CP-Router: An Uncertainty-Aware Router Between LLM and LRM

Jiayuan Su[1*], Fulin Lin[1*], Zhaopeng Feng[1*], Han Zheng[1], Teng Wang[2], Zhenyu Xiao[3], Xinlong Zhao[4], Zuozhu Liu[1†], Lu Cheng[5], Hongwei Wang[1†]

[1]Zhejiang University
[2]University of Hong Kong
[3]Tsinghua University
[4]Peking University
[5]University of Illinois Chicago

Abstract

Recent advances in large reasoning models (LRMs) have significantly enhanced long-chain reasoning capabilities over standard large language models (LLMs). However, LRMs often produce unnecessarily lengthy outputs even for simple queries, leading to inefficiencies or even accuracy degradation compared to LLMs. To address this, we propose CP-Router, a training-free, model-agnostic routing framework that dynamically selects between an LLM and an LRM, demonstrated with multiple-choice question answering (MCQA) prompts. The routing decision is guided by the prediction uncertainty estimates derived via Conformal Prediction (CP), which provides rigorous coverage guarantees. To improve uncertainty differentiation across inputs, we introduce Full and Binary Entropy (FBE), a novel entropy-based criterion that adaptively selects the appropriate CP threshold. Experiments across MCQA and QA benchmarks—including mathematics, logical reasoning, and Chinese chemistry—demonstrate that CP-Router efficiently reduces token usage while maintaining or even improving accuracy compared to using LRM alone. We further demonstrate the generality and robustness of CP-Router by extending it to diverse model pairings beyond the LLM–LRM setting.

Extended version — https://arxiv.org/abs/2505.19970

Introduction

Recent progress in reinforcement learning has substantially improved the reasoning capabilities of standard large language models (LLMs), leading to the emergence of large reasoning models (LRMs)—large models that exhibit markedly enhanced long-range reasoning abilities, e.g., OpenAI's o1/o3 (Jaech et al. 2024) and DeepSeek-R1 (Guo et al. 2025). While these models often outperform LLMs by leveraging increased reasoning tokens, they frequently generate unnecessarily redundant reasoning for straightforward prompts—i.e., inputs that are direct and require no multi-step inference. This inefficiency, commonly referred to as overthinking (Sui et al. 2025; Chen et al. 2025; Ballon, Algaba, and Ginis 2025), has become a growing concern.

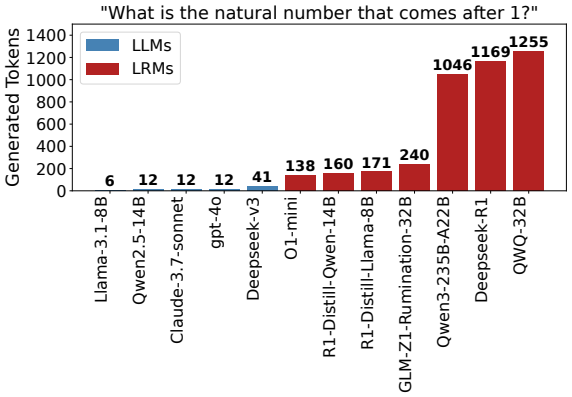

Figure 1: Token consumption for "What is the natural number that comes after 1?" LLMs provide correct answers with concise token usage, whereas LRMs consume significantly more tokens, suggesting a potential "overthinking" issue.

Recent studies (Su et al. 2025; Wu et al. 2025) indicate that excessive reasoning not only incurs additional computational overhead but can harm accuracy. In contrast, LLMs can produce concise and often more accurate responses with significantly lower cost (as illustrated in Figure 1 and Figure 4a). It reveals a critical inefficiency in current LRM behavior and raises the central question of our work: *How can we dynamically route prompts between an LLM and an LRM to minimize computational cost without sacrificing—and potentially even improving—accuracy?*

The exploration of effective LLM routing strategies remains limited. Pan et al. (2024) propose a majority-voting-based method to assess response consistency, routing inconsistent ones to a more deliberate reasoning path. However, it incurs high computational costs due to repeated sampling. More recently, advances such as Claude-3.7-Sonnet (Anthropic 2025) and Qwen3 (Yang et al. 2025a) empower users with manual controls to switch between reasoning and standard modes. However, this places a burden on users to make effective routing decisions. These limitations highlight the

*These authors contributed equally to this work.
†Corresponding author (hongweiwang@intl.zju.edu.cn).

