correspond to weihx@sustech.edu.cn.

ing, which constructs a composite model \hat{f} that selectively switches between the thinking model f and its cheaper non-thinking counterpart \tilde{f} . Concretely, PAC reasoning determines a switching threshold on a calibration dataset via a PAC calibration procedure (Algorithm 1), and during testing (Algorithm 2), it accepts the output of \hat{f} if its uncertainty score is below the threshold; otherwise, it resorts to f for reliability. In this way, PAC reasoning complements existing switching approaches by improving efficiency while offering statistical guarantees on performance loss.

Theoretically, we show that PAC reasoning achieves distribution-free control of performance loss with probabilistic guarantees. We formalize this through the composite model \hat{f} , whose key property is that the loss function is monotonic in the uncertainty score. This monotonicity enables the construction of an upper confidence bound on performance loss and the derivation of a valid threshold. Under mild regularity conditions, we prove that PAC reasoning keeps the loss below the user-specified tolerance, thereby satisfying PAC efficiency.

We then present comprehensive experimental results ¹ in Section 4 that rigorously evaluate the PAC reasoning across diverse reasoning benchmarks, including MATH-500 (Lightman et al., 2023), ZebraLogic(Lin et al., 2025), and Arena-Hard (Li et al., 2024). The results demonstrate that our approach effectively controls the performance loss and significantly reduces inference cost. For example, on Arena-Hard with performance tolerance $\epsilon = 0.08$ for the logits uncertainty score, our method controls the average empirical performance loss at 0.06 (below the tolerance), and achieves token savings exceeding 40%. We also find that the logits-based uncertainty score provides more stable performance loss control compared to the verbalized-based score.

Our contributions are as follows:

- We introduce the concept of (ε, α) -**PAC efficient**, a novel notion for quantifying performance loss of efficiency improvement under PAC-style guarantees in LRM, and to the best of our knowledge, the first formal such guarantee in this setting.
- We propose **PAC reasoning**, a method that combines a thinking-mode model with its nonthinking counterpart via an uncertainty-based mechanism to improve efficiency. The method is model-agnostic and provides *distribution-free* performance guarantees.
- We provide comprehensive experiments on mathematical reasoning, logical deduction, and text generation, demonstrating that PAC reasoning achieves efficiency gains while satisfying the statistical validity of the PAC efficient guarantee.

Notations We begin by introducing key notations. The first is the LRM with thinking-mode f , which is computationally expensive but delivers high performance on its answers. Given an input prompt x , f produces an output $y = f(x)$, which we regard as the “expert answer”. The second is the nonthinking LRM \tilde{f} , which is computationally cheaper but potentially less accurate, and $\tilde{y} = \tilde{f}(x)$. And for any input x , we use y^{gold} as its “gold reference”. Let $\mathcal{I}_{cal} = \{1, \dots, n\}$ and $\mathcal{I}_{test} = \{n + 1, \dots, n + N\}$ denote the indices of the calibration and test sets, respectively. Accordingly, we define the calibration dataset and the test dataset as :

$$\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i \in \mathcal{I}_{cal}}, \quad \mathcal{D}_{test} = \{(x_i, y_i)\}_{i \in \mathcal{I}_{test}},$$

where each x_i is an input prompt. It is worth noting that the y_i is not a ground-truth label, but the “expert answer” provided by the LRM f . Finally, let $y_i = (y_{i,1}, \dots, y_{i,l_{y_i}})$ denote an answer consisting of l_{y_i} tokens, with $y_{i,j}$ representing the j -th token of y_i .

2 PROBABLY APPROXIMATELY CORRECT REASONING

2.1 PAC EFFICIENT MODEL

We aim to build a more efficient LRM, denoted by \hat{f} , that provides probably approximately correct guarantees for its performance loss while improving efficiency. Specifically, given an error tolerance ϵ and a confidence level $1 - \alpha$, \hat{f} ensures its performance loss relative to the thinking-mode LRM f does not exceed ϵ with probability at least $1 - \alpha$. We formulate the PAC guaranteed \hat{f} as follows:

¹The reproducibility code is placed at [an anonymous link](#).

Definition 1 ((ϵ, α)-PAC efficient). An LRM \hat{f} is called an (ϵ, α) -probably approximately correct (PAC) efficient model (with respect to loss ℓ),² if for given $\epsilon > 0, \alpha \in (0, 1)$ it satisfies

$$\mathbb{P} \left(R(\hat{f}) \leq \epsilon \right) \geq 1 - \alpha,$$

where $R(\hat{f}) = \mathbb{E}_{x \sim P} [\ell(\hat{f}(x), f(x))]$ is the risk function, $\ell(\cdot, \cdot)$ is a loss function, x denotes an input prompt drawn from the underlying task distribution P .

Remark 2. Here, the loss function can be a 0-1 loss for verifiable tasks or a semantic loss for generative tasks. The positive value $\epsilon > 0$ is called the error tolerance, and $1 - \alpha$ is termed the confidence level. We sometimes term (ϵ, α) -PAC efficient model simply as PAC efficient model.

2.2 PAC REASONING

Constructing such a controllable LRM \hat{f} is straightforward intuitively. Given an LRM with thinking mode f and a fast LRM without thinking \tilde{f} , we create an intermediate model that selectively uses either the LRM with thinking or not based on certain conditions. This condition acts like a “sliding rheostat” that allows us to tune the performance trade-off by adjusting the “position” of the intermediate. We can obtain a model \hat{f} that achieves the desired error tolerance heuristically. However, this approach lacks statistical guarantees on the underlying distribution of performance loss. To build a model with statistical guarantees, a hypothesis test will be used to determine an optimal threshold that balances computational efficiency with output quality while maintaining statistical confidence.

Motivated by this, we present the **PAC reasoning**, which constructs a composite LRM \hat{f} that improves the efficiency of an LRM with thinking f . The composite LRM \hat{f} provides PAC guarantees for the efficiency improvement. The core idea is to use **uncertainty scores** to build an **upper confidence bound** for the performance loss. We could use the upper confidence bound to measure the uncertainty of performance loss for each value of the uncertainty score. Then we **calibrate** an uncertainty threshold to switch between the thinking and nonthinking models. We use the nonthinking LRM \tilde{f} on most inputs and strategically invoke the expensive LRM with thinking f only for inputs whose generation by \tilde{f} has high uncertainty. Next, we provide the details of the PAC reasoning.

2.2.1 UNCERTAINTY SCORES AND CUMULATIVE ERROR

We assume that for each input prompt x_i , the nonthinking LRM \tilde{f} produces an output \tilde{y}_i , and that there exists a score $U_i \in [0, 1]$ to quantify its uncertainty. This score should ideally correlate with the likelihood of disagreement with the reference model f . The core idea is to use these uncertainty scores to selectively use the expensive model with thinking f . We aim to find a threshold, \hat{u} , and accept the nonthinking LRM’s output \tilde{y}_i for the instances such that $U_i < \hat{u}$, while querying the model with thinking f for the cases where $U_i \geq \hat{u}$. To formalize, we define the cumulative error function conditioned on the uncertainty threshold u :

$$L(u) = \frac{1}{N} \sum_{i=n+1}^{n+N} \ell(y_i, \tilde{y}_i) \mathbf{1}\{U_i \leq u\}. \quad (1)$$

This function measures the average error for test data points with uncertainty scores no greater than u . If we could compute $L(u)$ for all u , we would choose the largest threshold u^* such that $L(u^*) \leq \epsilon$. However, computing $L(u)$ requires access to all expert answers $y_i = f(x_i)$ in the test set, which is computationally expensive. We try an alternative way to build a bound $\hat{L}_u(\alpha)$ for $\mathbb{E}L(u)$ satisfying the following inequality pointwise w.r.t. α :

$$\mathbb{P}(\hat{L}_u(\alpha) \geq \mathbb{E}L(u)) \geq 1 - \alpha. \quad (2)$$

Because the cumulative error function is monotone, we can easily obtain the PAC guarantee as in Definition 1. The monotonicity is a key property in our method, and it allows us to test fixed-sequences single-start without additional corrections (Angelopoulos et al., 2025b); we discuss it in Appendix C, and summarize it in Assumption 3.1.

