Quantifying and Understanding Uncertainty in Large Reasoning Models

Yangyi Li*, **Chenxu Zhao***, **Mengdi Huai**
Department of Computer Science, Iowa State University
{liyangyi, cxzhao, mdhuai}@iastate.edu

Abstract

Large Reasoning Models (LRMs) have recently demonstrated significant improvements in complex reasoning. While quantifying generation uncertainty in LRMs is crucial, traditional methods are often insufficient because they do not provide finite-sample guarantees for reasoning-answer generation. Conformal prediction (CP) stands out as a distribution-free and model-agnostic methodology that constructs statistically rigorous uncertainty sets. However, existing CP methods ignore the logical connection between the reasoning trace and the final answer. Additionally, prior studies fail to interpret the origins of uncertainty coverage for LRMs as they typically overlook the specific training factors driving valid reasoning. Notably, it is challenging to disentangle reasoning quality from answer correctness when quantifying uncertainty, while simultaneously establishing theoretical guarantees for computationally efficient explanation methods. To address these challenges, we first propose a novel methodology that quantifies uncertainty in the reasoning-answer structure with statistical guarantees. Subsequently, we develop a unified example-to-step explanation framework using Shapley values that identifies a provably sufficient subset of training examples and their key reasoning steps to preserve the guarantees. We also provide theoretical analyses of our proposed methods. Extensive experiments on challenging reasoning datasets verify the effectiveness of the proposed methods.

1 Introduction

Recent advancements have further evolved toward Large Reasoning Models (LRMs), exemplified by the OpenAI o-series (OpenAI, 2024) and DeepSeek-R1 series (DeepSeek-AI et al., 2025), capable of tackling complex tasks (Shen et al., 2025). Unlike standard LLMs that directly map inputs to outputs, LRMs distinguish themselves by generating an explicit reasoning trace prior to the final answer. By scaling inference-time computations, these models acquire the capability to explore diverse solution paths, reflect on potential errors, and refine intermediate steps. This rigorous process allows LRMs to decompose intricate problems in challenging domains such as mathematics and coding, justifying their conclusions with logical derivations that reflect human problem-solving.

Ensuring the reliable deployment of such models in real-world applications requires accurate uncertainty quantification. However, conventional approaches typically lack rigorous statistical guarantees (Tian et al., 2023), limiting their trustworthiness in critical scenarios. While Conformal Prediction (CP) (Vovk et al., 2005; Shafer and Vovk, 2008) offers a distribution-free and model-agnostic framework to address this, traditional CP methods are insufficient for LRMs. Existing approaches either calibrate the final answer while ignoring the reasoning trace (Ye et al., 2024), or treat the generation as a whole without verifying whether the reasoning logically supports the final answer (Quach et al., 2024). Consequently, by failing to account for the intrinsic reasoning-answer structure of LRMs, these methods cannot guarantee that a correct answer is supported by a valid reasoning trace. This limitation necessitates the design of a valid uncertainty quantification framework explicitly tailored to this special structure.

Beyond establishing valid uncertainty sets, interpreting the origins of quantified uncertainty is equally critical. For LRMs, identifying which training examples and steps are sufficient to guarantee uncertainty coverage provides essential insights into the reliability of the reasoning trace and guides further model refinement. However, existing literature largely overlooks the connection between uncertainty quantification and its explanation. Standard explanation methods (Selvaraju et al., 2017)

*Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32741–32754

typically focus on label prediction and lack the mechanisms to trace the validity of uncertainty sets back to the training stage. Moreover, they often treat training samples as single units, failing to pinpoint which reasoning steps within an example are critical for ensuring that the model generates valid reasoning-answer pairs. Consequently, even when uncertainty is rigorously guaranteed, the training factors behind this reliability remain opaque.

However, addressing these limitations presents several unique challenges. First, formulating a rigorous uncertainty guarantee for valid reasoning-answer pairs is inherently difficult. Specifically, it requires disentangling the quality of the reasoning trace from the correctness of the final answer while simultaneously capturing their interdependence within a rigorous statistical framework. Second, attributing valid prediction uncertainty to specific training data faces significant computational hurdles. This complexity escalates dramatically when extending attribution to the fine-grained level of reasoning steps, as rigorous attribution typically necessitates evaluating model performance across an exponentially growing number of data subsets. While approximation methods are required to mitigate these costs, establishing rigorous theoretical guarantees for such approximations presents a critical challenge. Specifically, it is challenging to certify that the extracted data subset identified through approximation is sufficient to preserve the statistical risk control of the original model.

To address the aforementioned challenges, in this paper, we first propose **Co**nformal **R**easoning-**A**nswer **P**rediction (**CoRAP**), a novel framework designed to jointly quantify uncertainty of the reasoning-answer structure. By defining distinct quality functions to evaluate logical interdependence of reasoning-answer pairs and employing a rigorous statistical calibration procedure, CoRAP constructs uncertainty sets with finite-sample guarantees, ensuring that the set contains a generated sequence in which the true answer is supported by a valid reasoning trace with a user-specified probability. In addition, we propose a unified example-to-step explanation framework using Shapley values to identify the most influential training factors contributing to the uncertainty coverage. To resolve computational intractability, we employ a hierarchical Monte Carlo approximation that first identifies pivotal training examples and subsequently isolates critical reasoning steps. Crucially, our framework extracts an example or step subset that is provably sufficient to ensure the uncertainty guarantees. Furthermore, we provide detailed theoretical analysis to certify the validity of our proposed uncertainty quantification and explanation methods. Extensive experiments across complex reasoning tasks demonstrate the effectiveness of our approach in ensuring theoretically guaranteed and interpretable uncertainty quantification.

2 Related Work

Large Reasoning Models (LRMs) represent a significant evolution by generating explicit reasoning traces to decompose intricate problems (An et al., 2025). Leveraging this capability, they play a pivotal role in various applications, including mathematical reasoning (Zhang, 2025; Minegishi et al., 2025), web search (Li et al., 2025b), and code generation (Li et al., 2025a). While uncertainty quantification is crucial for ensuring model trustworthiness (Qian et al., 2025), existing work developed for LLMs remains inadequate for LRMs (Han et al., 2025). Heuristic approaches (Zhang and Zhang, 2025; Li and Huai, 2025) relying on token-level probabilities or verbalized confidence scores, fundamentally lack rigorous statistical guarantees. Conversely, frameworks employing CP offer finite-sample coverage guarantees (Li et al., 2024; Wang et al., 2025) but fail to account for the intrinsic reasoning-answer structure of LRMs. Specifically, existing approaches either calibrate the final answer while ignoring the intermediate reasoning trace (Ye et al., 2024), or treat the generation process as a whole without verifying whether the reasoning logically supports the final answer (Quach et al., 2024; Su et al., 2025). Consequently, these methods often ensure coverage for the output but disregard the logical validity of the reasoning.

Moreover, explaining the sources of uncertainty is equally critical for transparency. Traditional approaches (Watson et al., 2023; Slack et al., 2021; Qian et al., 2024; Chen et al., 2024; Li and Huai, 2026) attribute uncertainty to static input features but are intractable for LRMs due to the prohibitive cost of long-sequence sampling and the complexity of dynamic reasoning. While natural language explanations are widely used to interpret LLM predictions (Kumar and Talukdar, 2020), they often fail to account for uncertainty. To address this, (Liu et al., 2025) attempts to categorize uncertainty sources to generate inquiries. However, this heuristic approach lacks both the step-level granularity essential for complex reasoning and rigorous sta-

tistical guarantees for explanation validity. In this work, we address these limitations by proposing a framework that jointly quantifies the uncertainty of the reasoning-answer structure with finite-sample guarantees. To identify influential factors, we introduce a unified example-to-step explanation method based on Shapley values. We resolve computational intractability via a hierarchical Monte Carlo approximation that efficiently isolates pivotal training examples and reasoning steps. Crucially, our framework extracts a subset provably sufficient to maintain the established uncertainty guarantees.

3 Methodology

In this section, we introduce our proposed framework designed to quantify and explain uncertainty in LRMs. While this task is critical, existing methods fail to disentangle reasoning traces from answers to guarantee the validity of reasoning-answer pair uncertainty, and they lack computationally feasible mechanisms to attribute uncertainty to specific data sources with guarantees. To address these challenges, we first propose a novel method that constructs statistically valid uncertainty sets for reasoning-answer pairs. Subsequently, we describe our example-to-step explanation framework, which utilizes Shapley values to identify the specific training examples and reasoning steps responsible for ensuring valid uncertainty quantification.

Formally, we consider a supervised reasoning task using a training set $\mathcal{D}_{\text{tr}} = \{z_j = (x_j, q_j, r_j, y_j)\}_{j=1}^{n_{\text{tr}}}$, where each instance consists of an input image x_j, a query q_j, a reasoning trace r_j including steps, and a final answer y_j. We denote a complete answer as $a = r \| y$. An LRM parameterizes a generation policy π_θ that models the joint probability of reasoning and answer tokens conditioned on the input. We use $\hat{a}$ to denote an answer generated by the model.

