Don't Miss the Forest for the Trees: In-Depth Confidence Estimation for LLMs via Reasoning over the Answer Space

Ante Wang[◇], **Weizhi Ma**[◇] and **Yang Liu**[♣,◇]
[♣]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[◇]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China

Abstract

Knowing the reliability of a model's response is essential in application. With the strong generation capabilities of LLMs, research has focused on generating verbalized confidence. This is further enhanced by combining chain-of-thought reasoning, which provides logical and transparent estimation. However, how reasoning strategies affect the estimated confidence is still under-explored. In this work, we demonstrate that predicting a verbalized probability distribution can effectively encourage in-depth reasoning for confidence estimation. Intuitively, it requires an LLM to consider all candidates within the answer space instead of basing on a single guess, and to carefully assign confidence scores to meet the requirements of a distribution. This method shows an advantage across different models and various tasks, regardless of whether the answer space is known. Its advantage is maintained even after reinforcement learning, and further analysis shows its reasoning patterns are aligned with human expectations.

1 Introduction

Despite significant progress on various tasks, Large Language Models (LLMs, Achiam et al. 2023; Jiang et al. 2023; Yang et al. 2025; Liu et al. 2024; Guo et al. 2025) still inevitably make errors, preventing their deployment in high-stakes applications such as healthcare, law, and finance (Jiang et al., 2012; Bojarski et al., 2016). A promising solution to this limitation is the development of well-calibrated confidence estimates, which would allow LLMs to differentiate between correct and incorrect answers and produce confidence scores that match their true empirical accuracy. This capability is essential for building trustworthy AI systems. For instance, if an LLM expresses low confidence in a clinical diagnosis, it could automatically trigger a more careful analysis by a human physician.

Traditional calibration methods often rely on the classification probabilities generated by the model (Guo et al., 2017; Kumar et al., 2019). However, these approaches are primarily designed for classification tasks and struggle to generalize to open-ended text generation. Furthermore, research (Tian et al., 2023) demonstrates that modern LLMs, particularly those fine-tuned with Reinforcement Learning from Human Feedback (RLHF, Ouyang et al. 2022), may sacrifice well-calibrated predictions to better adhere to user instructions. To address this, verbalized confidence (Tian et al., 2023; Yang et al.), where a model expresses its uncertainty directly in natural language, has been proposed. Nevertheless, empirical analyses (Xiong et al.; Li et al., 2024) reveal that this method can suffer from overconfidence and its reliability is often dependent on model scale.

Inspired by the success of chain-of-thought reasoning (Wei et al., 2022), recent work (Zhang and Zhang, 2025; Yoon et al., 2025; Damani et al., 2025) has shown that LLMs can reason about their own uncertainty to produce more accurate confidence estimates. These methods prompt the model to analyze its own prediction before providing a confidence score. However, it remains unclear what constitutes effective reasoning content for this purpose. To address this gap, we investigate verbalized probability distribution, which is motivated by the intuition that confidence in a given prediction should be constrained by the probabilities assigned to alternative options, forming a coherent probability distribution. This naturally incentivizes more in-depth reasoning than earlier approaches. By reasoning over the entire space of possible answers, the LLM can contextualize the likelihood of its primary prediction (the "*tree*") by considering all candidate answers (the "*forest*"), thereby mitigating the short-sightedness that leads to overconfidence.

We evaluate our approach on datasets encompassing both multiple-choice tasks (with a known

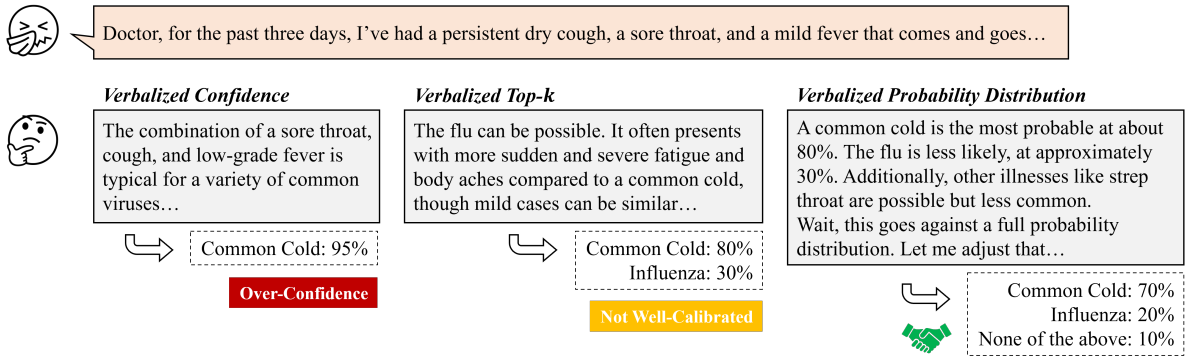

Figure 1: An example for illustrating the difference in the three verbalization-based methods.

answer space) and open-ended tasks (with an unknown answer space) across various LLMs. For multiple-choice, we use MedQA (Jin et al., 2021), MMLU-Pro (Wang et al., 2024c), and MedXpertQA (Zuo et al.). For open-ended generation, we use HotpotQA (Yang et al., 2018) and MedCaseReasoning (Wu et al., 2025). Results indicate that verbalized probability distribution achieves competitive performance on simpler datasets and significantly outperforms baselines on more challenging ones, using only training-free prompting. We further enhance the reasoning process via reinforcement learning (RL), building on prior work (Damani et al., 2025). These experiments show that our method converges faster than baselines and demonstrates robust performance on out-of-domain datasets. Finally, we analyze the reasoning patterns produced by different verbalization-based methods and discuss the limitations of our approach.