²For simplicity, we often omit mentioning “with respect to ℓ ” since most tasks have their conventional loss functions

Algorithm 1 Compute Confidence Bound $\hat{L}_u(\alpha)$ Based on Central Limit Theorem

Input: Calibration set $\{(x_i, y_i)\}_{i=1}^n$, model with thinking f , model without thinking \tilde{f} , uncertainty scores $\{U_i\}_{i=1}^n$, sampling weights $\{\pi_i\}_{i=1}^n$, sampling size m , and confidence level α .

Output: The confidence upper bound $\hat{L}_u(\alpha)$.

- 1: Initialize an empty list of samples $\mathcal{Z} = []$.
- 2: Let $\tilde{y}_i = \tilde{f}(x_i)$ for all $i = 1, \dots, n$.
- 3: **for** $j = 1, \dots, m$ **do**
- 4: Sample an index $i_j \sim \text{Unif}(\{1, \dots, n\})$.
- 5: Sample a Bernoulli random variable $\xi_{i_j} \sim \text{Bern}(\pi_{i_j})$.
- 6: **if** $\xi_{i_j} = 1$ **then**
- 7: Query the true label y_{i_j} and compute the importance-weighted loss $Z_j = \ell(y_{i_j}, \tilde{y}_{i_j})/\pi_{i_j}$.
- 8: **else**
- 9: $Z_j = 0$.
- 10: **end if**
- 11: Append Z_j to \mathcal{Z} .
- 12: **end for**
- 13: For a threshold u , define the variables $Z_j(u) = Z_j \cdot \mathbf{1}\{U_{i_j} \leq u\}$ for $j = 1, \dots, m$.
- 14: $\hat{\mu}_Z(u) \leftarrow \frac{1}{m} \sum_{j=1}^m Z_j(u)$
- 15: $\hat{\sigma}_Z(u) \leftarrow \sqrt{\frac{1}{m-1} \sum_{j=1}^m (Z_j(u) - \hat{\mu}_Z(u))^2}$
- 16: $z_{1-\alpha} \leftarrow (1 - \alpha)$ -quantile of the standard normal distribution.
- 17: **Return** $\hat{L}_u(\alpha) \leftarrow \hat{\mu}_Z(u) + z_{1-\alpha} \frac{\hat{\sigma}_Z(u)}{\sqrt{m}}$.

2.2.2 CONSTRUCTING THE UPPER CONFIDENCE BOUND (UCB)

To compute a feasible average error approximation $\hat{L}_u(\alpha)$, we propose a procedure inspired by probably approximately correct labeling (Candès et al., 2025). Given a sampling size m , we first collect m indices $\{i_1, \dots, i_m\}$ by sampling uniformly with replacement from $\{1, \dots, n\}$. Then, for each selected index i_j , we decide whether to query its expert answer y_{i_j} by performing a Bernoulli trial $\xi_{i_j} \sim \text{Bern}(\pi_{i_j})$, where $\{\pi_1, \dots, \pi_n\}$ are sampling weights. This procedure yields a dataset of m i.i.d. random variables:

$$Z_j(u) = \ell(y_{i_j}, \tilde{y}_{i_j}) \frac{\xi_{i_j}}{\pi_{i_j}} \mathbf{1}\{U_{i_j} \leq u\}. \quad (3)$$

The expectation of $Z_j(u)$ is equals the target quantity $L(u)$, since $\mathbb{E}[\xi_{i_j}/\pi_{i_j}|i_j] = 1$. We can therefore estimate an upper bound for $L(u)$ by computing a confidence interval for the mean of $\{Z_j(u)\}_{j=1}^m$. We formally described it in the central limit theorem (CLT) based Algorithm 1.

Remark 3. The procedure in Algorithm 1 uses importance sampling to construct an unbiased estimator for the true error $L(u)$. For any fixed threshold u , the random variables $Z_j(u)$ are i.i.d. with expectation $\mathbb{E}[Z_j(u)] = L(u)$. This holds because the sampling process decouples the choice of index i_j from the decision to query the label y_{i_j} . Given the samples $\{Z_j(u)\}_{j=1}^m$, we can form an upper bound for $L(u)$. Algorithm 1 illustrates this using a CLT-based approach, valid for large m . See Appendix D for discussion about its UCB validation as Assumption 3.1. Alternatively, if the importance-weighted losses are bounded, one could use concentration inequalities like Hoeffding's or Bernstein's inequality to construct a valid confidence bound (Bentkus, 2004; Hao et al., 2019; Hoeffding, 1994; Learned-Miller & Thomas, 2020; Ramdas et al., 2022; Waudby-Smith & Ramdas, 2021; 2024), which may provide better guarantees for smaller sample sizes. We give an example in Algorithm 3 based on Hoeffding's inequality (Hoeffding, 1994).

2.2.3 CALIBRATION

Once the UCB $\hat{L}_u(\alpha)$ is constructed, the threshold \hat{u} is the highest uncertainty level for which this estimated error bound remains below the tolerance ϵ :

$$\hat{u} = \max\{u \in [0, 1] : \hat{L}_u(\alpha) \leq \epsilon\}. \quad (4)$$

We calibrate the \hat{u} on the calibration set \mathcal{D}_{cal} , and apply it to the test sample. If the uncertainty score is larger than \hat{u} , we use the thinking model to answer; otherwise, we use the nonthinking mode. This

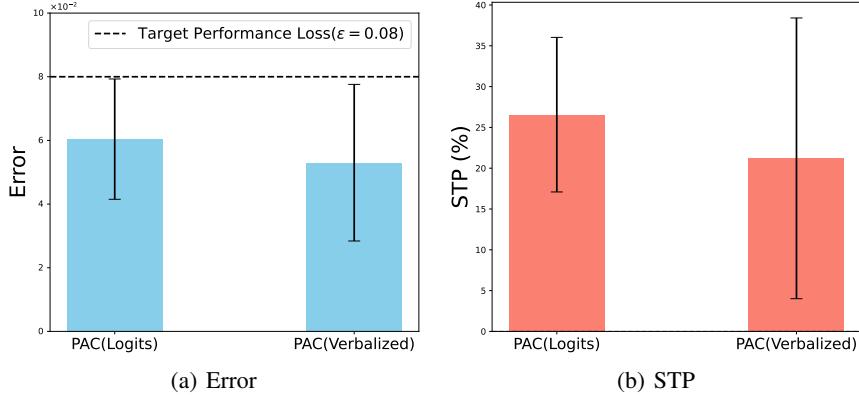


Figure 1: **Error control and saved token percentage (STP) of PAC reasoning**, with binary loss on ZebraLogic at a confidence level 95%. “PAC(Logits)” and “PAC(Verbalized)” present PAC reasoning using the logits-based score and the verbalized score. Both control the performance loss under the target 0.08 and save at least 20%. All experiments are repeated 100 times, and other details follow as the main experiment in Section 4.1 and Appendix H.

procedure ensures we can accept as many outputs from the nonthinking model \tilde{f} while controlling the overall performance loss with high probability. We summarize the PAC reasoning algorithm in Algorithm 2. And we show the error and saved token percentage on the dataset “ZebraLogic” using PAC reasoning with logits-based score and verbalized score in Figure 1. Our method controls the performance loss below the target loss 0.08 and saves token cost at least 20% with confidence 95%. As for the naive method discussed in Appendix I, switching model using a fixed threshold, it could not control the performance, and may incur more token usage, see details in Table 3.