3.1 Conformal Reasoning-Answer Prediction

Here, we propose a post-hoc and model-agnostic method that assigns provably valid uncertainty to the reasoning-answer structure given a calibration set $\mathcal{D}_{\text{cal}} = \{z_i = (x_i, q_i, r_i, y_i)\}_{i=1}^{n_{\text{cal}}}$. The challenge is to disentangle the reasoning trace r from the final answer y, and to guarantee that a correct answer is supported by a valid reasoning process.

To address this challenge, we first define distinct quality functions designed to disentangle the evaluation of reasoning and answers. Specifically, the plausibility of a full response is measured by the sequence quality function $Q(x_i, q_i, \hat{a}_i) = \frac{1}{|\hat{a}_i|} \log p_\theta(\hat{a}_i | x_i, q_i)$. We define the set confidence function as $F(\mathcal{C}) = \max_{a \in \mathcal{C}} Q(x_i, q_i, a)$. Furthermore, to validate the interdependence between the reasoning trace and the final answer, a conditional answer quality function $A(x_i, q_i, \hat{a}_i) = p_\theta(\hat{y}_i | x_i, q_i, \hat{r}_i)$ is employed to evaluate the model's confidence in $\hat{y}_i$ conditioned on $\hat{r}_i$. Intuitively, the sequence quality function Q filters out low-quality individual sequences, the set confidence function F ensures at least one candidate in the set is high-confidence, and the answer quality function A acts as a consistency check to ensure the answer is grounded in the reasoning trace.

We employ a sampling-based procedure to construct the prediction set by iteratively expanding a candidate pool. At each step k, we sample a sequence $\hat{a}_k$ from the model and include it in the output set $\mathcal{C}$ only if its sequence quality Q exceeds the threshold λ_1. This accumulation continues until the step $K \leq K_{\max}$, at which point two termination criteria are simultaneously met: (i) the set confidence $F(\mathcal{C}_K) \geq \lambda_2$, and (ii) the answer quality $A(x_i, q_i, a^*) \geq \lambda_3$ for the highest-scoring candidate a^* in the set $\mathcal{C}_K$. Considering a finite grid Λ of threshold tuples $\lambda = (\lambda_1, \lambda_2, \lambda_3)$, the final output set is denoted as $\mathcal{C}_\lambda^{\text{RA}}(x_i, q_i; \theta)$.

To assess the validity of the reasoning-answer pair uncertainty with finite-sample guarantees, we adopt a specialized loss function within a Learn Then Test (Angelopoulos et al., 2025) framework to identify the specific threshold configuration λ that ensures uncertainty guarantees on unseen data. The core idea is to define a failure event for any given instance. We now calibrate which threshold configuration to use so that the probability of such failures is controlled on test data. Specifically, the loss function is formulated to explicitly capture the failure of generating a correct answer supported by valid reasoning as follows:

$$L^{\text{RA}}(z_i; \lambda, \theta) = \mathbf{1}\{\nexists\, \hat{a}_k \in \mathcal{C}_\lambda^{\text{RA}}(x_i, q_i; \theta) : V(z_i, \hat{r}_k) = 1 \text{ and } \hat{y}_k = y_i\}, \quad (1)$$

where $V(z_i, \hat{r}_k)$ is a binary admission function that returns 1 if the generation reasoning trace is admitted for the input x_i, and 0 otherwise. Given a model θ, we evaluate every candidate $\lambda \in \Lambda$ on the calibration set by computing its empirical risk. To ensure statistical validity, we convert the estimated risk into a statistical p-value and apply family-wise error rate (FWER) control at level ε. The valid set $\Lambda_{\text{valid}}(\theta)$ retains only configurations

with sufficiently small FWER-adjusted p-values. With probability at least $1-\varepsilon$, the true risk of every retained configuration is bounded by α. If this set is empty, the procedure abstains. Otherwise, we select a final $\hat{\lambda}(\theta)$ from $\Lambda_{\text{valid}}(\theta)$ that minimizes the expected set size.

Theorem 3.1. *Assume calibration examples in $\mathcal{D}_{\text{cal}}$ and a test sample z_t are i.i.d., and CoRAP's sampling randomness is independent across examples. Let $\Lambda_{\text{valid}}(\theta)$ be obtained by applying an FWER-ε procedure to the p-values for target level $\alpha \in (0,1)$ on model θ. Then, for any $\hat{\lambda}(\theta) \in \Lambda_{\text{valid}}(\theta)$, we have the risk constraint*

$$\mathbb{E}[L^{\text{RA}}(z_t; \hat{\lambda}(\theta), \theta)] \leq \alpha, \tag{2}$$

with probability at least $1-\varepsilon$ over $\mathcal{D}_{\text{cal}}$.

In other words, this theorem guarantees that with probability at least $1-\varepsilon$ over the calibration data, the expected loss on z_t is strictly controlled below the user-specified target α. Since the loss indicates the failure to retrieve a valid reasoning-answer pair, this risk bound implies that the constructed uncertainty set covers a correct answer supported by valid reasoning with a probability of at least $1-\alpha$. We provide the detailed proof in Appendix B.

3.2 Example-to-Step Explanations for Conformal Reasoning-Answer Prediction

Based on the sets returned by CoRAP, we aim to identify the most influential training examples and reasoning steps that are crucial to satisfy the uncertainty guarantees in the theorem above for a given test instance. The primary challenge lies in the computational intractability of exact attribution methods, which typically require exponential retraining. Moreover, employing approximation methods introduces the critical obstacle of ensuring that the identified subset is statistically sufficient to ensure the rigorous risk control.

To overcome the challenges above, we develop a unified data valuation framework based on Shapley values (Shapley, 1953) with guarantees. At the example level, we quantify how each training example in $\mathcal{D}_{\text{tr}}$ contributes to covering y_t and select an appropriate subset $S^\star_{\text{ex}}$ that is provably sufficient for coverage at x_t. At the step level, conditioned on the selected examples $S^\star_{\text{ex}}$, we further attribute contribution to individual reasoning steps inside these examples and extract a step subset $T^\star$ that remains sufficient to achieve coverage. The two levels together produce an example-to-step explanation of CoRAP success with valid guarantees.

Formalizing the explanation problem. We define the contributing players at two levels. At the example level, let $\mathcal{P}_{\text{ex}} = [n_{\text{tr}}]$, where each player j corresponds to a training example. At the step level, we define the player set $\mathcal{P}_{\text{st}}$ as the set of all steps from the selected examples $S^\star_{\text{ex}}$. Next, we define value functions that measure the success of any coalition. Given a coalition of examples $S \subseteq \mathcal{P}_{\text{ex}}$, we train a model θ^S exclusively on the examples within S. Subsequently, we execute the CoRAP procedure at level α using this model to generate the uncertainty set $\mathcal{C}^{\text{RA}}_{\hat{\lambda}}(x_t, q_t; \theta^S)$. The example-level value is defined to measure whether a subset S suffices to ensure valid coverage:

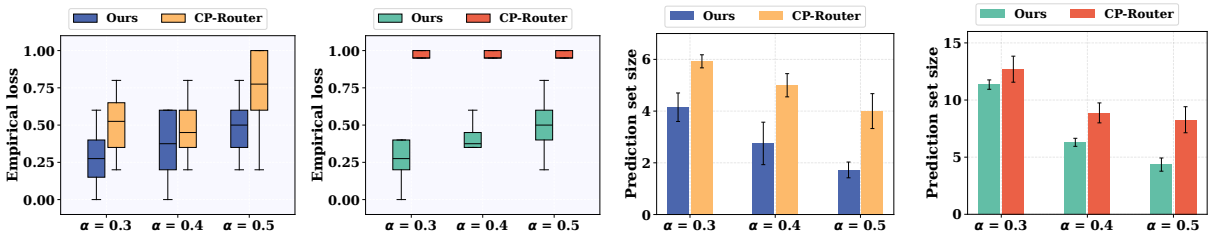
$$v_{\text{ex}}(S; z_t) := 1 - L^{\text{RA}}(z_t; \hat{\lambda}(\theta^S), \theta^S). \tag{3}$$

Conditioned on $S^\star_{\text{ex}}$, we consider per-example step selections. For the step universe $\mathcal{P}_{\text{st}}$ derived from $S^\star_{\text{ex}}$, consider a step coalition $T \subseteq \mathcal{P}_{\text{st}}$. The model θ^T is trained on the restricted steps in T, while all other steps are discarded. Running CoRAP under this training regime yields the set $\mathcal{C}^{\text{RA}}_{\hat{\lambda}}(x_t, q_t; \theta^T)$. From this, we define the conditional value to quantify the step-level contribution towards satisfying the guarantee:

$$v_{\text{st}}(T; z_t) := 1 - L^{\text{RA}}(z_t; \hat{\lambda}(\theta^T), \theta^T). \tag{4}$$

We specifically aim to identify the pivotal subset of training examples or steps that is sufficient to ensure the validity of the test sample by value functions. Overall, this hierarchical formulation allows us to first identify the influential examples and subsequently isolate the critical reasoning steps within them, effectively avoiding the computational overhead of searching the entire step universe.