2 Preliminary

2.1 Methods

We begin by introducing two conventional yet representative methods that use the model's generated token probabilities as confidence scores: Logit and p(True) (Kadavath et al., 2022).

The **Logit** method calculates confidence from the token probabilities that constitute the predicted answer. Formally, given an input x, a model response $\hat{y}$ is sampled from an LLM. This response comprises a reasoning process r and a predicted answer $\hat{a} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_L)$, where $L = |\hat{a}|$. The confidence score is the product of the conditional probabilities for each token:

$$\text{Conf}_{\text{Logit}} = \prod_{t=1}^{L} P(\hat{a}_t \mid x, r, \hat{a}_{<t}).$$

The **p(True)** method prompts the model to evaluate its own generated response, $\hat{y}$. The confidence score is the probability the model assigns to generating the token "True" in response to the judgment prompt p:

$$\text{Conf}_{\text{p(True)}} = P(\text{"True"} \mid x, \hat{y}, p).$$

This study primarily focuses on verbalization-based approaches, which employ natural language reasoning to articulate confidence values. This paradigm effectively harnesses the reasoning capabilities of the latest LLMs and is applicable to black-box models where token probabilities are inaccessible. Following previous work (Tian et al., 2023), we introduce Verbalized Confidence and Verbalized Top-k.

Verbalized Confidence directly prompts an LLM to reason about both an answer and its confidence score. **Verbalized Top-k**, in contrast, prompts the model to produce k guesses alongside a probability for each. The highest-probability prediction is then selected as the final output. Following Tian et al. (2023) and Tao et al. (2024), we set $k = 2$ across all experiments. Our empirical results (Table 8) demonstrate that increasing k does not lead to substantial improvement.

In this work, we propose **Verbalized Probability Distribution**, a method that prompts LLMs to reason about a probability distribution over the entire space of possible answers, thereby encouraging a comprehensive confidence quantification. While it shares the conceptual similarity of considering alternatives with Verbalized Top-k, it is distinguished by its emphasis on reasoning over the complete answer space instead of a fixed number of guesses. A comparison is shown in Figure 1. For open-ended questions where enumerating all answers is infeasible, we include a "None of the above" option

to aggregate the probability of all low-likelihood candidates. We provide used prompts in the Appendix A.

2.2 Metrics

Following established literature, we evaluate our method using the Area Under the Receiver Operating Characteristic curve (AUROC), the Expected Calibration Error (ECE), and the Brier Score. These metrics provide complementary views on model performance.

The **AUROC** measures a model's ability to discriminate between positive and negative classes, independent of the classification threshold. It is calculated as the area under the plot of the True Positive Rate (TPR) against the False Positive Rate (FPR). For a set of n predictions, it can be computed as:

$$\text{AUROC} = \frac{\sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \mathbb{I}(s_i^+ > s_j^-)}{n_+ \cdot n_-},$$

where s_i^+ and s_j^- are the confidence scores for the i-th positive and j-th negative instance, respectively, and n_+ and n_- are the total number of positive and negative instances.

The **ECE** quantifies calibration, which is the alignment between predicted confidence and empirical accuracy. It is computed by grouping predictions into M bins $(B_1, \ldots, B_M)$ based on their confidence scores:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ denote the accuracy and average confidence within bin B_m. A key limitation of ECE is that it does not assess a model's discriminative power; a model can be perfectly calibrated yet fail to separate the classes.

To provide a holistic assessment, we also report the **Brier Score**, which simultaneously evaluates both calibration and discrimination. It is defined as the mean squared error between the predicted probabilities and the true labels:

$$\text{Brier Score} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2,$$

where $\hat{p}_i$ is the predicted probability for instance i and $y_i \in \{0, 1\}$ is the corresponding true label. A lower Brier score indicates better overall performance.

2.3 Evaluation Setup

Datasets. We focus on tasks that involve identifying the best answer from multiple alternatives, rather than simple true-or-false verification. The medical domain serves as an important and suitable case study. For example, in medical diagnosis, a model must generate a reasonable differential diagnosis to aid in patient care. We also include tasks from other domains to validate the generalization of different methods.