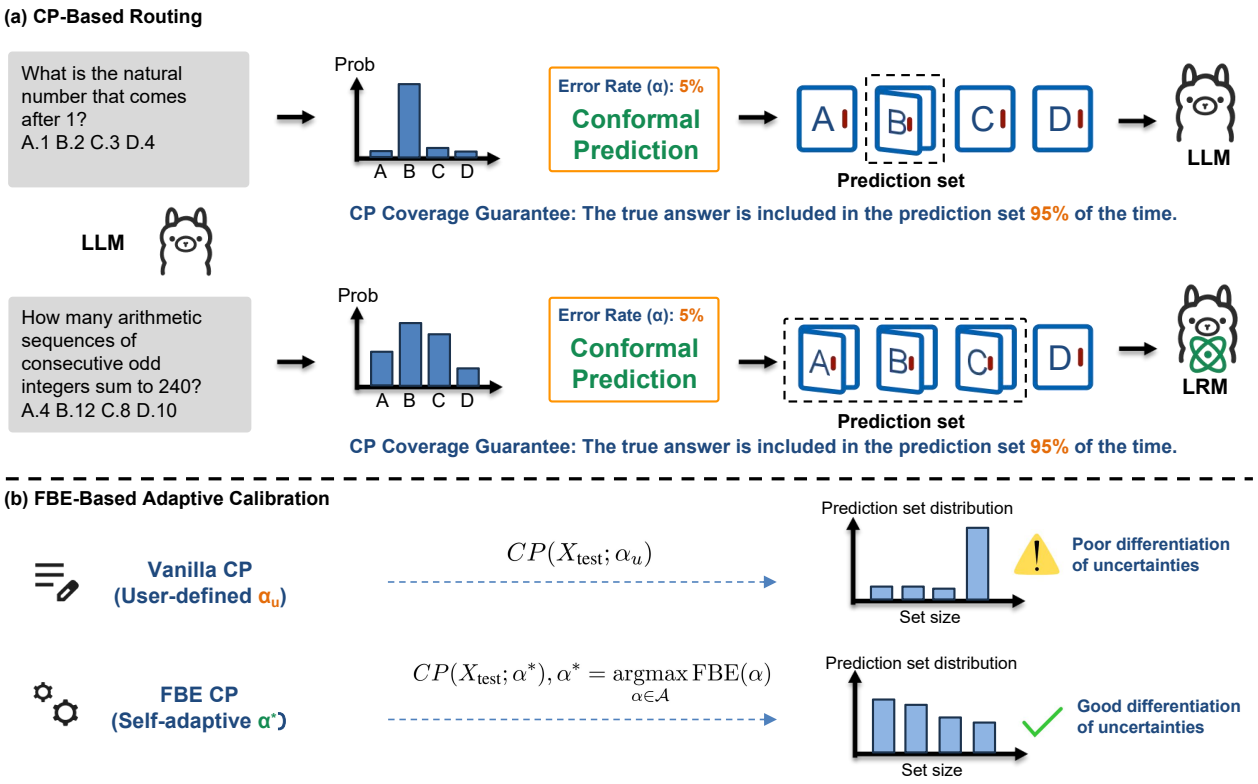

Figure 2: Key Components of the CP-Router. (a) CP-Based Routing. For each prompt, it applies CP with a target error rate α to generate a prediction set based on LLM output probabilities. Prompts with small prediction sets are routed to an LLM, while those with large sets are routed to an LRM. This enables uncertainty-aware, dynamic routing. (b) FBE-Based Adaptive Calibration. Instead of using a fixed error rate, CP-Router leverages FBE to select the optimal α that maximizes uncertainty separability, enabling more effective differentiation between easy and hard prompts—crucial for adaptive routing decisions.

urgent need for a practical and generalizable routing mechanism. An effective router is expected to exhibit the following key features: (1) **self-adaptive** — it should automatically determine whether a prompt is better suited for an LLM or an LRM, and adapt its decisions across diverse datasets; (2) **model-agnostic** — it should be compatible with model pairings comprising different families of LLMs and LRMs. (3) **lightweight** — it requires no training while maintaining low token overhead and high inference-time accuracy.

To design an effective routing strategy, we adopt Conformal Prediction (CP) (Vovk, Gammerman, and Shafer 2005; Angelopoulos and Bates 2021), a statistical framework that quantifies the prediction uncertainty of a model for a given input. CP is training-free and model-agnostic, naturally supporting the key requirements of a router. Crucially, CP also provides strong, finite-sample theoretical guarantees: given a user-specified error rate α, it constructs a *prediction set* that contains the true label with probability at least $1-\alpha$ (Vovk, Gammerman, and Shafer 2005). This coverage guarantee holds under the exchangeability assumption and distinguishes CP from other uncertainty quantification methods that may lack such formal correctness. Moreover, the size of the calibrated prediction set serves as a direct and interpretable proxy for uncertainty—larger sets naturally correspond to higher uncertainty. These properties make CP not just a practical option, but a principled and theoretically grounded choice for prompt routing.

We thus use CP as the backbone of our routing strategy. For each input, we obtain a prediction set from the LLM and use its size as a proxy for uncertainty. Prompts with small output sets (i.e., low uncertainty) are processed directly using the prediction from the LLM, while others are routed to the LRM. This strategy is illustrated in Figure 2a. Our empirical results in Figure 4 further support this strategy: low-uncertainty prompts are typically better handled by LLMs, while high-uncertainty prompts benefit more from LRMs.

One major challenge of applying CP in our task is selecting an appropriate error rate α for CP (Angelopoulos et al. 2020). Therefore, we propose a novel automatic calibration method that leverages a novel Full and Binary Entropy (FBE) metric to dynamically adjust α. FBE integrates both the global and binary entropy of the prediction set size distribution, enabling a more adaptive routing strategy, as illustrated in Figure 2b. Experiments across diverse benchmarks demonstrate that CP-Router achieves favorable accuracy and efficiency: it effectively reduces token overhead while maintaining or even improving accuracy compared to using LRM alone. We further extend CP-Router to diverse model pairings and the open-ended QA task, validating the generality of CP-Router. Our main contributions are:

- We introduce CP-Router as a novel uncertainty-aware routing mechanism that dynamically routes prompts between LLM and LRM based on estimated prediction uncertainty. CP offers strong coverage guarantees and produces interpretable prediction sets, making it particularly well-suited for prompt routing without requiring any additional training or model-specific tuning.
- We propose a novel automatic calibration method that leverages the FBE metric to adaptively select an appropriate CP error rate, ensuring well-separated prediction uncertainties and balancing accuracy and token efficiency.
- Experimental results demonstrate that CP-Router reduces token consumption compared to routing all prompts to LRMs, while maintaining comparable accuracy—and even surpassing it in some cases. We validate the effectiveness of FBE and show that CP-Router generalizes well across a broad range of models and tasks.