Algorithm 2 PAC Reasoning

Input: Calibration set $\{(x_i, y_i)\}_{i=1}^n$, test prompts $x_i, i \in \mathcal{I}_{test}$, model without thinking \tilde{f} , model with thinking f , loss function ℓ , error tolerance ϵ , confidence level α

Output: PAC reasoning model \hat{f} on test set \mathcal{I}_{test}

- 1: $\forall i \in \mathcal{I}_{cal}$, compute outputs without thinking $\tilde{y}_i = \tilde{f}(x_i)$ and uncertainty scores U_i .
 - 2: Compute confidence bound function $\hat{L}_u(\alpha)$ via Algorithm 1 on calibration data $\{(x_i, y_i)\}_{i=1}^n$.
 - 3: Determine threshold $\hat{u} = \max\{u \in \{U_i\}_{i \in \mathcal{I}_{cal}} : \hat{L}_u(\alpha) \leq \epsilon\}$.
 - 4: $\forall i \in \mathcal{I}_{test}$, compute outputs without thinking $\tilde{y}_i = \tilde{f}(x_i)$ and uncertainty scores U_i independently.
 - 5: $\hat{y}_i \leftarrow f(x_i) \mathbb{1}\{U_i \geq \hat{u}\} + \tilde{y}_i \mathbb{1}\{U_i < \hat{u}\}$ for all $i \in \mathcal{I}_{test}$.
 - 6: Define the composite model \hat{f} by its outputs: $\hat{f}(x_i) = \hat{y}_i$ for all $i \in \mathcal{I}_{test}$.
 - 7: **Return** \hat{f} .
-

3 THEORETICAL ANALYSIS

In this section, we aim to introduce the PAC guarantee. Our PAC reasoning builds upon the theoretical foundation established by the distribution-free risk control framework (Angelopoulos et al., 2025b). As discussed in Appendix B, our PAC reasoning problem can be viewed as a generalization of the distribution-free risk control method. It provides the mathematical foundation for our risk control type method on the characteristics of PAC reasoning. In detail, if the performance loss $L(u)$ is bounded by a UCB, and the UCB based on CLT or concentration inequality is valid as Assumption 3.1, we can prove the PAC guarantee as follows.

3.1 PAC GUARANTEE

First, noting that while the confidence bound $\hat{L}_u(\alpha)$ is constructed to hold for a single, pre-specified threshold u , our algorithm selects the threshold \hat{u} based on the calibration data. Let \mathcal{D}_{cal} be a calibration set, used to construct a threshold \hat{u} , and let \mathcal{D}_{test} be an independent test set with i.i.d. samples as \mathcal{D}_{cal} . For any threshold u , recalling the deployment strategy $T_u(x)$ as Eq. (6), we could see that $\hat{f} = T_{\hat{u}}$ is the composite model. We re-parameterize \hat{f} and its risk with respect to u , with a slight abuse of notation. Its population risk $R(\hat{f})$ is re-parameterized as:

$$R(u) = \mathbb{E}[\ell(y, T_u(x))],$$

and the empirical risk is re-parameterized as:

$$\widehat{R}(u) = \frac{1}{N} \sum_{i \in \mathcal{I}_{test}} \ell(y_i, T_u(x_i)).$$

Then we list the assumptions of the PAC reasoning.

Assumption 3.1 (UCB validity). *For each threshold u and any $\alpha \in (0, 1)$, there exists a UCB $\hat{L}_u(\alpha)$, computed on \mathcal{D}_{cal} , such that*

$$\mathbb{P}(R(u) \leq \hat{L}_u(\alpha)) \geq 1 - \alpha.$$

As discussed in Remark 3, we can build the UCB $\hat{L}_u(\alpha)$ for $\mathbb{E}L(u)$ in two main ways: using the central limit theorem, or using bounds like Hoeffding's or Bernstein's inequality (Bentkus, 2004; Hoeffding, 1994; Waudby-Smith & Ramdas, 2021; 2024), which may provide better guarantees for smaller sample sizes. We prove the validity of the CLT-based method in Appendix D. The risk function $R(u)$ is naturally non-decreasing with u . As u increases, the condition $U(x) \geq u$ becomes harder to satisfy, so we defer to the expert less often. Then, reducing deference to the expert can only increase the total risk under Assumption 3.1. We gather the above assumptions and the monotonicity, and provide the PAC guarantee of our proposed method:

Theorem 4 (PAC guarantee). *Let \hat{u} be the threshold selected by the PAC reasoning algorithm (Algorithm 2). If calibration set and test set are i.i.d. and Assumption 3.1 holds, then the composite model \hat{f} constructed by Algorithm 2 satisfies the (ϵ, α) -PAC guarantee, i.e.,*

$$\mathbb{P}(R(\hat{f}) \leq \epsilon) \geq 1 - \alpha.$$

We prove it in Appendix C. If the loss is bound in $[a, b]$, we provide an empirical version:

Theorem 5 (Empirical risk PAC guarantee). *Assume Assumption 3.1 holds, and the test batch \mathcal{D}_{test} is independent of the calibration data \mathcal{D}_{cal} . Given $\epsilon, \alpha \in (0, 1)$, $\ell \in [a, b]$, and \hat{u} defined as in Theorem 4, then any $t > 0$,*

$$\mathbb{P}(\widehat{R}(\hat{u}) \leq \epsilon + t) \geq 1 - \alpha - \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

Remark 6. A common special case is a bounded loss $\ell \in [0, 1]$, e.g., 0-1 loss for binary verifiable answers. Then $b-a=1$ and the bound simplifies to $\mathbb{P}(\widehat{R}(\hat{u}) \leq \epsilon + t) \geq 1 - \alpha - e^{-2Nt^2}$. It provides exact risk control for \hat{f} with probability at least $1 - \alpha$ by some slacks t .

We prove it in Appendix E. If the i.i.d. assumption does not hold, PAC reasoning can be extended to a transductive setting. This extension is discussed in Appendix G.

4 EXPERIMENTS

In this section, we present the experimental results that evaluate the performance loss and computational savings of the proposed PAC reasoning across diverse benchmarks, including mathematical reasoning (Lightman et al., 2023), logical deduction (Lin et al., 2025), and text generation tasks (Li et al., 2024). We aim to verify whether the proposed method could control the performance loss with a confidence level while saving computational resources. We evaluate PAC reasoning under different uncertainty estimators (logits-based and verbalized uncertainty score), as well as efficiency

metrics (expert calling percentage and saved token percentage). Experimental setup details are provided in Section 4.1. Across diverse benchmarks, PAC reasoning consistently controls performance loss under the desired tolerance while saving computation, regardless of the uncertainty estimator or loss function, as shown in Figure 2 and Table 3. We also discuss the behaviors of PAC reasoning and a naive method among different loss functions, uncertainty scores in Section 4.2.

4.1 SETUP

Large language models In this study, we evaluate the PAC reasoning based on the Qwen3 series models (Yang et al., 2025a). Specifically, we employ the “Qwen3-4B-Thinking-2507” as the thinking model and “Qwen3-4B-Instruct-2507” as the lower-performance nonthinking model. The sampling temperature and other hyperparameters for both LLMs are configured following the settings in the original paper. Details can be found in Appendix H.

Uncertainty score We adopt two complementary perspectives to quantify the uncertainty of \tilde{y} : a white-box score derived from model logits and a black-box score obtained from verbalized self-reports. We leverage token-level probabilities computed from the prediction logits (Kwon et al., 2023; Zheng et al., 2024) for the white-box score. Furthermore, we define the uncertainty score of y_i as its average token probability (Hao et al., 2023; Huang et al., 2025), given by:

$$U_{\text{logits}}(y_i) = 1 - \frac{1}{l_{y_i}} \sum_{j=1}^{l_{y_i}} \mathbb{P}(y_{i,j}|y_{i,1}, \dots, y_{i,j-1}, x_i),$$

where $\mathbb{P}(y_{i,j}|y_{i,1}, \dots, y_{i,j-1}, x_i)$ is the conditional probability of token $y_{i,j}$. Moreover, we also consider verbalized uncertainty scores from nonthinking models (Xiong et al., 2023; Tian et al., 2023; Yang et al., 2024), where the model explicitly states its self-reported confidence. The verbalized uncertainty score is mainly applicable in black-box scenarios, where access to generation logits is restricted, especially in the case of closed-source LLMs. In this study, we report the average confidence over 10 trials, and the corresponding prompts are listed in Appendix H.

Datasets We evaluate PAC reasoning on a series of real datasets spanning reasoning and open-ended generation tasks. Specifically, our evaluation covers a high-level mathematics benchmark, MATH-500 (Lightman et al., 2023), a text-based logical reasoning task, ZebraLogic (Lin et al., 2025), and an alignment-focused open-ended writing benchmark, Arena-Hard (Li et al., 2024). For each dataset, we partition the original test set into a calibration subset and a held-out test subset randomly. Table 2 in Appendix H provides details on the specific splitting strategies.

Loss functions We consider two types of loss functions for evaluating the PAC guarantee of our method: the semantic cosine distance and binary 0-1 loss. Details are placed in Appendix H.