Quantifying contribution and constructing provable explanations. To identify a subset sufficient to satisfy the coverage guarantee, we quantify the contribution of each player (an example or a step) using Shapley values. Conceptually, this game-theoretic metric measures a player's importance as its average marginal contribution to the coalition's value across all permutations. Since computing the exact Shapley value requires summing over exponentially many coalitions, it is computationally intractable. To address the computational challenge, we employ a Monte Carlo (MC) approximation. We sample M independent and uniformly random permutations $\tau_1, \ldots, \tau_M$ of $\mathcal{P}$. For each τ_m, let $S^{(m)}_u$ denote the players appearing before u in the sequence. The unbiased estimator for the Shapley value is defined as:



(a) Validity on CLEVR-Math (b) Validity on ScienceQA (c) Efficiency on CLEVR-Math (d) Efficiency on ScienceQA

Figure 1: Empirical loss and efficiency results across datasets.

$$\widehat{\phi}_u = \frac{1}{M}\sum_{m=1}^{M}\left(v(S_u^{(m)} \cup \{u\}) - v(S_u^{(m)})\right). \quad (5)$$

To quantify statistical reliability, we leverage the fact that the marginal contribution $v(S \cup \{u\}) - v(S)$ is bounded in $[-1, 1]$. Applying Hoeffding's inequality and a union bound over a player universe $\mathcal{P}$, with probability at least $1-\delta$ we have simultaneously for all $u \in \mathcal{P}$ that $\phi_u \geq \widehat{\phi}_u - b(M, \delta, |\mathcal{P}|)$, where $b(M, \delta, |\mathcal{P}|) = 2\sqrt{\log(2|\mathcal{P}|/\delta)/(2M)}$ represents the estimation uncertainty radius. We define $\mathcal{B}_u := \widehat{\phi}_u - b(M, \delta, |\mathcal{P}|)$ as the lower confidence bound. Intuitively, $\mathcal{B}_u$ serves as a conservative estimate of a player's true importance, penalizing the empirical mean $\widehat{\phi}_u$ by the sampling noise to prevent overestimation. Then we can construct our provable explanation by finding a subset of players whose cumulative certified contributions are sufficient to ensure the risk control.

At the example level, we compute $\widehat{\phi}_u^{\text{ex}}$ on $\mathcal{P}_{\text{ex}}$ using m_{ex} permutations and a failure budget δ_{ex}. We order the players by non-increasing lower confidence bounds $\mathcal{B}_u^{\text{ex}}$ to prioritize the most reliably influential examples. To identify a sufficient subset, we select the smallest number of top-ranked examples k such that their cumulative certified contribution satisfies $\sum_{i=1}^{k} \mathcal{B}_{u_i}^{\text{ex}} \geq 1-\alpha$, and report this set as $S_{\text{ex}}^\star$. Conditioned on the selected examples $S_{\text{ex}}^\star$, we refine the explanation to the step level using the value function v_{st}. We compute $\widehat{\phi}_u^{\text{st}}$ on the step universe $\mathcal{P}_{\text{st}}$ with m_{st} permutations and budget δ_{st}. Similarly, we sort the steps by non-increasing $\mathcal{B}_u^{\text{st}}$ and select the smallest k steps necessary to meet the condition $\sum_{i=1}^{k} \mathcal{B}_{u_i}^{\text{st}} \geq 1-\alpha$. The resulting set of reasoning steps is reported as $T^\star$.

Theorem 3.2. *Let z_t be a test example that is i.i.d. with the calibration examples in $\mathcal{D}_{\text{cal}}$, and let $\mathcal{H} := \{ L^{\text{RA}}(z_t; \hat{\lambda}(\theta_0), \theta_0) = 1 \}$ denote the event that the base model θ_0 fails on z_t. Condition on the validity of the thresholds used in evaluating v_{ex} and v_{st}. Assume that for a target level $\alpha \in (0,1)$ and error bounds $\xi_{\text{ex}}, \xi_{\text{st}} \geq 0$, the condition $v(\mathcal{P} \setminus U) \geq v(\mathcal{P}) - \sum_{u \in U} \phi_u - \xi$ holds for $(v, \mathcal{P}, \xi) = (v_{\text{ex}}, \mathcal{P}_{\text{ex}}, \xi_{\text{ex}})$ and $(v, \mathcal{P}, \xi) = (v_{\text{st}}, \mathcal{P}_{\text{st}}, \xi_{\text{st}})$.*
(i) (Example-level). *With probability at least $1 - \delta_{\text{ex}}$ over the Shapley MC, the example subset $S_{\text{ex}}^\star$ identified by the stopping condition $\sum_{u \in S_{\text{ex}}^\star} \mathcal{B}_u^{\text{ex}} \geq 1-\alpha$ is sufficient to achieve the risk control:*

$$\mathbb{E}[L^{\text{RA}}(z_t; \hat{\lambda}(\theta^{S_{\text{ex}}^\star}), \theta^{S_{\text{ex}}^\star}) \mid \mathcal{H}] \;\leq\; \alpha + \xi_{\text{ex}}. \quad (6)$$

(ii) (Step-level). *Conditionally on $S_{\text{ex}}^\star$, with probability at least $1-\delta_{\text{st}}$ over the step-level Shapley MC, the step subset $T^\star$ identified by the stopping condition $\sum_{u \in T^\star} \mathcal{B}_u^{\text{st}} \geq 1-\alpha$ is sufficient to achieve the risk control:*

$$\mathbb{E}[L^{\text{RA}}(z_t; \hat{\lambda}(\theta^{T^\star}), \theta^{T^\star}) \mid \mathcal{H}] \;\leq\; \alpha + \xi_{\text{st}}. \quad (7)$$

Theorem 3.2 provides a two-level guarantee on our unified explanation framework. Part (i) certifies the example-level explanation by stating that if the cumulative certified contribution of the selected examples $S_{\text{ex}}^\star$ meets the threshold $1-\alpha$, the subset is sufficient to ensure the risk control in Eq. (6) with probability $1-\delta_{\text{ex}}$. This implies that a model trained solely on $S_{\text{ex}}^\star$ satisfies the valid coverage requirements established in Theorem 3.1. Part (ii) extends this certification to the step-level explanation given $S_{\text{ex}}^\star$. Specifically, provided that the cumulative certified contribution of the selected steps $T^\star$ exceeds the threshold $1-\alpha$, the resulting model trained exclusively on the reasoning steps in $T^\star$ within $S_{\text{ex}}^\star$ is guaranteed to achieve the risk control in Eq. (7) with probability $1-\delta_{\text{st}}$. We also provide the detailed proof in Appendix B.

Discussion. We analyze the time complexity for a single test instance and propose several strategies to further reduce the computational overhead. At the example level, the total baseline complexity is $O(m_{\text{ex}}(n_{\text{tr}}^2 + n_{\text{tr}} K_{\max}) + m_{\text{st}}(|\mathcal{P}_{\text{st}}|^2 + |\mathcal{P}_{\text{st}}| K_{\max}) + |\Lambda|\, n_{\text{cal}} K_{\max})$. By leveraging influence functions and machine unlearning techniques (Zhao et al., 2024; Chen et al., 2025), we can substitute computationally expensive retraining with efficient approximate updates. With a

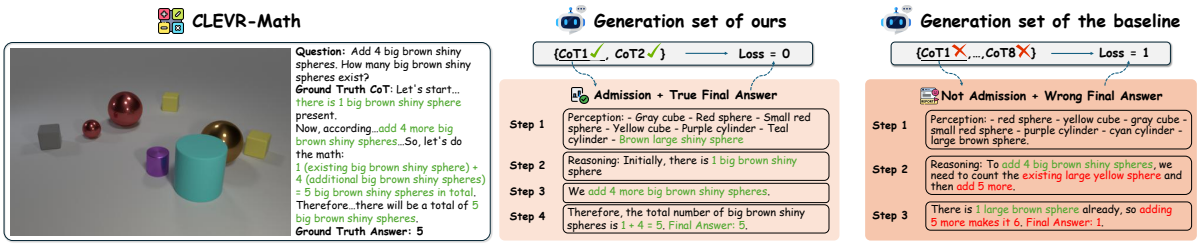


Figure 2: Visualization of our proposed conformal prediction framework on CLEVR-Math.

one-time setup cost I (e.g., gradient caching), the complexity reduces to $O(I+m_{\text{ex}}(n_{\text{tr}}+n_{\text{tr}}K_{\max})+m_{\text{st}}(|\mathcal{P}_{\text{st}}|+|\mathcal{P}_{\text{st}}|K_{\max})+|\Lambda|\,n_{\text{cal}}K_{\max})$. To further optimize, we partition the n_{tr} players into G disjoint groups. We employ a warm-start preselection with m_0 permutations to identify one active group by the value function $v=1$, restricting the intensive estimation to this sole group. The final optimized complexity is $O(I+m_0(G+GK_{\max})+m_{\text{ex}}K_{\max}+m_{\text{st}}(|\mathcal{P}_{\text{st}}|+|\mathcal{P}_{\text{st}}|K_{\max})+|\Lambda|\,n_{\text{cal}}K_{\max})$, which yields significant savings when the average group size $\bar{g}=n_{\text{tr}}/G \ll n_{\text{tr}}$. This optimization effectively decouples the inference-time cost from the quadratic scaling of the training data, enabling the practical application of our method to large-scale reasoning models.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed methods. More experimental details and results (e.g., experimental results about uncertainty on various models) are deferred to the Appendix of the paper.