Conformal Prediction

CP (Vovk, Gammerman, and Shafer 2005; Angelopoulos and Bates 2021) is a statistical framework to estimate/quantify uncertainty. Given a user-specified error rate α (e.g., 10%), CP constructs a *prediction set* for each input in a principled manner (detailed in CP Steps), ensuring that the true label is included with probability approximately $1-\alpha$ (e.g., 90%). This guarantee is formally established in Theorem 0.1.

Terminologies

Score Function In CP, a score function (or nonconformity measure) $S(x, y) \in \mathbb{R}$ is defined to quantify the agreement (or "conformity") between an input x and a candidate output y. In our work, we focus on the multiple-choice question answering (MCQA) setting, and define $S(x, y)$ as $1 - f(y)$ (Sadinle, Lei, and Wasserman 2019), where $f(y)$ denotes the probability assigned to option y after applying the softmax over the logits of the candidate choices (e.g., 'A', 'B', 'C', and 'D'), conditioned on the prompt. Intuitively, higher scores indicate y is more plausible given x.

Prediction Set CP outputs a *prediction set*, which contains all candidate labels deemed plausible for a given input under a user-defined error rate. The set for an input x is defined as

$$C(x) = \{y \in \mathcal{Y} : S(x, y) \leq \hat{q}\}, \tag{1}$$

where $S(x, y)$ is the score function and $\hat{q}$ is the quantile threshold computed from the calibration set. The size of the prediction set can be interpreted in two complementary ways. First, under the same error rate, a larger prediction set indicates higher uncertainty for the input. Second, the distribution of prediction set sizes across different inputs reflects the adaptivity of the CP procedure, as shown in Figure 3. A wider spread of set sizes is desirable, since it means that the procedure is effectively distinguishing between easy and hard inputs(Angelopoulos and Bates 2021).

Formal Definition

Let (X, Y) be a sample, where X represents features and Y represents the outcome. Given a calibration set and a test set drawn independently and identically from the same underlying distribution, we denote the calibration set as $(X_i, Y_i)_{i=1,\ldots,n}$ and the test set as $(X_{\text{test}}, Y_{\text{test}})$. CP presents the following nesting property:

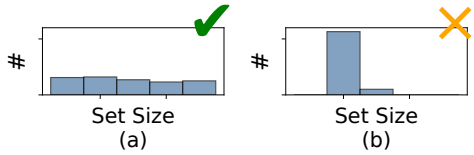

Figure 3: Different prediction set size distributions. A manually chosen error rate α might lead to a poor spread, as in (b), whereas (a) illustrates a desirable distribution.

$$\alpha_1 > \alpha_2 \Rightarrow C_{1-\alpha_1}(X) \subseteq C_{1-\alpha_2}(X). \tag{2}$$

Theorem 0.1 (Conformal Coverage Guarantee) *Suppose $(X_i, Y_i)_{i=1,\ldots,n}$ and (X_{test}, Y_{test}) are independent and identically distributed (i.i.d.). $C : \mathcal{X} \to 2^{\mathcal{Y}}$ is a set-valued mapping satisfying the nesting property in Eq. 2. The following holds:*

$$\mathbb{P}\left(Y_{test} \in \mathcal{C}(X_{test})\right) \geq 1 - \alpha \tag{3}$$

where $\alpha \in (0, 1)$ is the user-defined error rate, and $\mathcal{C}(X_{test})$ is the prediction set for input X_{test}.

CP Steps

The standard procedure involves four steps:

- Establish heuristic uncertainty notions, e.g., token logits/probabilities in LLM.
- Define the score function $S(x, y) \in \mathbb{R}$. We define the score as $1 - f(y)$, where $f(y)$ denotes the probability assigned to the option y.
- Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the scores, where n is the number of calibration samples, and α is a user-defined error rate.
- Use $\hat{q}$ to generate prediction sets for new samples: $C(X_{\text{test}}) = \{Y : S(X_{\text{test}}, Y) \leq \hat{q}\}$. The sets may vary in size depending on the uncertainty of the predictions.

Finally, we obtain statistically grounded prediction sets that quantify prediction uncertainty and guide **CP-Router**, which uses their size to route between LLMs and LRMs.

CP-Router

We first verify that the prediction set produced by CP serves as a reliable proxy for uncertainty. Then, we employ an FBE-guided calibration strategy to adaptively tune the error rate, enhancing uncertainty separation across inputs. The calibrated error rate is used to perform CP, constructing prediction sets that enable effective routing and achieve a strong balance between accuracy and efficiency.

Uncertainty Estimation via CP

Building on the CP framework, we estimate prediction uncertainty for each input under the LLM to determine whether a question should be routed to the LLM or the LRM. The key intuition is that questions with low prediction uncertainty

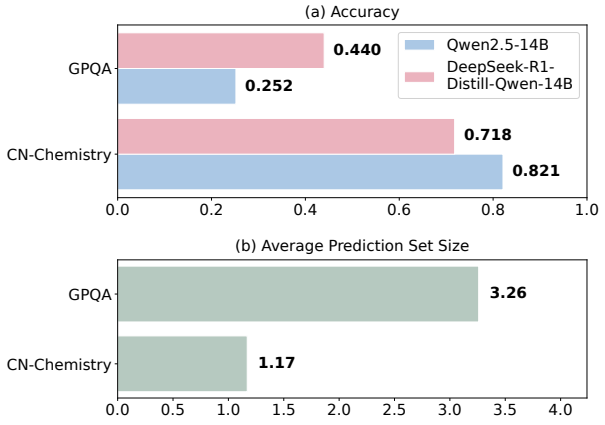

Figure 4: (a) Accuracy comparison between Qwen2.5-14B (LLM) and DeepSeek-R1-Distill-Qwen-14B (LRM) on GPQA and CN-Chemistry. LRM performs better on GPQA, while LLM achieves higher accuracy on CN-Chemistry. (b) Average Prediction Set Size (APSS) of Qwen2.5-14B under an error rate of 0.2 on both datasets. GPQA exhibits a larger APSS, suggesting higher uncertainty.