Metrics To evaluate the effectiveness of the PAC reasoning in optimizing budget usage, we define two key metrics: *Expert Call Percentage* (ECP) and *Saved Token Percentage* (STP). These metrics are formally defined as follows:

$$\text{ECP} := \frac{|\{i : U_i \geq \hat{u}, i \in \mathcal{I}_{\text{test}}\}|}{N} \times 100\%, \quad \text{STP} := \frac{1}{N} \sum_{i \in \mathcal{I}_{\text{test}}} \frac{l_{\tilde{y}_i} + \mathbb{1}\{U_i \geq \hat{u}\}l_{y_i}}{l_{y_i}} \times 100\%, \quad (5)$$

where $l_{\tilde{y}_i}$ and l_{y_i} represent the number of tokens in the candidate answer \tilde{y}_i and the reference answer y_i , respectively. The ECP measures the proportion of test cases requiring expert intervention. At the same time, the STP quantifies the token efficiency by comparing the token counts of candidate and reference answers, accounting for cases where expert calls are triggered.

We repeat each experiment 100 times and report the mean and standard deviation of the budget savings. We fix $\alpha = 0.05$ throughout all experiments while varying ϵ , and set the sampling weight $\pi = \pi_i = 0.5$ for each $i \in \mathcal{I}_{\text{cal}}$ and the sample size $m = n \times \frac{1}{\pi}$.

4.2 RESULTS

PAC reasoning improves the efficiency and guarantees the performance loss. Under the semantic loss (Figure 2), PAC reasoning consistently ensures validity of error across all the benchmarks:

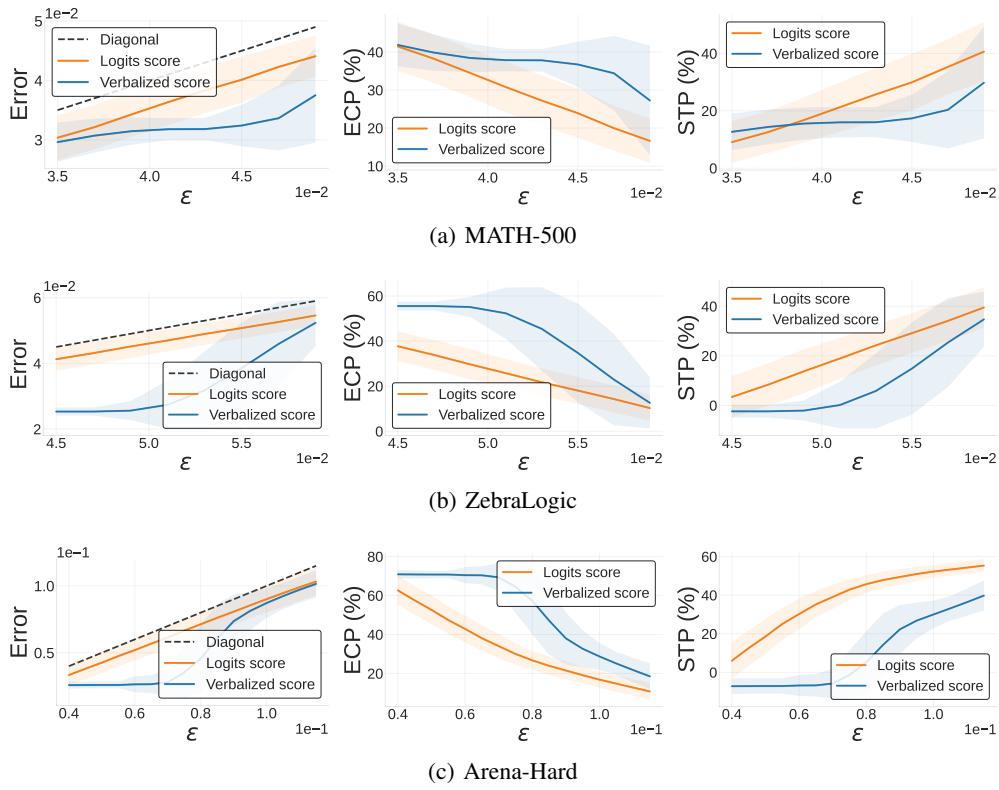


Figure 2: **Error control, ECP and STP of PAC reasoning** for semantic loss across three benchmarks at a confidence level of $\alpha = 0.05$. Uncertainty score includes the logits-based score and the verbalized score. All experiments are repeated 100 times, and the shaded areas represent standard deviations.

the empirical error remains below the target risk level ϵ while achieving substantial budget savings. For instance, on Arena-Hard with $\epsilon = 0.08$ for the logits uncertainty score, the average empirical error is approximately 0.06, the ECP is about 20%, and the STP is around 40%. Under the 0-1 binary loss (see Figure 1 and Table 3), PAC reasoning also maintains error rates within the target risk level and saves computational budget. For example, for the MATH-500 dataset, PAC reasoning saves ECP by 22.50% and STP by 23.13%. In summary, PAC reasoning bounds the performance loss within the target risk level and significantly improves inference efficiency.

Logits-based uncertainty score is more stable. From the results in Figure 2 and Table 3, the logits-based uncertainty score consistently shows smoother and more stable behavior, with lower variance in both ECP and STP. In contrast, the verbalized uncertainty score exhibits larger fluctuations due to its sparse and clustered distribution. For instance, on ZebraLogic (see Table 3), the standard deviations of ECP and STP under the verbalized score are 20.68 and 17.20, considerably higher than the corresponding 7.47 and 9.90 values for the logits-based score. Similar patterns can also be observed in Figure 2, where the variance of the verbalized score is consistently higher than that of the logits-based score across different settings. Moreover, in Figure 2, the verbalized score exhibits overly conservative error control and fails to improve inference efficiency when ϵ is small. Therefore, although the verbalized score occasionally achieves stronger risk control or higher savings, its calibration is less reliable, leading to less consistent performance.

5 RELATED WORK

Our work intersects efficiency improvement for reasoning models and the distribution-free inference for risk control of its performance loss with confidence.

Efficiency improvement for reasoning models Large Reasoning Models (LRMs) have recently become a research hotspot due to their outstanding performance in handling complex tasks (Yue et al., 2025). However, the problem of overthinking has emerged (Sui et al., 2025; Chen et al., 2025), where LRMs tend to engage in unnecessarily long reasoning chains and redundant computational steps. This increases latency and cost, and may even cause error accumulation through extended reasoning paths. For example, in mathematical problems, LRMs may explore irrelevant solution branches or perform excessive intermediate calculations that do not contribute to the final answer. To alleviate this issue, recent studies propose efficient reasoning strategies such as Early Exit of thinking (Yang et al., 2025b; Jiang et al., 2025) and adaptive switching between “fast” and “slow” thinking modes to reduce reasoning tokens and avoid redundant steps (Cheng et al., 2025; Chung et al., 2025; Fang et al., 2025; Li et al., 2025; Liang et al., 2025; Ma et al., 2025; Paliotta et al., 2025; Pan et al., 2024a;b; 2025; Xiao et al., 2025; Yao et al., 2024; Yong et al., 2025). Despite their empirical effectiveness, these techniques lack theoretical guarantees on the performance loss when using the nonthinking mode. We fill this gap by introducing a PAC-based reasoning that provides statistically guaranteed performance loss for efficient reasoning.

PAC learning and Learn-then-Test framework PAC learning (Valiant, 1984), and the LTT framework (Bates et al., 2021; Angelopoulos et al., 2025b) provide a theoretical foundation for reliable machine learning. PAC learning establishes how algorithms can generalize from training data with probabilistic guarantees, addressing questions of sample complexity and error bounds. The LTT framework is a useful distribution-free inference method (Angelopoulos et al., 2025a; Gibbs et al., 2025; Lei & Wasserman, 2014; Lei et al., 2018), offering an approach for calibrating predictive algorithms to achieve risk control, enabling practitioners to specify error tolerances and confidence levels while constructing systems that satisfy these requirements with high probability. The key innovation of LTT is separating the learning phase from the testing phase (Zeng et al., 2025), where performance guarantees are established. This framework has been extended to conformal risk control (Angelopoulos et al., 2025c) and its applications in language modeling (Quach et al., 2024), allowing for flexible control of monotone loss functions while maintaining distribution-free guarantees. This work applies the LTT framework to reasoning model efficiency improvement and rigorously validates its theoretical properties. This application enables us to provide formal guarantees for the trade-off between computational efficiency and reasoning accuracy for LRMs.