Datasets and models. In experiments, we evaluate our method on two challenging multimodal reasoning datasets. CLEVR-Math (Lindström and Abraham, 2022) assesses mathematical reasoning in visual contexts, while ScienceQA (Lu et al., 2022) provides a diverse set of science-domain questions based on images. We follow the preprocessing pipeline from (Tan et al., 2025). We consider various LRMs, including LLaVA-CoT (Xu et al., 2025), LMM-R1 (Peng et al., 2025), and R1-Onevision (Yang et al., 2025) models.

Baselines. For conformal prediction, we compare against *CP-Router* (Su et al., 2025), a recent uncertainty method tailored to LRMs. For explanation and data attribution methods in LRMs, to the best of our knowledge there is no dedicated prior method. We therefore include a *Random* baseline that uniformly samples a subset of examples (or steps), treats them as important, and fine-tunes under identical training and inference budgets.

Implementation details. In experiments, we use a temperature of 1.2 and top-p sampling of 0.85 to produce candidate sequences, where the sampling budget $K_{max}=16$ per input. For each dataset, we sample 1500 calibration examples from the original training corpus and 100 test examples from the test set, using the remaining training examples as the training data. All experiments are run for 8 trials, and we report the averaged results.

4.1 Conformal Prediction for LRMs

Validity. We evaluate the validity of our proposed uncertainty quantification framework on large reasoning models by constructing valid uncertainty sets for reasoning-answer pairs. Experiments are conducted on CLEVR-Math and ScienceQA using LMM-R1, under target significance levels $\alpha=0.3, 0.4$, and 0.5. As illustrated in Fig. 1a and 1b, the empirical losses of our methods remain below the target level α, demonstrating that our method satisfies Eq. (2) across various α and thus upholds the validity guarantee about the reasoning-answer structure established in Theorem 3.1. In contrast, CP-Router shows empirical losses exceeding the target levels in multiple settings, indicating that it fails to maintain the desired risk control.

Efficiency. In Fig. 1c and 1d, we examine the efficiency of our uncertainty quantification framework by prediction set size across different LRMs. Experiments are conducted on CLEVR-Math and ScienceQA using LMM-R1 under significance levels $\alpha=0.3, 0.4$, and 0.5, maintaining consistency with the validity evaluation. As shown, our proposed method consistently produces compact uncertainty sets over reasoning steps compared with CP-Router, enabling efficient interpretation of the model's thought process. For instance, on CLEVR-Math, our method achieves an average set size of only 2.76 with LMM-R1 at $\alpha=0.4$, meaning that fewer than three reasoning traces are typically sufficient to cover the correct answer while retaining valid coverage guarantees. In contrast, CP-Router achieves an average set size of 5.00 in the same

Model	Method	Top-1 Shapley value	Success rate
LMM-R1	Random	0.042 ± 0.026	0.000 ± 0.000
	Ours	$\mathbf{0.433 \pm 0.116}$	$\mathbf{0.833 \pm 0.167}$
R1-Onevision	Random	0.021 ± 0.038	0.000 ± 0.000
	Ours	$\mathbf{0.356 \pm 0.058}$	$\mathbf{1.000 \pm 0.000}$

Table 1: Data-level explanations of the uncertainty sets.

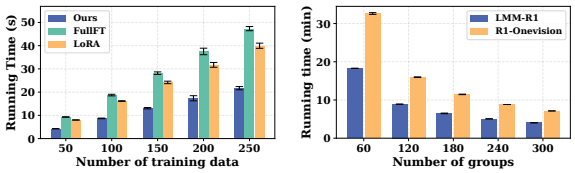
Model	Method	Top-1 Shapley value	Success rate
LMM-R1	Random	0.020 ± 0.016	0.000 ± 0.000
	Ours	$\mathbf{0.739 \pm 0.128}$	$\mathbf{0.875 \pm 0.125}$
R1-Onevision	Random	0.006 ± 0.005	0.000 ± 0.000
	Ours	$\mathbf{0.589 \pm 0.119}$	$\mathbf{0.750 \pm 0.164}$

Table 2: Step-level explanations of the uncertainty sets.

setting. Fig. 2 visually demonstrates this efficiency on a CLEVR-Math example. Our method generates a compact set containing only two reasoning traces. As shown in the green-coded section, these traces pass the admission function and achieve a true final answer, resulting in zero loss. In contrast, CP-Router generates a large set with eight traces. Its example trace, containing severe reasoning flaws highlighted in red, is rejected by the admission function and produces a wrong final answer, leading to a loss of 1. These combined results show that our method efficiently quantifies uncertainty by producing compact and high-quality sets of reasoning traces, making it a practical tool for analyzing complex visual reasoning tasks.

4.2 Explanations of Prediction Sets in LRMs

Data-level explanations. As shown in Table 1, our proposed example-level explanations substantially outperform a random retrieval baseline on CLEVR-Math using LMM-R1 and R1-Onevision when $\alpha = 0.5$, with the value of top-1 increasing from 0.042 to 0.433 using LMM-R1 and from 0.021 to 0.356 using R1-Onevision. Using the selected data to fine-tune changes the outcome of the model, so the uncertainty set for the test data shifts from excluding the true final answer to including it on both models, while fine-tuning on randomly chosen examples does not achieve this. This demonstrates that our method accurately identifies pivotal training evidence via Shapley value, providing a suitable explanation for why the uncertainty set ultimately includes the true final answer. Furthermore, to empirically validate the theoretical guarantee in Theorem 3.2 (i), we set the example-level failure parameter $\delta_{\text{ex}} = 0.25$. The success rate in Table 1, which measures the empirical frequency of successfully identifying a sample subset sufficient to maintain the CoRAP risk control, is 0.833 for LMM-R1 and 1.000 for R1-Onevision. Both results are greater than the required level $1 - \delta_{\text{ex}}$, empirically verifying the effectiveness of our proposed example-level explanation guarantee.



(a) Fine-tuning methods (b) Group numbers

Figure 3: Analysis of running time.

Step-level explanations. As shown in Table 2, our proposed step-level explanations substantially outperform a random step-selection baseline on CLEVR-Math using LMM-R1 and R1-Onevision when $\alpha = 0.5$, with the top-1 Shapley value of steps increasing from 0.020 to 0.739 using LMM-R1 and from 0.006 to 0.589 using R1-Onevision. Similar to the data-level analysis, we further empirically validate the theoretical guarantee for step-level attribution in Theorem 3.2 (ii). Setting the step-level failure probability $\delta_{\text{st}} = 0.25$, the success rate in Table 2 is 0.875 for LMM-R1 and 0.750 for R1-Onevision. Both results surpass the required threshold $1 - \delta_{\text{st}} = 0.75$, confirming the validity of our step-level explanation guarantee. These results present the influence of intermediate reasoning traces. We find that the uncertainty set's shift from excluding to including the true final answer is driven by critical reasoning steps within the model's generation process. Our method accurately identifies these pivotal internal reasoning steps by Shapley value, thereby providing suitable explanations for why the uncertainty set of test data ultimately includes the true final answer at the step-level for LRMs.

4.3 Ablation Study

First, we compare the running time of our proposed method against full fine-tuning (FullFT) and LoRA (Hu et al., 2022) on CLEVR-Math, with their methods run for 3 epochs. The results presented in Fig. 3a reveal that our method achieves a significantly lower computational cost when adopting the influence function methods. When the number of training data increases, the running time required by ours is substantially less than that of both FullFT and LoRA, demonstrating its superior efficiency over these traditional techniques.

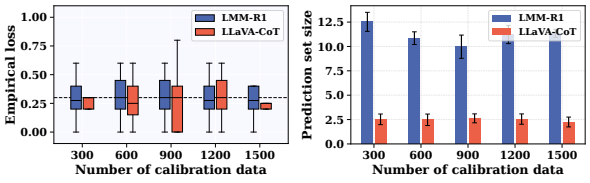
Then, we show the influence of group numbers for the running time of our proposed method using LMM-R1 and R1-Onevision on CLEVR-Math.