can be confidently handled by the LLM, while those with high uncertainty should be routed to the LRM for further consideration. To validate this intuition, we conduct experiments using Qwen2.5-14B (Yang et al. 2025b) (LLM) and DeepSeek-R1-Distill-Qwen-14B (Guo et al. 2025) (LRM) on two benchmarks: GPQA (Rein et al. 2024) and CN-Chemistry (Zhang et al. 2024). We first evaluate the accuracy of both models on these tasks, and then apply CP under Qwen2.5-14B to compute the Average Prediction Set Size (APSS). As illustrated in Figure 4, Qwen2.5-14B achieves higher accuracy on CN-Chemistry, which exhibits a smaller APSS—suggesting low prediction uncertainty. In contrast, DeepSeek-R1-Distill-Qwen-14B performs better on GPQA, where the APSS is larger—indicating higher uncertainty. These results empirically support our intuition.

Since our framework is designed for MCQA prompts, we extract the logits corresponding to the answer choices (e.g., 'A', 'B', 'C', and 'D') for each prompt and apply the softmax function to obtain a probability distribution over the options. These probabilities define the nonconformity scores used in CP procedure. CP then constructs a prediction set that includes the plausible answer options under a given error rate α. We adopt a prediction set size-aware routing rule: if the prediction set is small (e.g., of size 1), this indicates low uncertainty, and the prompt is handled by the LLM. Otherwise, it is routed to the LRM for more deliberate processing.

Adaptive CP Routing

However, a central challenge in using CP for routing lies in selecting the error rate α (Angelopoulos and Bates 2021), which determines the target coverage of the prediction sets. Although α is user-defined, there are no established principles for tuning it specifically for routing purposes. As illustrated in the section of CP, the choice of α influences the prediction set sizes and, in turn, affects routing decisions. An inappropriate α can lead to suboptimal adaptivity: a small α tends to produce large APSS, causing the LLM to reject many prompts and overburden the LRM; conversely, a large α may yield overly small APSS, reducing the engagement of the LRM even when necessary. Additionally, certain α values can result in prediction set distributions with poor separation between high- and low-uncertainty prompts (Angelopoulos and Bates 2021), as shown in Figure 3b, limiting the flexibility of the router. In our experiments, we explore the impact of different error rates in our experiments (Table 3), and observe that the choice of α plays a critical role in balancing accuracy and token efficiency.

To address the sensitivity to the choice of error rate α, we propose an adaptive calibration strategy based on a novel criterion: Full and Binary Entropy (FBE). FBE is designed to capture the distribution of prediction set sizes from two complementary perspectives. The full entropy term measures the overall diversity across all prediction set sizes, helping to ensure well-separated uncertainty estimates. The binary entropy term captures the balance between singleton and non-singleton prediction sets, which is critical for avoiding routing load imbalance. Formally, FBE is defined as:

$$\begin{aligned}\text{FBE} &= \beta * H_{\text{full}} + H_{\text{binary}} \\ &= -(\beta * \sum_i p_i \log p_i \\ &\quad + \left[p_{\text{s}=1} \log p_{\text{s}=1} + p_{\text{s}\neq 1} \log p_{\text{s}\neq 1}\right]),\end{aligned} \tag{4}$$

where p_i denotes the normalized frequency of prediction set size i, $p_{\text{s}=1}$ is the total frequency of singleton prediction sets (size = 1), and $p_{\text{s}\neq 1}$ is the combined frequency of all non-singleton sets. H_{full} promotes diversity in uncertainty, while H_{binary} encourages a meaningful split between certain and uncertain cases. In our implementation, we perform a grid search over candidate values of α and compute FBE based on the prediction sets generated on the test set. We then select the α that maximizes FBE, as shown below:

$$\alpha^* = \underset{\alpha \in \mathcal{A}}{\text{argmax}}\, \text{FBE}. \tag{5}$$

With the calibrated error rate α^* selected via FBE, we can finalize the routing procedure: For each test prompt $x^{(j)}$, we use the LLM f_θ to compute a softmax distribution over the answer choices and construct the prediction set $C^{(j)}$. We then use the prediction set size-aware routing rule.

Experiments

Experimental Setup

Backbone LLMs and LRMs. For LLMs, we include Llama-3-8B (Grattafiori et al. 2024), Qwen-2.5-14B (Yang et al. 2024), and DeepSeek-V3 (Liu et al. 2024). For LRMs, we use DeepSeek-R1-Distilled-Llama-3-8B, DeepSeek-R1-Distilled-Qwen-2.5-14B, and DeepSeek-R1 (Guo et al. 2025).

Benchmarks. We evaluate our method on six MCQA and a QA benchmarks spanning several domains. For mathematics, we include MMLU-STEM (Hendrycks et al. 2020),