6 CONCLUSION

We propose PAC reasoning, which is designed to provide rigorous, theoretically grounded, and practical efficiency improvement for reasoning models while maintaining probabilistic correctness guarantees. Our approach constructs a composite model that selectively uses either an expert model or a candidate model based on a constructed confidence bound. The PAC reasoning contributes by delivering statistical assurances for model performance and demonstrating that it reliably achieves the specified error rate with probability. We validate our method through extensive experiments on real-world reasoning tasks, showing it can significantly reduce computational costs while maintaining user-specified error tolerances with confidence. Future research will focus on developing more advanced uncertainty estimation techniques, exploring tighter theoretical bounds, and broadening the method’s applicability to other large language model efficiency-improvement strategies.

Limitations Though our PAC reasoning provides rigorous statistical guarantees, there are several limitations worth noting: First, the method requires access to uncertainty scores from the model without thinking, which may not always be available or reliable. Second, the sampling-based confidence bound estimation introduces additional computational overhead during calibration. Finally, the method assumes that the calibration and test distributions are i.i.d. meaning that the calibration set is drawn from the same distribution as the test set. Significant distribution shifts could invalidate the guarantees. These limitations suggest future work improving the method’s robustness and applicability to a broader range of reasoning models and tasks.

REFERENCES

- Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical Foundations of Conformal Prediction. *arXiv preprint arXiv:2411.11824*, 2025a.
- Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19:1641–1662, 2025b.
- Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2025c.
- Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.
- Vidmantas Bentkus. On Hoeffding’s inequalities. *The Annals of Probability*, 32, 2004.
- Emmanuel J. Candès, Andrew Ilyas, and Tijana Zrnic. Probably approximately correct labels. *arXiv preprint arXiv:2506.10908*, 2025.
- Chen, Qin, Liu, Peng, Guan, Wang, and et al. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. *arXiv preprint arXiv:2503.09567*, 2025.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*, 2025.
- Stephen Chung, Wenyu Du, and Jie Fu. Thinker: Learning to think fast and slow. *arXiv preprint arXiv:2505.21097*, 2025.
- DeepSeek-AI, Guo, Yang, Zhang, Song, Zhang, and et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: LLM learns when to think. *arXiv preprint arXiv:2505.13379*, 2025.
- Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, 2025.
- Hao, Gu, Ma, Hong, Wang, Wang, and et al. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In N. I. Fisher and P. K. Sen (eds.), *The Collected Works of Wassily Hoeffding*, pp. 409–426. Springer, New York, NY, 1994. ISBN 978-1-4612-0865-5. doi: 10.1007/978-1-4612-0865-5_26.
- Huang, Song, Wang, Zhao, Chen, Juefei-Xu, and et al. Look before you leap: An exploratory study of uncertainty measurement for large language models. *IEEE Transactions on Software Engineering*, 51:413–429, 2025.
- Guochao Jiang, Guofeng Quan, Zepeng Ding, Ziqin Luo, Dixuan Wang, and Zheng Hu. FlashThink: An early exit method for efficient reasoning. *arXiv preprint arXiv:2505.13949*, 2025.
- Kwon, Li, Zhuang, Sheng, Zheng, Yu, and et al. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP ’23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0229-7. doi: 10.1145/3600006.3613165.

-
- Erik Learned-Miller and Philip S. Thomas. A new confidence interval for the mean of a bounded random variable. *arXiv preprint arXiv:1905.06208*, 2020.
- Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:71–96, 2014.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 2018.
- Li, Chiang, Frick, Dunlap, Wu, Zhu, and et al. From crowdsourced data to high-quality benchmarks: Arena-hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.
- Li, Wei, Huang, Yan, Chen, Kwok, and et al. DynamicMind: A tri-mode thinking system for large language models. *arXiv preprint arXiv:2506.05936*, 2025.
- Guosheng Liang, Longguang Zhong, Ziyi Yang, and Xiaojun Quan. ThinkSwitcher: When to think hard, when to think fast. *arXiv preprint arXiv:2505.14183*, 2025.
- Lightman, Kosaraju, Burda, Edwards, Baker, Lee, and et al. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- Lin, Bras, Richardson, Sabharwal, Poovendran, Clark, and et al. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *Forty-Second International Conference on Machine Learning*, 2025.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.
- NVIDIA, Basant, Khairnar, Paithankar, Khattar, Renduchintala, and et al. NVIDIA nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.
- Paliotta, Wang, Pagliardini, Li, Bick, Kolter, and et al. Thinking slow, fast: Scaling inference compute with distilled reasoners. *arXiv preprint arXiv:2502.20339*, 2025.
- Pan, Zhang, Zhang, Liu, Wang, Li, and et al. DynaThink: Fast or slow? A dynamic decision-making framework for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14686–14695, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.814.
- Pan, Ji, Ding, Li, Chen, Wang, and et al. A survey of slow thinking-based reasoning LLMs using reinforced learning and inference-time scaling law. *arXiv preprint arXiv:2505.02665*, 2025.
- Jiabao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, and Haizhou Li. DynaThink: Fast or slow? A dynamic decision-making framework for large language models. *arXiv preprint arXiv:2407.01009*, 2024b.
- Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, and et al. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2024.
- Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2022.
- Roller, Dinan, Goyal, Ju, Williamson, Liu, and et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020.
- Sui, Chuang, Wang, Zhang, Zhang, Yuan, and et al. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. *Transactions on Machine Learning Research*, 2025.

-
- Tian, Mitchell, Zhou, Sharma, Rafailov, Yao, and et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330.
- L. G. Valiant. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing - STOC '84*, pp. 436–445, Not Known, 1984. ACM Press. ISBN 978-0-89791-133-7. doi: 10.1145/800057.808710.
- Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *arXiv preprint arXiv:2006.04347*, 2021.
- Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86:1–27, 2024.
- Xiao, Gan, Dai, He, Huang, Li, and et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.
- Xiong, Hu, Lu, Li, Fu, He, and et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2023.
- Yang, Li, Yang, Zhang, Hui, Zheng, and et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Yang, Si, Duan, Zhu, Zhu, Li, and et al. Dynamic early exit in reasoning models. *arXiv preprint arXiv:2504.15895*, 2025b.
- Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. *arXiv preprint arXiv:2412.14737*, 2024.
- Wenlin Yao, Haitao Mi, and Dong Yu. HDFlow: Enhancing LLM complex problem-solving with hybrid thinking and dynamic workflows. *arXiv preprint arXiv:2409.17433*, 2024.
- Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. Think or not? Exploring thinking efficiency in large reasoning models via an information-theoretic lens. *arXiv preprint arXiv:2505.18237*, 2025.
- Yue, Du, Wang, Gao, Yao, Wang, and et al. Don't overthink it: A survey of efficient R1-style large reasoning models. *arXiv preprint arXiv:2508.02120*, 2025.
- Hao Zeng, Kangdao Liu, Bingyi Jing, and Hongxin Wei. Parametric scaling law of tuning bias in conformal prediction. In *Forty-Second International Conference on Machine Learning*, 2025.
- Zhang, Roller, Goyal, Artetxe, Chen, Chen, and et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Zheng, Yin, Xie, Sun, Huang, Yu, and et al. SGLang: Efficient execution of structured language model programs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

A LLM USAGE

This work was developed with the help of large language models (LLMs) include GPT 5 and Gemini 2.5 pro, which helped enhance the clarity, coherence, and technical precision of the writing. The authors' original research contributions, intellectual property, and key arguments remain entirely their own. All content has undergone thorough review and validation by the authors to ensure accuracy and scientific integrity.