Dataset	Level ε	Probability	Prediction set size
CLEVR-Math	0.2	0.875 ± 0.125	2.775 ± 0.821
	0.4	0.625 ± 0.183	2.750 ± 0.429
	0.6	0.500 ± 0.189	2.575 ± 0.301
	0.8	0.500 ± 0.189	2.525 ± 0.285
ScienceQA	0.2	0.875 ± 0.125	6.525 ± 0.360
	0.4	0.625 ± 0.183	6.375 ± 0.799
	0.6	0.500 ± 0.189	6.025 ± 0.671
	0.8	0.375 ± 0.183	5.850 ± 0.568

Table 3: Empirical probability and prediction set size of prediction sets on LMM-R1 across different levels ε.

The results are depicted in Fig. 3b, where we test configurations with the number of groups ranging from 60 to 300. For both models, the running time demonstrates a significant decrease as the number of groups increases. This empirically confirms that our groupwise operation provides substantial computational savings by partitioning groups, which aligns with our methodology. Furthermore, LMM-R1 consistently outperforms R1-Onevision across all tested group sizes, which is expected as LMM-R1 is a smaller model with fewer parameters.

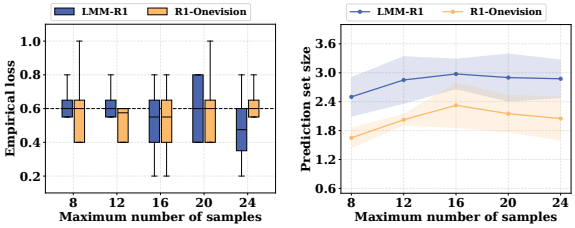
We also empirically examine the influence of the FWER control level ε on the uncertainty set introduced in Theorem 3.1. It controls the confidence level $1-\varepsilon$ that the configurations retained in $\Lambda_{\text{valid}}(\theta)$ satisfy the risk guarantee in Eq. (2). As shown in Table 3, our results across four different ε validate on CLEVR-Math and ScienceQA using LMM-R1. The average empirical probability computed over random subsets of the calibration data consistently meets or exceeds the $1-\varepsilon$ target for both datasets (e.g., 0.875 observed on ScienceQA when $\varepsilon = 0.2$). Moreover, these results illustrate the trade-off controlled by ε, as allowing a higher FWER by increasing ε makes the uncertainty set less conservative, which in turn leads to producing modestly smaller prediction set sizes.



(a) Validity (b) Efficiency

Figure 4: Influence of calibration set size for CoRAP.

We further explore how the calibration set size influences CoRAP in terms of validity and efficiency by varying the number of examples from 300 to 1500 in increments of 300 using LMM-R1 and LLaVA-CoT on ScienceQA. Results in Fig. 4a confirm that CoRAP consistently keeps the empirical loss below the target threshold $\alpha = 0.3$. Increasing the calibration size reduces loss variance and indicates improved stability in empirical loss on both models. Regarding the efficiency of our proposed methods, Fig. 4b reveals that the prediction set size stabilizes without significant fluctuation. These findings confirm that CoRAP provides uncertainty guarantees while maintaining stable prediction set sizes with limited calibration data.



(a) Validity (b) Efficiency

Figure 5: Influence of maximum numbers of samples.

We further investigate how the maximum sampling budget $K_{\max}$ influences CoRAP's performance. Fig. 5a confirms that CoRAP strictly adheres to the reliability constraint, with the empirical loss consistently fluctuating near the target $\alpha = 0.6$, regardless of the sampling scale. Notably, while LMM-R1 shows slightly higher variance in risk control at $K_{\max} = 20$, the overall trend remains well-calibrated. From an efficiency perspective in Fig. 5b, we observe that the prediction set sizes for both models initially increase and then stabilize as $K_{\max}$ expands. Moreover, R1-Onevision consistently produces more compact prediction sets than LMM-R1, indicating a higher density of correct reasoning paths within its early samples. The convergence of set sizes at higher $K_{\max}$ values demonstrates that CoRAP is computationally economical, providing rigorous uncertainty guarantees without necessitating exhaustive sampling.

5 Conclusion

In this paper, we design a novel framework to quantify and explain the uncertainty in LRMs with guarantees. To address the challenge of ensuring logical interdependent generations, we first propose CoRAP, which quantifies the uncertainty of the reasoning-answer structure with rigorous guarantees. Subsequently, to resolve the computational complexity in tracing the origins of uncertainty, we design a hierarchical example-to-step explanation framework using Shapley values with guarantees. Furthermore, we conduct the theoretical analysis for our proposed uncertainty quantification and explanation methods. Extensive experiments demonstrate the effectiveness of our approach in ensuring valid and interpretable uncertainty quantification

across complex reasoning tasks.

6 Limitations

Our experiments demonstrate that our methods achieve rigorous coverage guarantees and provide provable explanations for uncertainty, confirming the practical efficacy of our framework. However, our current study has several limitations. First, the experimental results are primarily based on specific datasets, so additional experiments on other types of data (e.g., legal or medical) are needed to verify generality. Second, the present study considers multimodal settings using Vision-Language Models. An important next step is to extend our evaluation to text-only settings, applying the framework to standard Large Language Models to assess whether the efficiency and stability gains remain consistent in unimodal tasks.

Acknowledgments

This work is supported in part by the US National Science Foundation under grants CNS-2350332 and IIS-2442750. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Sohyun An, Ruochen Wang, Tianyi Zhou, and Cho-Jui Hsieh. 2025. Don't think longer, think wisely: Optimizing thinking dynamics for large reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2025. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662.

Aobo Chen, Yangyi Li, Wei Qian, Kathryn Morse, Chenglin Miao, and Mengdi Huai. 2024. Modeling and understanding uncertainty in medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–567. Springer.

Aobo Chen, Yangyi Li, Chenxu Zhao, and Mengdi Huai. 2025. A survey of security and privacy issues of machine unlearning. *AI Magazine*, 46(1):e12209.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. Attributes as textual genes: Leveraging LLMs as genetic algorithm simulators for conditional synthetic data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19367–19389, Suzhou, China. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, and Dayiheng Liu. 2025a. Teaching language models to reason with tools. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yangyi Li, Aobo Chen, Wei Qian, Chenxu Zhao, Divya Lidder, and Mengdi Huai. 2024. Data poisoning attacks against conformal prediction. In *Forty-first International Conference on Machine Learning*.

Yangyi Li and Mengdi Huai. 2025. Quantifying uncertainty in natural language explanations of large language models for question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25005–25013, Suzhou, China. Association for Computational Linguistics.

Yangyi Li and Mengdi Huai. 2026. Uncertainty-aware language guidance for concept bottleneck models. *arXiv preprint arXiv:2602.23495*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Jingyu Liu, JingquanPeng JingquanPeng, Xiaopeng Wu, Xubin Li, Tiezheng Ge, Bo Zheng, and Yong Liu. 2025. Do not abstain! identify and solve the uncertainty. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17177–17197, Vienna, Austria. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Topology of reasoning: Understanding large reasoning models through reasoning graph properties. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

OpenAI. 2024. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Wei Qian, Chenxu Zhao, Yangyi Li, Fenglong Ma, Chao Zhang, and Mengdi Huai. 2024. Towards modeling uncertainties of self-explaining neural networks via conformal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14651–14659.

Wei Qian, Chenxu Zhao, Yangyi Li, Wenqian Ye, and Mengdi Huai. 2025. Towards unveiling predictive uncertainty vulnerabilities in the context of the right to be forgotten. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5130–5135.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Lloyd S Shapley. 1953. A value for n-person games.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025. DAST: Difficulty-adaptive slow-thinking for large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2322–2331, Suzhou (China). Association for Computational Linguistics.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404.

Jiayuan Su, Fulin Lin, Zhaopeng Feng, Han Zheng, Teng Wang, Zhenyu Xiao, Xinlong Zhao, Zuozhu Liu, Lu Cheng, and Hongwei Wang. 2025. Cp-router: An uncertainty-aware router between llm and lrm. *arXiv preprint arXiv:2505.19970*.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. 2025. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. 2023. Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36:7330–7350.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *Preprint*, arXiv:2503.10615.