Method	Threshold	Elementary Math			High School Math			College Math			LogiQA			GPQA			STEM-MCQA		
		Acc	TRR	U_{token}	Acc	TRR	U_{token}	Acc	TRR	U_{token}	Acc	TRR	U_{token}	Acc	TRR	U_{token}	Acc	TRR	U_{token}
Llama-3.1-8B and **DeepSeek-R1-Distill-Llama-8B**																			
LLM	-	41.6	-	-	31.9	-	-	31.2	-	-	35.9	-	-	29.6	-	-	24.4	-	-
LRM	-	79.9	-	38.3	70.4	-	38.5	58.8	-	27.6	46.4	-	10.5	34.0	-	4.4	65.0	-	40.6
Random	0.2	72.3	19.8	38.3	63.0	19.3	38.5	53.8	16.0	26.9	44.7	19.7	11.0	33.1	19.9	4.4	57.3	19.4	40.8
	0.3	69.0	29.4	38.8	59.1	29.4	38.5	51.5	25.3	27.2	43.5	29.1	10.8	32.7	29.4	4.4	53.4	28.9	40.8
	0.4	65.7	39.3	39.8	55.6	39.1	38.8	49.3	34.5	27.6	42.8	39.3	11.4	33.1	38.7	5.7	49.4	38.9	40.9
	0.5	61.9	49.4	40.1	51.5	49.0	38.4	45.5	45.0	26.0	41.8	48.9	11.5	32.1	48.4	4.8	45.9	42.5	37.4
Top-1	0.6	79.2	10.9	42.2	69.9	0.9	38.4	58.8	1.3	27.9	48.0	8.5	13.2	34.0	5.0	4.6	65.2	0.4	41.0
	0.7	79.5	6.9	40.8	70.4	0.5	38.7	58.8	1.3	27.9	47.0	3.1	11.5	34.0	0.6	4.4	65.2	0	40.8
	0.8	80.5	2.6	40.0	70.4	0	38.5	58.8	0	27.6	46.6	1.0	10.9	34.0	0.6	4.4	65.2	0	40.8
Entropy	1.0	80.2	2.6	39.6	70.4	0	38.5	58.8	0	27.6	46.6	1.0	10.9	34.0	0.6	4.4	65.2	0	40.8
	1.2	79.9	4.6	40.1	70.4	0.5	38.7	58.8	1.3	27.9	46.6	1.2	10.9	34.0	0.6	4.4	65.2	0	40.8
	1.4	79.9	5.6	40.5	70.4	0.5	38.7	58.8	1.3	27.9	47.2	2.1	11.6	34.0	0.6	4.4	65.2	0	40.8
Explicit	-	41.6	0	0	31.9	0	0	31.3	0	0.1	35.7	1.7	-0.2	29.6	1.3	0	24.4	0.4	0
Dynathink	-	50.3	1.1	8.8	61.3	2.9	30.2	55.3	7.6	26.1	45.1	9.1	10.1	34.0	3.1	4.5	50.1	2.2	26.3
CP-Router	-	78.2	14.2	**42.7**	68.5	7.9	**39.7**	60.0	8.8	**31.6**	48.2	10.0	**13.7**	35.2	5.7	**5.9**	63.3	6.4	**41.6**
Qwen-2.5-14B and **DeepSeek-R1-Distill-Qwen-2.5-14B**																			
LLM	-	46.9	-	-	31.5	-	-	36.2	-	-	44.9	-	-	25.2	-	-	30.4	-	-
LRM	-	93.4	-	46.5	84.3	-	52.8	83.8	-	47.6	59.9	-	15.0	44.0	-	18.8	79.0	-	48.6
Random	0.2	84.4	19.8	46.8	75.4	19.3	54.3	76.8	16.0	48.3	56.8	19.7	14.8	40.4	19.9	18.9	70.5	19.4	49.8
	0.3	79.9	29.4	46.8	70.9	29.4	55.9	71.8	25.3	47.6	55.7	29.1	15.2	39.0	29.4	19.6	65.9	28.9	50.0
	0.4	75.6	39.3	47.3	65.7	39.1	56.1	66.0	34.5	45.5	54.6	39.3	16.0	37.2	38.7	19.6	61.3	38.9	50.5
	0.5	70.5	49.4	46.6	60.7	49.0	57.1	62.8	45.0	48.3	53.7	48.9	17.3	36.2	45.4	20.1	57.0	46.7	49.9
Top-1	0.6	89.8	23.3	55.9	85.2	5.8	57.0	81.3	8.8	49.4	60.1	16.5	18.2	45.2	3.4	20.7	79.4	0.6	49.3
	0.7	92.1	20.8	57.0	85.7	1.9	55.2	83.8	3.8	49.4	59.9	14.8	17.6	44.0	1.3	19.1	79.4	0.2	49.1
	0.8	93.4	11.6	52.6	86.6	0.9	55.6	83.8	2.5	48.8	59.9	6.0	15.9	44.0	0	18.8	79.4	0	49.0
Entropy	1.0	93.7	11.6	53.0	86.6	0.9	55.6	83.8	2.5	48.8	59.7	5.2	15.6	44.0	0	18.8	79.4	0	49.0
	1.2	92.7	16.8	55.1	85.7	1.9	55.2	83.8	3.8	49.4	59.9	10.8	16.8	44.0	0	18.8	79.4	0	49.0
	1.4	91.8	19.1	55.4	85.7	3.2	42.8	82.5	6.3	49.4	59.7	18.7	18.2	44.7	3.1	20.1	79.4	0.2	49.1
Explicit	-	67.0	45.5	36.9	57.4	44.0	46.3	57.5	41.3	36.3	51.8	38.4	11.2	38.1	38.4	21.0	58.3	46.0	**51.7**
Dynathink	-	77.3	1.3	30.8	71.3	5.3	42.0	72.5	5.8	38.5	52.3	17.0	8.0	36.2	12.6	12.6	62.1	1.0	32.0
CP-Router	-	92.4	22.8	**58.9**	84.7	7.3	**57.4**	82.5	7.5	**50.1**	59.5	21.7	**18.7**	44.7	7.9	**21.2**	77.3	7.5	50.7

Table 1: Main results for the Llama pairing (Llama-3.1-8B and DeepSeek-R1-Distill-Llama-8B) and the Qwen pairing (Qwen-2.5-14B and DeepSeek-R1-Distill-Qwen-14B) across various benchmarks. CP-Router achieves the highest token utility across all evaluated benchmarks with the Llama pairing, and on 5 out of 6 benchmarks with the Qwen pairing. Even in the single case where the Explicit baseline surpasses CP-Router in token utility, its accuracy is 19% lower than that of CP-Router.

which covers elementary mathematics, high school statistics, and college mathematics; STEM-MCQA (Wordsmiths 2024); and GSM8K (Cobbe et al. 2021). For logical reasoning, we use GPQA (Rein et al. 2024) and LogiQA (Liu et al. 2021). We also include CN-Chemistry (Zhang et al. 2024), a Chinese-language chemistry benchmark.