B FROM LEARN-THEN-TEST TO PAC REASONING

The LTT framework considers a family of post-processing mappings $\{T_\lambda\}_{\lambda \in \Lambda}$. It aims to select a parameter $\hat{\lambda}$ such that the risk $R(T_{\hat{\lambda}})$ is controlled with confidence. Our PAC reasoning problem naturally fits into this framework through the following construction:

Mapping into LTT framework We define the parameter space as $\Lambda = \{U_1, \dots, U_n\}$, where U_i are the uncertainty scores on the calibration set. This discretization to sample points is consistent with the implementation of our algorithm. For each threshold $u \in \Lambda$, we define the mapping:

$$T_u(x_i) := \begin{cases} y_i, & U_i \geq u \quad (\text{query the model with thinking}), \\ \tilde{y}_i, & U_i < u \quad (\text{use the model without thinking}). \end{cases} \quad (6)$$

This mapping T_u , where $u \in \Lambda$, corresponds precisely to our uncertainty-based selection strategy in the PAC reasoning framework. We define the empirical risk as the average performance loss:

$$\hat{R}(T_u) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, T_u(x_i)) = L(u), \quad (7)$$

where $L(u)$ is our cumulative error function defined in (2). The LTT framework requires valid p-values for each parameter λ . Our confidence bound construction in Algorithm 1 provides this. For each threshold u , our confidence upper bound $\hat{L}_u(\alpha)$ satisfies

$$\mathbb{P}(\mathbb{E}L(u) \leq \hat{L}_u(\alpha)) \geq 1 - \alpha.$$

This can be converted to a valid p-value for testing the null hypothesis $H_u : L(u) > \epsilon$ by defining

$$p_u := \inf\{\gamma \in (0, 1) : \hat{L}_u(\gamma) \leq \epsilon\}.$$

Under the null hypothesis H_u , this p-value is stochastically dominated by the uniform distribution on $[0, 1]$, satisfying the validity requirement of the LTT framework.

Monotonicity and fixed-sequence testing A crucial property of our setup is that the risk function $u \mapsto L(u)$ is monotonically non-decreasing in u . This monotonicity allows us to use the fixed-sequence testing procedure from the LTT framework without additional multiple testing corrections (e.g., Benjamini-Hochberg procedure, Bonferroni correction, or Holm's step-down method). Specifically, our threshold selection $\hat{u} = \max\{u \in \Lambda : \hat{L}_u(\alpha) \leq \epsilon\}$ is equivalent to the fixed-sequence single-start procedure in LTT, which maintains the family-wise error rate control at level α . Therefore, the theoretical guarantees of our PAC reasoning inherit the rigorous foundation of the LTT framework, ensuring that our risk control is valid despite the data-dependent threshold selection.

C PROOF OF THEOREM 4

Proof. Let

$$u^* := \inf\{u \in \Lambda : R(u) > \epsilon\}.$$

As $R(u)$ is non-decreasing, it holds that

$$R(u) \leq \epsilon \text{ for all } u \leq u^*,$$

and

$$R(u^*) > \epsilon.$$

If $\hat{u} > u^*$, then

$$\hat{L}_{u^*}(\alpha) \leq \epsilon < R(u^*),$$

which contradicts Assumption 3.1 except with probability at most α .

Therefore,

$$\mathbb{P}(\hat{u} \leq u^*) \geq 1 - \alpha,$$

and because $R(u)$ is non-decreasing,

$$R(\hat{u}) \leq \epsilon \text{ on this event.}$$

□

D VALIDITY OF CLT-BASED UPPER CONFIDENCE BOUND

We show that a CLT-based upper confidence bound computed on the calibration set satisfies Assumption 3.1 asymptotically.

Proposition 7 (Asymptotic validity of UCB based on CLT). *Fix a threshold u and let $Z_j(u)$ be the i.i.d. random variables defined in Algorithm 1 for $j = 1, \dots, m$, with mean $L(u)$ and variance $\sigma_Z^2(u) > 0$. Let $\hat{\mu}_Z(u) = m^{-1} \sum_{j=1}^m Z_j(u)$ and $\hat{\sigma}_Z^2(u) = (m-1)^{-1} \sum_{j=1}^m (Z_j(u) - \hat{\mu}_Z(u))^2$. Define the UCB based on CLT*

$$\hat{L}_u^{\text{CLT}}(\alpha) := \hat{\mu}_Z(u) + z_{1-\alpha} \sqrt{\hat{\sigma}_Z^2(u)/m},$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$ -quantile of the standard normal distribution. Then

$$\liminf_{m \rightarrow \infty} \mathbb{P}(L(u) \leq \hat{L}_u^{\text{CLT}}(\alpha)) \geq 1 - \alpha.$$

Proof. By the classical Lindeberg–Feller central limit theorem,

$$\frac{\sqrt{m} (\hat{\mu}_Z(u) - L(u))}{\sigma_Z(u)} \rightarrow \mathcal{N}(0, 1)$$

in distribution as $m \rightarrow \infty$ because the variables are i.i.d. Since $\hat{\sigma}_Z(u) \xrightarrow{p} \sigma_Z^2(u)$ by the weak law of large numbers, Slutsky's theorem yields

$$\frac{\sqrt{m} (\hat{\mu}_Z(u) - L(u))}{\sqrt{\hat{\sigma}_Z(u)}} \rightarrow \mathcal{N}(0, 1).$$

Therefore

$$\mathbb{P}\left(\frac{L(u) - \hat{\mu}_Z(u)}{\sqrt{\hat{V}_Z(u)/m}} \leq z_{1-\alpha}\right) \rightarrow 1 - \alpha.$$

Equivalently, define $d_m := (L(u) - \hat{\mu}_Z(u))/\sqrt{\hat{V}_Z(u)/m}$ and observe that

$$\{d_m \leq z_{1-\alpha}\} = \left\{L(u) \leq \hat{\mu}_Z(u) + z_{1-\alpha} \sqrt{\hat{V}_Z(u)/m}\right\}.$$

Hence,

$$\liminf_{m \rightarrow \infty} \mathbb{P}(L(u) \leq \hat{L}_u^{\text{CLT}}(\alpha)) \geq 1 - \alpha.$$

□

E PROOF OF THEOREM 5

E.1 NOTATION RECALLING AND LEMMA

We present PAC guarantees for the empirical test risk under precise Hoeffding conditions, making explicit the roles of calibration and test randomness. For any threshold u , the deployment rule T_u predicts with the expert when $U(x) \geq u$ and otherwise uses the fast model. The population risk is

$$R(u) = \mathbb{E}_{(x,y) \sim P} [\ell(y, T_u(x))].$$

Given an independent test set $\mathcal{D}_{\text{test}}$ drawn i.i.d. from P , the empirical test risk is

$$\widehat{R}(u) = \frac{1}{N} \sum_{j=n+1}^{n+N} \ell(y_j, T_u(x_j)).$$

Our guarantee for the empirical test risk relies on Assumption 3.1 from the main text, in addition to the following lemma.

Lemma 8 (Conditional Hoeffding bound). *Let \hat{u} be a random variable determined by the calibration set \mathcal{D}_{cal} . Assume that, conditioned on \hat{u} , the test losses $Z_j(\hat{u}) := \ell(y_j, T_{\hat{u}}(x_j))$ for $j \in \mathcal{I}_{\text{test}}$ are i.i.d. and bounded in $[a, b]$. Then for any $t > 0$,*

$$\mathbb{P}(\widehat{R}(\hat{u}) - R(\hat{u}) > t \mid \hat{u}) \leq \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

Proof. Let $Z_j(\hat{u}) = \ell(y_j, T_{\hat{u}}(x_j))$ for $j \in \mathcal{I}_{\text{test}}$. The model \hat{u} is fixed when we condition on it. Since the test set $\mathcal{D}_{\text{test}}$ consists of i.i.d. samples and is independent of \hat{u} , the random variables $Z_1(\hat{u}), \dots, Z_N(\hat{u})$ are conditionally independent and identically distributed.