Algorithm 1 Construction of Conformal Reasoning–Answer Prediction (CoRAP) Set

Input: Input x, query q, model π_θ, functions Q, F, A, thresholds $\hat{\lambda}(\theta) = (\hat{\lambda}_1(\theta), \hat{\lambda}_2(\theta), \hat{\lambda}_3(\theta))$, max samples $K_{\max}$

Output: Prediction set $\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta)$

Initialize $\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta) \leftarrow \emptyset$
for $k \leftarrow 1$ **to** $K_{\max}$ **do**
  Sample $\hat{a}_k = \hat{r}_k \| \hat{y}_k \sim \pi_\theta(\cdot \mid x, q)$
  if $Q(x, q, \hat{a}_k) < \hat{\lambda}_1(\theta)$ **then**
    continue
  end if
  $\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta) \leftarrow \mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta) \cup \{\hat{a}_k\}$
  if $\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta) \neq \emptyset$ **then**
    $a^* \leftarrow \arg\max_{a \in \mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x,q;\theta)} Q(x, q, a)$, where $a^* = r^* \| y^*$
    if $F(\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta)) \geq \hat{\lambda}_2(\theta)$ **and** $A(x, q, a^*) \geq \hat{\lambda}_3(\theta)$ **then**
      break
    end if
  end if
end for
return $\mathcal{C}^{\mathrm{RA}}_{\hat{\lambda}(\theta)}(x, q; \theta)$

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385.

Boxuan Zhang and Ruqi Zhang. 2025. CoT-UQ: Improving response-wise uncertainty quantification in LLMs with chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26114–26133, Vienna, Austria. Association for Computational Linguistics.

Jiaxin Zhang. 2025. Confidence-aware reasoning: Optimizing self-guided thinking trajectories in large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2081–2095, Suzhou (China). Association for Computational Linguistics.

Chenxu Zhao, Wei Qian, Yangyi Li, Aobo Chen, and Mengdi Huai. 2024. Rethinking adversarial robustness in the context of the right to be forgotten. In *Forty-first International Conference on Machine Learning*.

A Algorithm for CoRAP

In this section, we provide the detailed procedure for CoRAP. As described in Section 3.1, we discretize the threshold space into a finite grid Λ. For a fixed model θ and each candidate threshold $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in \Lambda$, we systematically evaluate its performance on the calibration set $\mathcal{D}_{\mathrm{cal}} = \{z_i\}_{i=1}^{n_{\mathrm{cal}}}$ with $z_i = (x_i, q_i, r_i, y_i)$ by running Algorithm 1 to construct the prediction set $\mathcal{C}^{\mathrm{RA}}_{\lambda}(x_i, q_i; \theta)$ and computing the binary loss $L^{\mathrm{RA}}(z_i; \lambda, \theta) \in \{0, 1\}$, where 1 indicates miscoverage. To formally certify rigorous risk control, we treat each λ as a hypothesis test with the null $H_{0,\lambda} : \mathbb{E}[L^{\mathrm{RA}}(z; \lambda, \theta)] > \alpha$. Since the total number of failures $S(\lambda) := \sum_{i=1}^{n_{\mathrm{cal}}} L^{\mathrm{RA}}(z_i; \lambda, \theta)$ follows a Binomial distribution under i.i.d. sampling, we compute the Binomial-tail p-value $p^{\mathrm{BT}}_{\lambda} = \Pr(\mathrm{Binom}(n_{\mathrm{cal}}, \alpha) \leq S(\lambda))$ to quantify the likelihood of observing at most $S(\lambda)$ failures if the true risk exceeded α. We then apply standard family-wise error rate (FWER) control at level ε to $\{p^{\mathrm{BT}}_{\lambda}\}_{\lambda \in \Lambda}$ to obtain a valid subset of thresholds $\Lambda_{\mathrm{valid}}(\theta) \subseteq \Lambda$. If $\Lambda_{\mathrm{valid}}(\theta)$ is empty, the procedure abstains; otherwise, we select an optimal threshold $\hat{\lambda}(\theta) = (\hat{\lambda}_1(\theta), \hat{\lambda}_2(\theta), \hat{\lambda}_3(\theta))$ based on a chosen efficiency criterion. Finally, as detailed in Algorithm 1, this $\hat{\lambda}(\theta)$ is subsequently deployed during inference to dynamically filter sampled reasoning-answer pairs via $\hat{\lambda}_1$ and terminate the sampling process early if the set-level and answer-level quality scores satisfy $\hat{\lambda}_2$ and $\hat{\lambda}_3$, respectively, or when the maximum sampling budget $K_{\max}$ is exhausted.

B Proofs of Theorems

Theorem 3.1. *Assume calibration examples in $\mathcal{D}_{\mathrm{cal}}$ and a test sample z_t are i.i.d., and CoRAP's sampling randomness is independent across examples. Let $\Lambda_{\mathrm{valid}}(\theta)$ be obtained by applying an FWER-ε procedure to the p-values for target level $\alpha \in (0, 1)$ on model θ. Then, for any $\hat{\lambda}(\theta) \in \Lambda_{\mathrm{valid}}(\theta)$, we have the risk constraint*

$$\mathbb{E}[L^{\mathrm{RA}}(z_t; \hat{\lambda}(\theta), \theta)] \leq \alpha, \tag{8}$$

with probability at least $1 - \varepsilon$ over $\mathcal{D}_{\mathrm{cal}}$.

Proof. Fix a model θ and a finite grid of candidate thresholds Λ. For each $\lambda \in \Lambda$, let $R(\lambda; \theta) := \mathbb{E}[L^{\mathrm{RA}}(z; \lambda, \theta)]$ denote the population miscoverage risk. On the calibration set $\mathcal{D}_{\mathrm{cal}} = \{z_i\}_{i=1}^{n_{\mathrm{cal}}}$, define the empirical risk $\widehat{R}(\lambda; \theta) := \frac{1}{n_{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{cal}}} L^{\mathrm{RA}}(z_i; \lambda, \theta)$ and the count of failures $S(\lambda) := n_{\mathrm{cal}} \widehat{R}(\lambda; \theta)$. By the i.i.d. assumption and the independence of CoRAP's sampling randomness across examples, for each fixed λ we have $S(\lambda) \sim \mathrm{Binom}(n_{\mathrm{cal}}, R(\lambda; \theta))$.

For the target level α, consider testing the one-sided null $H_{0,\lambda} : R(\lambda;\theta) > \alpha$. We use the binomial-tail p-value $p_\lambda^{\mathrm{BT}} := \Pr(\mathrm{Binom}(n_{\mathrm{cal}},\alpha) \le S(\lambda))$. Under $H_{0,\lambda}$, this p-value is super-uniform: for any $u \in [0,1]$, letting $c(u)$ be the largest integer with $\Pr(\mathrm{Binom}(n_{\mathrm{cal}},\alpha) \le c(u)) \le u$, we have

$$\begin{aligned}\Pr(p_\lambda^{\mathrm{BT}} \le u) &= \Pr(S(\lambda) \le c(u)) \\ &\le \Pr(\mathrm{Binom}(n_{\mathrm{cal}},\alpha) \le c(u)) \le u,\end{aligned} \tag{9}$$

where the inequality uses that the binomial CDF at a fixed count is non-increasing in the success probability, and $R(\lambda;\theta) > \alpha$ under the null.

Let $\Lambda_{\mathrm{valid}}(\theta)$ be the set returned by any FWER-ε procedure applied to $\{p_\lambda^{\mathrm{BT}}\}_{\lambda\in\Lambda}$. By FWER control and the super-uniformity above, with probability at least $1-\varepsilon$ over $\mathcal{D}_{\mathrm{cal}}$, no $\lambda \in \Lambda_{\mathrm{valid}}(\theta)$ violates the target risk, i.e., $R(\lambda;\theta) \le \alpha$ for all $\lambda \in \Lambda_{\mathrm{valid}}(\theta)$. In particular, this holds for any (possibly data-dependent) choice $\hat\lambda(\theta) \in \Lambda_{\mathrm{valid}}(\theta)$.

Finally, since z_t is an independent draw from the same distribution, $\mathbb{E}[L^{\mathrm{RA}}(z_t;\hat\lambda(\theta),\theta)] = R(\hat\lambda(\theta);\theta)$. With probability at least $1-\varepsilon$ over $\mathcal{D}_{\mathrm{cal}}$, we have $R(\hat\lambda(\theta);\theta) \le \alpha$, and hence

$$\mathbb{E}[L^{\mathrm{RA}}(z_t;\hat\lambda(\theta),\theta)] \le \alpha. \tag{10}$$