Baselines. Baselines fall into two categories:

- **Single Model:** Using only an LLM or an LRM.
- **Hybrid Models with different routing strategies:** Including random routing, top-1 probability routing (inspired by Chuang et al. (2025)), response entropy routing, Dynathink (majority voting based routing) (Pan et al. 2024) and Explicit (explicit self-awareness routing). A detailed description of baselines can be found in Appendix.

Metrics. We evaluate effectiveness and efficiency using three metrics: (1) *Acc* (Accuracy), which measures the correctness of final predictions; (2) *TRR* (Token Reduction Ratio), defined as the proportion of output tokens saved compared to LRM, including those used for routing and inference/reasoning; and (3) U_{token} (Token Utility), which captures the improvement in prediction accuracy per unit of token usage, defined in Equation 6 as the ratio of accuracy gain over token usage ratio of a given method relative to LRM. U_{token} is introduced to facilitate a fairer comparison across methods with varying levels of token consumption.

$$U_{\text{token}} = \frac{Acc \text{ - } Acc_{LLM}}{1 \text{ - } TRR}. \tag{6}$$

Main Results

The performance of CP-Router on seven MCQA benchmarks, using Qwen pairing and Llama pairing, is summarized in Table 1. Our method achieves the highest token utility on all 6 benchmarks with Llama pairing, and on 5 out of 6 with Qwen pairing, while maintaining competitive accuracy compared to all baselines. For example, on College Math with Llama pairing, we surpass LRM's accuracy by 1.2% while also reducing token consumption.

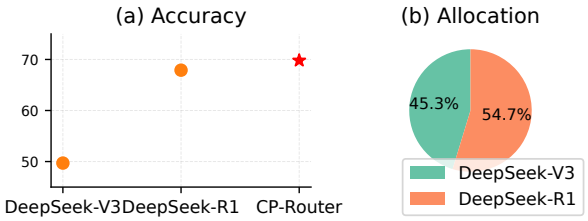

Figure 5: Accuracy and prompt allocation of CP-Router on the GPQA benchmark using **DeepSeek-V3** and **DeepSeek-R1**. CP-Router improves overall accuracy while reducing token consumption by avoiding routing a portion of the prompts to the more expensive R1 model.

On the STEM-MCQA benchmark with Qwen pairing, although the Explicit baseline attains the highest token utility, its accuracy is significantly lower than that of CP-Router, with a 19% decrease. This indicates that our method not only adapts to reduce token usage but also maintains—or even improves—the accuracy of the LRM. We also observe that different benchmarks correspond to different TRR values. For example, on Elementary Math, the Qwen pairing achieves a higher TRR of 22.8%. In contrast, on High School Math and College Math—which are intuitively more difficult and require greater reasoning—lower TRR values are observed. This demonstrates that CP-Router can dynamically adjust its routing based on prediction uncertainty, making it adaptive rather than fixed.

Analysis and Ablations

We conduct additional experiments and analyses on routing performance with large proprietary LLMs, as well as cases where the LLM is stronger than the LRM. We also present ablation studies on FBE and extend our framework to QA tasks and diverse model pairings beyond LLM–LRM combinations. Further details on hyperparameters, additional results, and case analyses are provided in the Appendix.

Routing Larger LLMs

Due to computational constraints, experiments involving larger LLMs such as DeepSeek-V3 and DeepSeek-R1 are conducted only on the GPQA benchmark. The results in Figure 5 show that the CP-Router outperforms both the standalone V3 and R1 models in terms of accuracy. Furthermore, it substantially reduces token consumption by routing only 54.7% of the prompts to R1. These results demonstrate that the proposed method is effective not only for 7B and 14B models, but also scales well to larger, more robust ones, underscoring its model-agnostic nature.

Performance when LLM is Stronger than LRM

As shown in Table 2, the CN-Chemistry benchmark presents a case where LLM achieves higher accuracy in the 14B setting. We evaluate CP-Router under this condition and observe that it outperforms LLM by more than 0.8% in terms of accuracy. In the 8B pairing, LRM achieves higher standalone performance, yet CP-Router routes effectively and still improves overall accuracy by 1.6%. These results demonstrate that CP-Router is flexible and robust—it can adaptively leverage the strengths of the available models and even surpass the stronger LLM in the pairing. Furthermore, since CN-Chemistry is a Chinese-language benchmark, this further highlights the multilingual robustness of CP-Router.

Ablations on FBE Components

To assess the contribution of FBE in CP-Router, we conduct a comprehensive ablation study on its two components: full entropy and binary entropy. As shown in Table 3, removing either component degrades routing performance in distinct ways. Excluding full entropy achieves substantial token savings (over 50% TRR), but at the cost of a notable accuracy drop—for instance, from 68.5% to 51.9% on the High School Math subset. In contrast, using only binary entropy preserves accuracy (65.0% on Elementary Math) but results in minimal token reduction (as low as 0.9% TRR). Including both components strikes a favorable balance between efficiency and accuracy. Our results demonstrate that different entropy-based methods (lead to varying error rates), which affect prediction set sizes and, consequently, overall performance. This highlights the importance of reasonable automatic calibration. Based on these findings, we adopt a 3:1 weighting between full and binary entropy in experiments.