By the boundness of the loss function, each $Z_j(\hat{u})$ is bounded in $[a, b]$. The conditional expectation of each $Z_j(\hat{u})$ is $\mathbb{E}[Z_j(\hat{u}) \mid \hat{u}] = \mathbb{E}[\ell(y_j, T_{\hat{u}}(x_j)) \mid \hat{u}]$. Since (x_j, y_j) is independent of \hat{u} , this is equal to the unconditional expectation over the data distribution, $\mathbb{E}_{(x,y) \sim P} [\ell(y, T_{\hat{u}}(x))]$, which is the definition of the true risk $R(\hat{u})$. The empirical risk is the sample mean:

$$\widehat{R}(\hat{u}) = \frac{1}{N} \sum_{j=1}^N Z_j(\hat{u}).$$

Its conditional expectation is

$$\mathbb{E}[\widehat{R}(\hat{u}) \mid \hat{u}] = \frac{1}{N} \sum_{j=1}^N \mathbb{E}[Z_j(\hat{u}) \mid \hat{u}] = R(\hat{u}).$$

We can now apply Hoeffding's inequality (Hoeffding, 1963) to the conditional i.i.d. bounded variables $Z_j(\hat{u})$. For any $t > 0$, the one-sided version states that

$$\mathbb{P}\left(\frac{1}{N} \sum_{j=n+1}^{n+N} Z_j(\hat{u}) - \mathbb{E}\left[\frac{1}{N} \sum_{j=n+1}^{n+N} Z_j(\hat{u}) \mid \hat{u}\right] > t \mid \hat{u}\right) \leq \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

Substituting the empirical risk and true risk, we get

$$\mathbb{P}(\widehat{R}(\hat{u}) - R(\hat{u}) > t \mid \hat{u}) \leq \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

This completes the proof. □

E.2 PROOF DETAILS

Proof. The independence of $\mathcal{D}_{\text{test}}$ and \mathcal{D}_{cal} ensures Lemma 8 hold. Use the inclusion

$$\{\widehat{R}(\hat{u}) > \epsilon + t\} \subseteq \{R(\hat{u}) > \epsilon\} \cup \{\widehat{R}(\hat{u}) - R(\hat{u}) > t\}$$

and take probabilities. Combine the result of Theorem 4 (i.e., $\mathbb{P}(R(\hat{u}) > \epsilon) \leq \alpha$) with the law of total probability and Lemma 8 to conclude the claim:

$$\begin{aligned}\mathbb{P}(\hat{R}(\hat{u}) > \epsilon + t) &\leq \mathbb{P}(R(\hat{u}) > \epsilon) + \mathbb{P}(\hat{R}(\hat{u}) - R(\hat{u}) > t) \\ &= \mathbb{P}(R(\hat{u}) > \epsilon) + \mathbb{E}[\mathbb{P}(\hat{R}(\hat{u}) - R(\hat{u}) > t | \hat{u})] \\ &\leq \alpha + \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).\end{aligned}$$

This is equivalent to the stated guarantee. \square

F FINITE-SAMPLE CONFIDENCE BOUNDS FOR BOUNDED LOSSES

This section presents an alternative algorithm for computing confidence bounds that provides strict finite-sample guarantees under bounded loss assumptions. This approach leverages concentration inequalities such as the Hoeffding inequality or betting-based confidence intervals (Bentkus, 2004; Hao et al., 2019; Hoeffding, 1994; Learned-Miller & Thomas, 2020; Ramdas et al., 2022; Waudby-Smith & Ramdas, 2021; 2024) to achieve tighter bounds compared to the asymptotic normal approximation used in Algorithm 1.

Algorithm 3 Compute Confidence Bound $\hat{L}_u(\alpha)$ Based on Hoeffding Inequality

Input: Calibration set $\{(x_i, y_i)\}_{i=1}^n$, model with thinking f , model without thinking \tilde{f} , uncertainty scores $\{U_i\}_{i=1}^n$, sampling weights $\{\pi_i\}_{i=1}^n$, sampling size m , confidence level α , loss upper bound $B > 0$

Output: The finite-sample confidence upper bound $\hat{L}_u(\alpha)$.

- 1: Initialize an empty list of samples $\mathcal{Z} = []$.
 - 2: Let $\tilde{y}_i = \tilde{f}(x_i)$ for all $i = 1, \dots, n$.
 - 3: **for** $j = 1, \dots, m$ **do**
 - 4: Sample an index $i_j \sim \text{Unif}(\{1, \dots, n\})$.
 - 5: Sample a Bernoulli random variable $\xi_{i_j} \sim \text{Bern}(\pi_{i_j})$.
 - 6: **if** $\xi_{i_j} = 1$ **then**
 - 7: Query the true label y_{i_j} and compute $Z_j = \min\left(\frac{\ell(y_{i_j}, \tilde{y}_{i_j})}{\pi_{i_j}}, \frac{B}{\pi_{i_j}}\right)$.
 - 8: **else**
 - 9: $Z_j = 0$.
 - 10: **end if**
 - 11: Append Z_j to \mathcal{Z} .
 - 12: **end for**
 - 13: For a threshold u , define the variables $Z_j(u) = Z_j \cdot \mathbf{1}\{U_{i_j} \leq u\}$ for $j = 1, \dots, m$.
 - 14: $\hat{\mu}_Z(u) \leftarrow \frac{1}{m} \sum_{j=1}^m Z_j(u)$
 - 15: $R \leftarrow \frac{B}{\min_i \pi_i}$
 - 16: $\delta_{\text{HB}}(\alpha) \leftarrow \sqrt{\frac{R^2 \log(2/\alpha)}{2m}}$
 - 17: **Return** $\hat{L}_u(\alpha) \leftarrow \hat{\mu}_Z(u) + \delta_{\text{HB}}(\alpha)$.
-

G TRANSDUCTIVE PAC REASONING

In this section, we introduce a transductive version of PAC reasoning. In this setting, the calibration set and the test set are identical. We consider a fixed dataset $\mathcal{D} = \mathcal{D}_{\text{test}} = \mathcal{D}_{\text{cal}} = \{x_1, \dots, x_n\}$, and the randomness only comes from the algorithm itself (e.g., which data points are selected to query the expert, the sampling design, and the internal randomization of the mean upper bound estimator). The goal is to provide a guarantee of empirical performance over this fixed dataset. Specifically, the algorithm ensures that the empirical average performance loss is controlled below a user-specified tolerance level ϵ with a confidence of at least $1 - \alpha$.

Let's update the setup for the transductive setting. For a given threshold u , the empirical risk on the dataset \mathcal{D} is defined as

$$L(u) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{y}_i) \mathbf{1}\{U_i \leq u\}.$$

This represents the average loss for the data points where the model is used (i.e., uncertainty is below the threshold u), and L_u is non-decreasing in u . The validity of our transductive PAC reasoning algorithm relies on the following assumptions. For any given threshold u and significance level α , there exists an upper confidence bound (UCB) $\hat{L}_u(\alpha)$, computable from samples drawn by the algorithm, that satisfies

$$\mathbb{P}(L(u) \leq \hat{L}_u(\alpha)) \geq 1 - \alpha.$$

In practice, we instantiate $\hat{L}_u(\cdot)$ using our UCB procedures. Concretely, one may compute $\hat{L}_u(\alpha')$ via Algorithm 1 (CLT-based) or Algorithm 3 (Hoeffding/Bentkus/betting-based), depending on sample size and desired conservativeness. We summarize our transductive style method in Algorithm 4.

Algorithm 4 Transductive PAC-Labeling

```

1: Inputs: Test ataset  $\mathcal{D} = \{x_1, \dots, x_n\}$ , uncertainty scores  $U_i, i = 1, \dots, n$ , model predictions  $\tilde{y}_i, \tilde{y}_n$ , tolerance  $\epsilon$ , significance level  $\alpha$ , number of trials  $m$ , sampling probabilities  $\pi_i, i = 1, \dots, n$ , UCB function UCB( $m, \alpha$ ).
2: Output: Labeled dataset  $\{(X_i, \tilde{Y}_i)\}_{i=1}^n$  and threshold  $\hat{u}$ .
3: Sampling phase: Draw up to  $m$  observations to estimate  $L_u$  via importance sampling.
4: for  $j = 1, \dots, m$  do
5:   Draw an index  $i_j$  according to the chosen sampling design (e.g., uniform or importance-based).
6:   With probability  $\pi_{i_j}$ , query the expert for  $y_{i_j}$ . Let  $\xi_j \sim \text{Bernoulli}(\pi_{i_j})$ .
7:   if  $\xi_j = 1$  then
8:     Observe the true label  $y_{i_j}$  and compute the loss  $\ell(y_{i_j}, \tilde{y}_{i_j})$ .
9:   else
10:    Mark as not-observed.
11:   end if
12: end for
13: For each candidate threshold, compute a weighted sample for UCB.  $\hat{L}_u(\alpha)$ 
14: Choose the estimated threshold.
15: Let  $\hat{u} := \max\{\hat{L}_u \leq \epsilon\}$ .
16: Final label assignment on the fixed dataset  $\mathcal{D}$ .
17: for  $i = 1, \dots, n$  do
18:   if  $U_i \geq \hat{u}$  then
19:     Ensure expert label  $y_i$  is obtained (query now if not already queried).
20:     Set  $\tilde{y}_i := y_i$ .
21:   else
22:     Set  $\tilde{y}_i := \hat{y}_i$ .
23:   end if
24: end for