The proof is complete. □

Theorem 3.2. *Let z_t be a test example that is i.i.d. with the calibration examples in $\mathcal{D}_{\mathrm{cal}}$, and let $\mathcal{H} := \{\, L^{\mathrm{RA}}(z_t;\hat\lambda(\theta_0),\theta_0) = 1 \,\}$ denote the event that the base model θ_0 fails on z_t. Condition on the validity of the thresholds used in evaluating v_{ex} and v_{st}. Assume that for a target level $\alpha \in (0,1)$ and error bounds $\xi_{\mathrm{ex}}, \xi_{\mathrm{st}} \ge 0$, the condition $v(\mathcal{P}\setminus U) \ge v(\mathcal{P}) - \sum_{u\in U}\phi_u - \xi$ holds for $(v,\mathcal{P},\xi) = (v_{\mathrm{ex}},\mathcal{P}_{\mathrm{ex}},\xi_{\mathrm{ex}})$ and $(v,\mathcal{P},\xi) = (v_{\mathrm{st}},\mathcal{P}_{\mathrm{st}},\xi_{\mathrm{st}})$.*
(i) (Example-level). *With probability at least $1-\delta_{\mathrm{ex}}$ over the Shapley MC, the example subset $S_{\mathrm{ex}}^\star$ identified by the stopping condition $\sum_{u\in S_{\mathrm{ex}}^\star}\mathcal{B}_u^{\mathrm{ex}} \ge 1-\alpha$ is sufficient to achieve the risk control:*

$$\mathbb{E}[L^{\mathrm{RA}}(z_t;\hat\lambda(\theta^{S_{\mathrm{ex}}^\star}),\theta^{S_{\mathrm{ex}}^\star}) \mid \mathcal{H}] \;\le\; \alpha + \xi_{\mathrm{ex}}. \tag{11}$$

(ii) (Step-level). *Conditionally on $S_{\mathrm{ex}}^\star$, with probability at least $1-\delta_{\mathrm{st}}$ over the step-level Shapley MC, the step subset $T^\star$ identified by the stopping condition $\sum_{u\in T^\star}\mathcal{B}_u^{\mathrm{st}} \ge 1-\alpha$ is sufficient to achieve the risk control:*

$$\mathbb{E}[L^{\mathrm{RA}}(z_t;\hat\lambda(\theta^{T^\star}),\theta^{T^\star}) \mid \mathcal{H}] \;\le\; \alpha + \xi_{\mathrm{st}}. \tag{12}$$

Proof. We condition throughout on the event that the conformal thresholds $\hat\lambda(\theta^S)$ and $\hat\lambda(\theta^T)$ used to evaluate $v_{\mathrm{ex}}(\cdot;z_t)$ and $v_{\mathrm{st}}(\cdot;z_t)$ are valid for all coalitions considered.

We first prove the example-level statement. Let ϕ_u^{ex} be the Shapley value of the set function $v_{\mathrm{ex}}(\cdot;z_t)$ for each $u \in \mathcal{P}_{\mathrm{ex}}$. By the uniform lower confidence bound, with probability at least $1-\delta_{\mathrm{ex}}$ over the Shapley Monte Carlo, all bounds hold simultaneously, i.e., $\phi_u^{\mathrm{ex}} \ge \mathcal{B}_u^{\mathrm{ex}}$ for all u. On this event, the stopping rule implies $\sum_{u\in S_{\mathrm{ex}}^\star}\phi_u^{\mathrm{ex}} \ge \sum_{u\in S_{\mathrm{ex}}^\star}\mathcal{B}_u^{\mathrm{ex}} \ge 1-\alpha$. Let $U := \mathcal{P}_{\mathrm{ex}} \setminus S_{\mathrm{ex}}^\star$. By the inequality assumed in Theorem 3.2, applied to $v_{\mathrm{ex}}(\cdot;z_t)$, removing U from the full set can decrease the value by at most $\sum_{u\in U}\phi_u^{\mathrm{ex}} + \xi_{\mathrm{ex}}$. Using Shapley efficiency (the total Shapley mass equals $v_{\mathrm{ex}}(\mathcal{P}_{\mathrm{ex}};z_t) - v_{\mathrm{ex}}(\varnothing;z_t)$), we obtain the lower bound

$$v_{\mathrm{ex}}(S_{\mathrm{ex}}^\star;z_t) \ge v_{\mathrm{ex}}(\varnothing;z_t) + \sum_{u\in S_{\mathrm{ex}}^\star}\phi_u^{\mathrm{ex}} - \xi_{\mathrm{ex}}. \tag{13}$$

On the miscoverage event $\mathcal{H}$, the empty coalition corresponds to the initial configuration, hence $v_{\mathrm{ex}}(\varnothing;z_t) = 1 - L^{\mathrm{RA}}(z_t;\hat\lambda(\theta_0),\theta_0) = 0$. Combining this with $\sum_{u\in S_{\mathrm{ex}}^\star}\phi_u^{\mathrm{ex}} \ge 1-\alpha$ yields $v_{\mathrm{ex}}(S_{\mathrm{ex}}^\star;z_t) \ge 1-\alpha-\xi_{\mathrm{ex}}$ on $\mathcal{H}$. By definition (3), this implies $L^{\mathrm{RA}}(z_t;\hat\lambda(\theta^{S_{\mathrm{ex}}^\star}),\theta^{S_{\mathrm{ex}}^\star}) \le \alpha + \xi_{\mathrm{ex}}$ on $\mathcal{H}$, and taking conditional expectation given $\mathcal{H}$ gives the desired bound

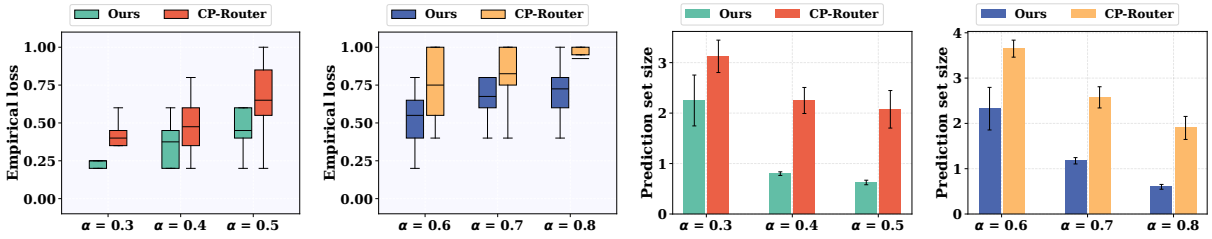
$$\mathbb{E}[L^{\mathrm{RA}}(z_t;\hat\lambda(\theta^{S_{\mathrm{ex}}^\star}),\theta^{S_{\mathrm{ex}}^\star}) \mid \mathcal{H}] \;\le\; \alpha + \xi_{\mathrm{ex}}. \tag{14}$$

We next prove the step-level statement. Condition on the realized $S_{\mathrm{ex}}^\star$ and consider the step universe $\mathcal{P}_{\mathrm{st}} = \{(j,s) : j \in S_{\mathrm{ex}}^\star,\ s \in [L_j]\}$. Let ϕ_u^{st} be the Shapley value of the set function $v_{\mathrm{st}}(\cdot;z_t)$ for each $u \in \mathcal{P}_{\mathrm{st}}$. By the same uniform-LCB argument, with probability at least $1-\delta_{\mathrm{st}}$ over the step-level Shapley Monte Carlo, we have $\phi_u^{\mathrm{st}} \ge \mathcal{B}_u^{\mathrm{st}}$ for all $u \in \mathcal{P}_{\mathrm{st}}$. On this event, the stopping rule implies $\sum_{u\in T^\star}\phi_u^{\mathrm{st}} \ge \sum_{u\in T^\star}\mathcal{B}_u^{\mathrm{st}} \ge 1-\alpha$.

Let $U := \mathcal{P}_{\mathrm{st}} \setminus T^\star$. By the same inequality in Theorem 3.2, applied to $v_{\mathrm{st}}(\cdot;z_t)$, removing U from the full set can decrease the value by at most $\sum_{u\in U}\phi_u^{\mathrm{st}} + \xi_{\mathrm{st}}$. Using Shapley efficiency gives

$$v_{\mathrm{st}}(T^\star;z_t) \ge v_{\mathrm{st}}(\varnothing;z_t) + \sum_{u\in T^\star}\phi_u^{\mathrm{st}} - \xi_{\mathrm{st}}. \tag{15}$$

On $\mathcal{H}$, the empty step coalition corresponds to training on no selected steps and retains the initial failure, so $v_{\mathrm{st}}(\varnothing;z_t) = 1 - L^{\mathrm{RA}}(z_t;\hat\lambda(\theta_0),\theta_0) = 0$.



(a) Validity on ScienceQA using LLaVA-CoT (b) Validity on CLEVR-Math using R1-Onevision (c) Efficiency on ScienceQA using LLaVA-CoT (d) Efficiency on CLEVR-Math using R1-Onevision

Figure 6: Empirical loss and efficiency results across datasets and models.