Extension to General QA Format

We further extend our framework to the QA benchmark GSM8K, using the Llama-3-8B and DeepSeek-R1-Distill-Llama-8B pairing for evaluation. To adapt this open-ended format to our routing setup, we first prompt the LLM to generate its four most likely answers and then append an additional option, 'Others', to form a five-choice MCQ. We then apply the CP-Router procedure in the same manner as for the MCQA setting. As shown in Figure 6, CP-Router achieves a competitive accuracy of 77.9%, trailing the LRM by only 1.1%, while saving reasoning-token consumption for 32.9% of the questions. These results demonstrate the strong generalization capability of CP-Router to broader QA formats. We also report wall-clock time in Table 4, which shows that CP-Router reduces inference time by 9.3% and total completion token usage by 33%.

Generalization Across Diverse Model Pairings

CP-Router is designed to be model-agnostic, capable of routing prompts across arbitrary model pairings. To evaluate

Pairings	Standard Acc	Reasoning Acc	CP-Router Acc	CP-Router TRR
Llama-8B	45.3	46.0	47.6	18.9
Qwen-14B	82.1	71.6	82.9	68.0

Table 2: Performance on CN-Chemistry. The LLM achieves higher accuracy in the 14B setting, while the LRM performs better in the 8B setting. In both cases, CP-Router further improves upon the stronger model's accuracy.

Full	Binary	Elementary Math		High School Math	
		Acc	TRR	Acc	TRR
✗	✓	65.0	50.8	51.9	52.7
✓	✗	78.2	10.6	69.0	0.9
1	1	77.9	14.5	58.3	33.4
2	1	78.2	14.2	66.2	13.4
3	1	78.2	14.2	68.5	7.9

Table 3: Ablation study on STEM-MMLU benchmark subsets Elementary Math and High School Math. Different entropy-based methods result in varying error rates, which affect prediction set sizes and influence overall performance.

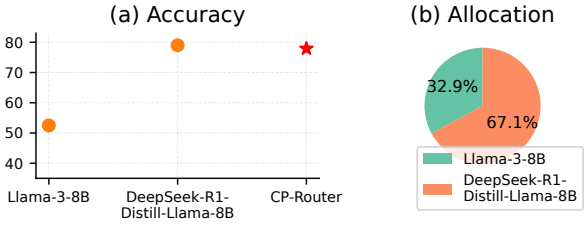

Figure 6: Accuracy and prompt allocation of CP-Router on the open-ended QA benchmark GSM8K, using **Llama-3-8B** and **DeepSeek-R1-Distill-Llama-8B**, demonstrating its generalization to broader QA formats.

its generalization ability, we conduct experiments on GPQA using both cross-family pairings (e.g., Qwen LLM–Llama LRM) and intra-capability pairings (e.g., Qwen LLM-Llama LLM), as shown in Figure 7. Across all configurations, CP-Router demonstrates robust performance. For instance, in the Llama-3-8B and Distill-Qwen-14B pairing, CP-Router even slightly surpassing LRM. In the Qwen-2.5-14B and Distill-Llama-8B pairing, it matches the LRM's accuracy (34.6%) while substantially outperforming LLM (25.2%). Even in the Qwen-2.5-14B and Llama-3-8B pairing—where both models are LLMs—CP-Router remains competitive. These results show that CP-Router generalizes effectively across both architectural and capability dimensions.

Related Work

LLM Routing

LLM routing has received attention but remains underdeveloped. Training-based methods such as Ong et al. (2024) and Aytes, Baek, and Hwang (2025) rely on classifiers to dispatch prompts, while Chuang et al. (2024) introduce uncertainty-specialized tokens to enable confidence-aware routing. These methods require substantial training, limiting generalizability across diverse pairings. Beyond learned classifiers, Pan et al. (2024) use majority voting to route prompts with consistent outputs to lightweight paths, though repeated sampling is computationally costly. Recently, models like Claude 3.7 Sonnet (Anthropic 2025) and Qwen3 (Yang et al. 2025a) offer manual control over reasoning modes, shifting the routing burden to users who may lack the expertise to make effective decisions.

Metric	LRM	CP-Router
Total Wall Time (s)	71,974.5	44,623.8
Avg. Wall Time per Item (s)	68.7	62.3
Total Completion Tokens	2,333,412	1,565,720

Table 4: Wall-clock time and token usage for the GSM8K benchmark using LLaMA-3-8B and DeepSeek-R1-Distill-LLaMA-8B across LRM and CP-Router methods.

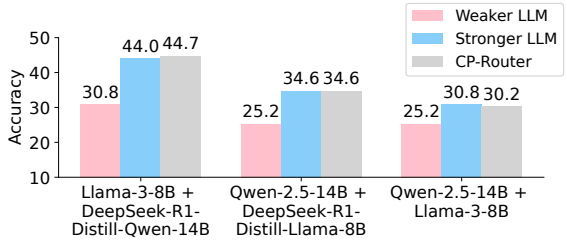

Figure 7: CP-Router performance on GPQA under different model pairings. We evaluate both cross-family (e.g., Qwen LLM–Llama LRM) and intra-capability (e.g., Llama LLM-Qwen LLM) pairings to validate the model-agnostic nature of CP-Router. Results show CP-Router generalizes effectively across diverse pairings.