```

Theorem 9 (Transductive PAC guarantee). *Suppose $\mathbb{P}(L(u) \leq \hat{L}_u(\alpha)) \geq 1 - \alpha$. Then the procedure in Algorithm 4, which selects $\hat{u} = \max\{u : \hat{L}_u(\alpha) \leq \epsilon\}$, achieves*

$$\mathbb{P}(L(\hat{u}) \leq \epsilon) \geq 1 - \alpha.$$

Proof. Let E be the event that $L_u \leq \hat{L}_u(\alpha)$ holds simultaneously for all $u \in \Lambda$. Then $\mathbb{P}(E) \geq 1 - \alpha$. On E , for any u with $\hat{L}_u(\alpha) \leq \epsilon$ we have $L(u) \leq \epsilon$. By the selection rule, either the set is non-empty and \hat{u} is the maximal such u , which implies $L_{\hat{u}} \leq \epsilon$ by monotonicity, and we take $\hat{u} = -\infty$, in which case $L_{-\infty} = 0 \leq \epsilon$ since no model predictions are used. Therefore $\mathbf{1}\{L(\hat{u}) \leq \epsilon\} = 1$ on E , and hence $\mathbb{P}(L(\hat{u}) \leq \epsilon) \geq \mathbb{P}(E) \geq 1 - \alpha$. \square

Remark 10 (Comparison with the inductive setting). In the inductive setting, the calibration and test sets are drawn independently from the same distribution. The goal is to guarantee performance

Table 1: Prompt for the verbalized confidence scores.

System prompt: You are a reasoning assistant. For each question and proposed answer, you must estimate how likely the proposed answer is correct.
User prompt:
Question: {QUESTION}
Answer: {ANSWER}
Provide a probability (between 0.0 and 1.0) that your answer is correct. Only output the probability.

on future, unseen data points from that distribution. The guarantee is of the form $\mathbb{P}(\mathbb{E}L(\hat{u}) \leq \varepsilon) \geq 1 - \alpha$, where the probability is over the random draws of both datasets. In contrast, our transductive approach provides a guarantee for a specific, fixed dataset, which can be more suitable in applications where the set of items to be labeled is known in advance or the calibration and test datasets are not exchangeable.

H EXPERIMENTAL DETAILS

Hyperparameter settings of LLMs In this study, we configure the decoding parameters as follows: for Qwen/Qwen3-4B-Instruct-2507, we set *Temperature* = 0.7, *TopP* = 0.8, *TopK* = 20, and *MinP* = 0; for Qwen/Qwen3-4B-Thinking-2507, we set *Temperature* = 0.6, *TopP* = 0.95, *TopK* = 20, and *MinP* = 0. Experiments were run on one NVIDIA RTX A6000 Graphics Card.

The prompt for verbalized uncertainty score In Table 1, we present the prompt used to elicit the verbalized confidence scores. After ten trials, we obtained the average confidence score and defined the verbalized uncertainty score as 1 minus this average confidence.

Details of Datasets Table 2 summarizes the datasets employed in our experiments, together with their corresponding splitting strategies. For each dataset, we report its type, overall size, and the partitioning into PAC calibration and PAC test sets.

Table 2: The details of datasets and splitting settings for PAC experiments

Dataset	Dataset Type	Dataset Size	Split Setting	Size
MATH-500	Math Reasoning	500	PAC Calibration	300
			PAC Test	200
ZebraLogic	Text reasoning	1000	PAC Calibration	500
			PAC Test	500
Arena-Hard	Alignment Task	750	PAC Calibration	450
			PAC Test	300

Loss function We consider two types of loss functions for evaluating the PAC guarantee of our reasoning method. The first is a semantic cosine distance, which measures the semantic similarity between outputs in the embedding space. Formally, given reference output $y_i = f(x_i)$ and PAC reasoning output $\hat{y}_i = \hat{f}(x_i)$, we compute their embeddings v_{y_i} and $v_{\hat{y}_i}$, and define the loss as:

$$\ell(y_i, \hat{y}_i) = 1 - \frac{v_y \cdot v_{\hat{y}}}{\|v_y\| \|v_{\hat{y}}\|}. \quad (8)$$

For the semantic embedding model, we adopt “Qwen/Qwen3-Embedding-4B” (Yang et al., 2025a). Second, we employ a binary 0–1 loss, which captures the actual loss in answer accuracy when comparing the PAC reasoning output \hat{y} with the reference output y :

$$\ell(y_i, \hat{y}_i) = \ell(y_i, \hat{y}_i | y_{i,gold}) = \mathbb{1}\{\hat{y}_i \neq y_i^{gold}\} \mathbb{1}\{y_i = y_i^{gold}\} \quad (9)$$

where $y_{i,gold}$ is the ground-truth answer for the problem x_i .

The choice of loss functions The loss functions, shown as in Eq. (8) and Eq. (9), serve distinct purposes in evaluating the PAC reasoning. The semantic cosine distance captures the degree of semantic alignment between the PAC reasoning’s prediction and reference outputs. It is particularly suitable for tasks where nuanced differences in meaning are critical, such as natural language understanding or generation tasks. By leveraging the “Qwen/Qwen3-Embedding4B” model, we ensure that the embeddings capture rich contextual information, robustly comparing semantic content in high-dimensional spaces. In contrast, the binary 0–1 loss is designed for scenarios where the correctness of the generated answer is verifiable, such as in mathematical problem-solving or multiple-choice question answering. This loss function is particularly effective for evaluating the framework’s ability to produce exact matches to ground-truth answers, emphasizing precision in verifiable tasks. By testing the PAC reasoning on these two loss functions, we can assess the semantic quality and factual accuracy of the PAC reasoning across diverse tasks.

I MAIN RESULTS FOR THE BINARY LOSS

For the binary loss, we evaluate PAC reasoning on the verifiable datasets MATH-500 and ZebraLogic, with target risk levels set to $\epsilon = 0.03$ and $\epsilon = 0.08$, respectively (see Section 4.1 for experimental details). For comparison, we also consider a naive fixed-threshold baseline as well as the approach that relies solely on the nonthinking model.

As shown in Table 3, PAC reasoning consistently keeps the error rates below the target risk, while also achieving substantial efficiency gains. In contrast, the naive baseline exhibits unstable behavior across datasets: on ZebraLogic, although it attains a very small error with logits-based uncertainty, it violates efficiency by yielding a negative STP (-34.78%), meaning it requires even more tokens than fully using the thinking model. Meanwhile, on MATH-500 with verbalized uncertainty, the same method produces a large error (0.0346), which substantially exceeds the target risk $\epsilon = 0.03$. These results highlight that naive thresholding fails to provide reliable control over both loss and budget, often swinging between overly conservative and overly risky outcomes. In summary, PAC reasoning strikes a balanced trade-off, keeping the error within ϵ while delivering consistent savings across tasks and datasets.

Table 3: Experimental results of the binary loss function on verifiable datasets ($\alpha = 0.05$). For MATH-500, we set $\epsilon = 0.03$, and for ZebraLogic, we set $\epsilon = 0.08$.

Dataset	Metric	Logits-based score		Verbalized score		Non-thinking
		PAC reasoning	Naive ($U_i \geq 0.05$)	PAC reasoning	Naive ($U_i \geq 0.05$)	
MATH-500	Error	0.0206 ± 0.0126	0.0179 ± 0.0068	0.0209 ± 0.0141	0.0346 ± 0.0095	0.0435 ± 0.0107
	ECP (%) ↓	21.48 ± 17.85	14.44 ± 2.02	24.59 ± 20.48	2.83 ± 0.94	—
	STP (%) ↑	37.61 ± 23.19	43.58 ± 4.78	36.13 ± 26.44	66.67 ± 4.91	—
ZebraLogic	Error	0.0615 ± 0.0181	0.0062 ± 0.0026	0.0530 ± 0.0246	0.0631 ± 0.0074	0.1163 ± 0.0102
	ECP (%) ↓	22.50 ± 7.47	77.28 ± 1.36	26.95 ± 20.68	12.49 ± 1.07	—
	STP (%) ↑	23.13 ± 9.90	-34.78 ± 1.26	21.21 ± 17.20	32.70 ± 2.11	—