Therefore $v_{\text{st}}(T^\star; z_t) \geq 1 - \alpha - \xi_{\text{st}}$ on $\mathcal{H}$. By definition (4), this implies $L^{\text{RA}}(z_t; \hat{\lambda}(\theta^{T^\star}), \theta^{T^\star}) \leq \alpha + \xi_{\text{st}}$ on $\mathcal{H}$, and taking conditional expectation given $\mathcal{H}$ yields

$$\mathbb{E}[L^{\text{RA}}(z_t; \hat{\lambda}(\theta^{T^\star}), \theta^{T^\star}) \mid \mathcal{H}] \leq \alpha + \xi_{\text{st}}. \quad (16)$$

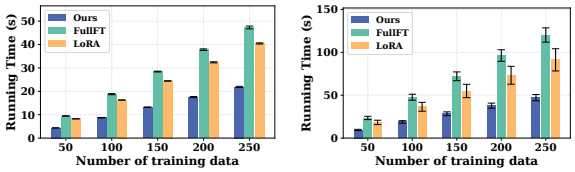
The proof is complete. □

C More Experimental Details

Datasets. We evaluate our method on two distinct benchmarks. First, CLEVR-Math is a mathematical dataset designed for visual arithmetic reasoning released under the CC BY 4.0 license. It consists of images depicting 3D geometric objects (such as spheres, cubes, and cylinders) paired with questions that require performing mathematical operations based on visual attributes. Second, ScienceQA is a large-scale multimodal benchmark covering natural, social, and language sciences distributed under the CC BY-NC-SA 4.0 license. It is publicly available and derived from standard educational materials. For our experiments, we specifically utilize the curated versions of these benchmarks from (Tan et al., 2025), which provide explicit Chain-of-Thought (CoT) annotations essential for LRMs. The utilized data comprises 35000 training samples for CLEVR-Math and 2112 training samples for ScienceQA. Given the synthetic nature of CLEVR-Math and the educational origin of ScienceQA, the data is free from personally identifying information and offensive material.

Parameter settings. In our experimental setup, we adopt 3B LMM-R1 (Peng et al., 2025), 7B R1-Onevision (Yang et al., 2025), and the 11B LLaVA-CoT (Xu et al., 2025) models on CLEVR-Math. Instead of full fine-tuning, we approximate parameter updates on the language model head using influence functions with a step size η of 5×10^{-6}. On ScienceQA, we similarly adopt the LLaVA-CoT, LMM-R1, and R1-Onevision models. These models utilize identical influence function configurations, applying a step size η of 5×10^{-6} to estimate the contribution of training factors. We set the default level ε to 0.2. To ensure rigorous explanation guarantees, we perform 256 Monte Carlo permutations with a failure probability δ of 0.25. For the admission function $V(z_i, \hat{r}_k)$, we admit a reasoning trace if its ROUGE-L (Lin, 2004) against the reference rationale is ≥ 0.2 (chosen empirically). V can be replaced by other verifiers. We implement our models with PyTorch (v2.5.1) and HuggingFace Transformers (v4.46.1). LoRA fine-tuning uses PEFT (v0.12.0) and ROUGE evaluation uses the official ROUGE implementation.

Machine configuration. The experiments are implemented using the PyTorch framework and run on a cluster equipped with AMD 32-core 2.6GHz CPUs and Nvidia A100 40/80GB GPUs.



(a) LMM-R1 (b) LLaVA-CoT

Figure 7: Influence of fine-tuning methods on ScienceQA.

D More Experimental Results

Validity. Figures 6a and 6b empirically present the validity of our proposed framework. We employ R1-Onevision on the CLEVR-Math dataset with target significance levels α set to 0.6, 0.7, and 0.8, while for LLaVA-CoT on ScienceQA, we utilize levels of 0.3, 0.4, and 0.5. Across these diverse configurations, the empirical losses of our method consistently stay within the specified thresholds. This confirms that our approach strictly adheres to the risk control requirements defined in Eq. (2) and fulfills the theoretical guarantees of Theorem 3.1. Conversely, the baseline CP-Router frequently exhibits empirical losses that violate the target α levels, signaling an inability to maintain reliable risk control in complex reasoning scenarios.

Efficiency. Beyond validity, we analyze the compactness of the constructed uncertainty sets in Figure 6c and 6d. Under the same experimental settings, our framework demonstrates superior efficiency compared to CP-Router by generating significantly smaller sets. For instance, when applying R1-Onevision to CLEVR-Math at $\alpha = 0.7$, our method yields a tight average set size of 1.175, implying that a minimal number of reasoning traces is sufficient to cover the true answer. In comparison, CP-Router produces much looser sets with an average size of 2.575 under identical conditions, resulting in higher redundancy. Overall, these results demonstrate that our CoRAP efficiently and effectively quantifies uncertainty by producing small and valid uncertainty sets.

Ablation study. In Fig. 7 we compare the running time of our proposed method against full fine-tuning (FullFT) and LoRA finetuning on the ScienceQA dataset, utilizing both LMM-R1 and LLaVA-CoT models, with the baselines run for 3 epochs. The results reveal that our method achieves a significantly lower computational cost when adopting the influence function methods. As the number of training data increases, the running time required by ours is substantially less than that of both FullFT and LoRA for both architectures, demonstrating its superior efficiency over these traditional techniques across different LRMs.

E Background

Large reasoning models (LRMs). We consider a supervised reasoning task using a dataset $\mathcal{D}_{\text{tr}} = \{(x_j, q_j, r_j, y_j)\}_{j=1}^{n_{\text{tr}}}$, where each instance consists of an input image x_j, a query q_j, a reasoning sequence $r_j = [r_{j,1}, \ldots, r_{j,L_j}]$, and a final answer y_j. We define the universe of all reasoning steps across training examples as $\mathcal{U}_{\text{st}} = \{(j, s) : j \in [n_{\text{tr}}],\ s \in [L_j]\}$. An LRM parameterizes a generation policy π_θ that models the joint probability of reasoning and answer tokens conditioned on the input. Representing each sample as a single autoregressive sequence $a_j = [r_{j,1}, \ldots, r_{j,L_j}, y_j]$, the supervised fine-tuning (SFT) objective maximizes the likelihood of generating the full sequence:

$$\mathcal{L}_{\text{SFT}} = \tag{17}$$
$$- \mathbb{E}_{(x,q,r,y)\sim\mathcal{D}_{\text{tr}}} \sum_{s=1}^{|a_j|} \log \pi_\theta(a_{j,s} \mid x_j, q_j.a_{j,<s}).$$

This objective provides the training setup used throughout the paper.

Split conformal prediction. Given a model θ, split conformal prediction (CP) provides reliability guarantees by outputting a prediction set $\mathcal{C}_\alpha$, rather than just one answer, which is guaranteed to contain the true answer y with a high and user-specified probability (e.g., $1 - \alpha$). This method works for any black-box model θ by using a separate calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, q_i, r_i, y_i)\}_{i=1}^{n_{\text{cal}}}$. The key component of split CP is a nonconformity score function $\mathcal{S}((x, q), y; \theta)$. This function measures how unusual or atypical a candidate answer y is for a given input (x, q) according to the model. A high score means the pair $((x, q), y)$ is non-conforming (i.e., the model finds it unlikely), while a low score means it is conforming. To provide its statistical guarantee, split CP requires that the calibration data $\mathcal{D}_{\text{cal}}$ and test data are exchangeable.

To determine the conformal prediction set for a test input (x_t, q_t), we test the nonconformity score for each potential answer $y \in \mathcal{Y}$ against a pre-defined significance level $\alpha \in (0, 1)$. The goal is to construct a prediction set $\mathcal{C}_\alpha$ that satisfies the marginal coverage guarantee $\mathbb{P}(y_t \in \mathcal{C}_\alpha(x_t, q_t; \theta)) \geq 1 - \alpha$, where (x_t, q_t, r_t, y_t) is a test point exchangeable with the calibration data.

Theorem E.1 ((Vovk et al., 2005)). *Assume that examples $z_i = (x_i, q_i, r_i, y_i)$ for $i = 1, \cdots, n_{\text{cal}}$ and z_t are exchangeable. For any nonconformity measure $\mathcal{S}$ and target risk $\alpha \in (0, 1)$, define the conformal set at (x_t, q_t) as $\mathcal{C}_\alpha(x_t, q_t; \theta) = \{y \in \mathcal{Y} : \mathcal{S}((x_t, q_t), y; \theta) \leq \hat{q}\}$, where $\hat{q} = Quantile(1 - \alpha; \{\mathcal{S}((x_i, q_i), y_i; \theta)\}_{i=1}^{n_{\text{cal}}} \cup \{\infty\})$. Then $\mathcal{C}_\alpha(x_t, q_t; \theta)$ satisfies*

$$\mathbb{P}(y_t \in \mathcal{C}_\alpha(x_t, q_t; \theta)) \geq 1 - \alpha. \tag{18}$$

In essence, this theorem provides the formal guarantee that the constructed prediction set $\mathcal{C}_\alpha(x_t, q_t; \theta)$ will fail to cover the true answer y_t with a probability of at most α, provided the calibration and test data are exchangeable.

F AI Assistance in Writing

During the preparation of this manuscript, we utilized AI tools (e.g., GPT-5.2, Gemini 3 Pro) exclusively for copy-editing purposes, including grammatical correction, phrasing refinement, and readability improvement. These tools were not employed to generate scientific concepts, conduct analyses, or formulate conclusions, nor did they influence the study's methodology or results. All substantive content was developed solely by the human authors, who retain full responsibility for the final text and editorial decisions.