Conformal Prediction in LLMs

CP (Vovk, Gammerman, and Shafer 2005; Zhou et al. 2025) has been applied to a range of natural language understanding tasks (Dey et al. 2022; Maltoudoglou, Paisios, and Papadopoulos 2020; Ernez et al. 2023) prior to the rise of LLMs. More recent efforts have extended CP to LLMs, aiming to provide statistically rigorous uncertainty estimates. Kumar et al. (2023) and Quach et al. (2024) are among the first to adapt CP for LLMs. Su et al. (2024) further addresses the limitations of API-based LLMs by introducing sampling-based nonconformity scores. Wang et al. (2024) proposes ConU to incorporate self-consistency into CP. Cherian, Gibbs, and Candes (2024) develops adaptive conformal inference techniques that offer conditional validity guarantees. While they improve CP's reliability for LLMs, its application in routing remains unexplored.

Conclusion

This paper introduces CP-Router, a training-free and model-agnostic routing mechanism grounded in CP for efficient prompt routing between LLM and LRM. CP-Router dynamically selects either an LLM or an LRM based on the size of CP prediction sets, with error rates calibrated via FBE to more effectively differentiate uncertainty across inputs. Extensive experiments across diverse tasks demonstrate that CP-Router substantially reduces token usage while maintaining—and occasionally exceeding—the accuracy of LRMs. We further show that CP-Router generalizes across model scales and families, and can be extended beyond the MCQA task. We hope our findings inspire future research in uncertainty-aware LLM routing.

Acknowledgments

We gratefully acknowledge that this work is supported by the National Key Research and Development Program of China (2024YFF0907800), the Zhejiang Provincial Natural Science Foundation of China (LDT23F02023F02), and the Research Fund for International Scientists of the National Natural Science Foundation of China (72350710798).

References

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anthropic. 2025. Claude 3.7 Sonnet System Card.

Aytes, S. A.; Baek, J.; and Hwang, S. J. 2025. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. *arXiv preprint arXiv:2503.05179*.

Ballon, M.; Algaba, A.; and Ginis, V. 2025. The Relationship Between Reasoning and Performance in Large Language Models–o3 (mini) Thinks Harder, Not Longer. *arXiv preprint arXiv:2502.15631*.

Chen, X.; Xu, J.; Liang, T.; He, Z.; Pang, J.; Yu, D.; Song, L.; Liu, Q.; Zhou, M.; Zhang, Z.; Wang, R.; Tu, Z.; Mi, H.; and Yu, D. 2025. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. arXiv:2412.21187.

Cherian, J.; Gibbs, I.; and Candes, E. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37: 114812–114842.

Chuang, Y.-N.; Yu, L.; Wang, G.; Zhang, L.; Liu, Z.; Cai, X.; Sui, Y.; Braverman, V.; and Hu, X. 2025. Confident or Seek Stronger: Exploring Uncertainty-Based On-device LLM Routing From Benchmarking to Generalization. *arXiv preprint arXiv:2502.04428*.

Chuang, Y.-N.; Zhou, H.; Sarma, P. K.; Gopalan, P.; Boccio, J.; Bolouki, S.; and Hu, X. 2024. Learning to Route with Confidence Tokens. *arXiv preprint arXiv:2410.13284*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dey, N.; Ding, J.; Ferrell, J.; Kapper, C.; Lovig, M.; Planchon, E.; and Williams, J. P. 2022. Conformal prediction for text infilling and part-of-speech prediction. *The New England Journal of Statistics in Data Science*.

Ernez, F.; Arnold, A.; Galametz, A.; Kobus, C.; and Ould-Amer, N. 2023. Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0). In Papadopoulos, H.; Nguyen, K. A.; Boström, H.; and Carlsson, L., eds., *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, 16–35. PMLR.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Kumar, B.; Lu, C.; Gupta, G.; Palepu, A.; Bellamy, D.; Raskar, R.; and Beam, A. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2021. LogiQA: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3622–3628.

Maltoudoglou, L.; Paisios, A.; and Papadopoulos, H. 2020. BERT-based conformal predictor for sentiment analysis. In *Conformal and Probabilistic Prediction and Applications*, 269–284. PMLR.

Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J. E.; Kadous, M. W.; and Stoica, I. 2024. RouteLLM: Learning to Route LLMs from Preference Data. In *The Thirteenth International Conference on Learning Representations*.

Pan, J.; Zhang, Y.; Zhang, C.; Liu, Z.; Wang, H.; and Li, H. 2024. DynaThink: Fast or Slow? A Dynamic Decision-Making Framework for Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 14686–14695. Miami, Florida, USA: Association for Computational Linguistics.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2024. Conformal Language Modeling.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

Su, J.; Healey, J.; Nakov, P.; and Cardie, C. 2025. Between Underthinking and Overthinking: An Empirical Study of

Reasoning Length and correctness in LLMs. *arXiv preprint arXiv:2505.00127*.

Su, J.; Luo, J.; Wang, H.; and Cheng, L. 2024. API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 979–995. Miami, Florida, USA: Association for Computational Linguistics.

Sui, Y.; Chuang, Y.-N.; Wang, G.; Zhang, J.; Zhang, T.; Yuan, J.; Liu, H.; Wen, A.; Chen, H.; Hu, X.; et al. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; Wang, Q.; Shi, X.; Xu, K.; Shen, H. T.; and Zhu, X. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6886–6898.

Wordsmiths, T. 2024. stemqa Dataset. https://huggingface.co/datasets/thewordsmiths/stem_mcqa.

Wu, Y.; Wang, Y.; Du, T.; Jegelka, S.; and Wang, Y. 2025. When More is Less: Understanding Chain-of-Thought Length in LLMs. *arXiv preprint arXiv:2502.07266*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025a. Qwen3 Technical Report. arXiv:2505.09388.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025b. Qwen2.5 Technical Report. arXiv:2412.15115.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.; et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Zhou, X.; Chen, B.; Gui, Y.; and Cheng, L. 2025. Conformal prediction: A data perspective. *ACM Computing Survey*.