The tasks used in this work can be categorized based on whether the answer space is known. For the known category, we evaluate three multiple-choice tasks requiring reasoning: MedQA (Jin et al., 2021), MMLU-Pro (Wang et al., 2024c), and MedXpertQA (Zuo et al.). The test sets comprise 1,273, 12,032, and 2,450 instances, respectively. MedQA and MedXpertQA are both medical benchmarks, with the latter being notably more difficult. MMLU-Pro offers a comprehensive evaluation across diverse domains, such as physical, business, and law. For the unknown category, we employ HotpotQA (Yang et al., 2018) and MedCaseReasoning (Wu et al., 2025). HotpotQA focuses on multi-hop question answering, while MedCaseReasoning comprises clinical diagnosis cases. We utilize the distractor subset of HotpotQA, from which we sample 1,000 test instances. No supporting facts are provided, because we notice this is too simple for recent LLMs. Regarding MedCaseReasoning, we filter the dataset to include only 644 instances that are solvable by the GPT-5, as the original set contains many cases that are prohibitively difficult for most LLMs. We use accuracy to evaluate performance across all datasets. For tasks with a known search space, we compute accuracy using an exact match. For tasks with an unknown search space, we employ LLM-as-a-Judge (Zheng et al., 2023) to compare the prediction against the ground truth. Evaluation prompts are presented in Appendix B.

Models. Our evaluation encompasses a diverse set of LLMs at various scales. This includes the latest models from the Qwen3 series (Qwen3-4B-Instruct-2507 and Qwen3-30B-A3B-Instruct-2507, Yang et al. 2025), the known Mistral-3.2-24B-Instruct (Jiang et al., 2023), and the medical-specific Baichuan-M2-32B-GPTQ-INT4 (Dou et al., 2025). Furthermore, we tested the more advanced closed-source models GPT-4.1 (Achiam et al., 2023) and DeepSeek-V3-

	MedQA				MMLU-Pro				MedXpertQA			
	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓
Qwen3-4B-Instruct	**0.764**	–	–	–	**0.720**	–	–	–	**0.176**	–	–	–
↪ + Logit	**0.764**	**0.675**	0.223	0.219	**0.720**	0.744	0.261	0.259	**0.176**	0.505	0.810	0.808
↪ + p(True)	**0.764**	0.642	0.215	0.199	**0.720**	0.535	0.334	0.323	**0.176**	0.497	0.534	0.572
↪ + Verbalized Conf.	0.735	0.663	0.238	0.224	0.711	0.764	0.250	0.243	0.160	0.548	0.789	0.808
↪ + Verbalized Top-k	0.749	**0.675**	0.207	0.181	0.716	0.772	0.236	0.230	0.169	0.518	0.733	0.768
↪ + Verbalized Distrib.	0.756	0.672	**0.173**	**0.044**	0.707	**0.773**	**0.174**	**0.091**	0.172	**0.564**	**0.395**	**0.472**
Qwen3-30B-A3B-Instruct	0.869	–	–	–	**0.797**	–	–	–	0.247	–	–	–
↪ + Logit	0.869	0.683	0.123	0.121	**0.797**	0.734	0.191	0.190	0.247	0.508	0.734	0.732
↪ + p(True)	0.869	**0.713**	0.127	0.127	**0.797**	0.684	0.191	0.189	0.247	0.510	0.734	0.736
↪ + Verbalized Conf.	0.866	0.679	0.117	0.088	0.794	0.724	0.179	0.162	**0.258**	0.565	0.672	0.689
↪ + Verbalized Top-k	**0.870**	0.655	**0.112**	**0.072**	0.790	0.755	0.172	0.152	0.252	0.525	0.658	0.683
↪ + Verbalized Distrib.	0.867	0.680	0.118	0.096	0.786	**0.777**	**0.140**	**0.059**	0.251	**0.582**	**0.339**	**0.359**
GPT-4.1	0.936	–	–	–	**0.808**	–	–	–	0.397	–	–	–
↪ + Verbalized Conf.	0.936	0.679	0.059	0.053	0.806	0.715	0.172	0.172	**0.417**	0.551	0.559	0.564
↪ + Verbalized Top-k	0.939	**0.822**	**0.052**	**0.041**	0.796	0.798	0.141	0.092	0.404	0.583	0.399	0.402
↪ + Verbalized Distrib.	**0.940**	0.812	0.056	0.073	0.802	**0.812**	**0.124**	**0.045**	0.400	**0.656**	**0.309**	**0.290**
DeepSeek-V3	0.880	–	–	–	0.770	–	–	–	0.307	–	–	–
↪ + Verbalized Conf.	**0.888**	0.750	**0.089**	**0.029**	**0.788**	0.769	0.162	0.120	**0.313**	0.548	0.542	0.573
↪ + Verbalized Top-k	0.882	0.796	**0.089**	0.044	0.771	0.776	0.151	0.074	0.309	0.550	0.434	0.467
↪ + Verbalized Distrib.	0.883	**0.807**	0.097	0.096	**0.788**	**0.817**	**0.135**	**0.059**	0.297	**0.574**	**0.302**	**0.275**

Table 1: Comparison of training-free approaches on MedQA, MMLU-Pro, and MedXpertQA test sets. We include results of Mistral-3.2-24B-Instruct and Baichuan-M2-32B in Table 10.

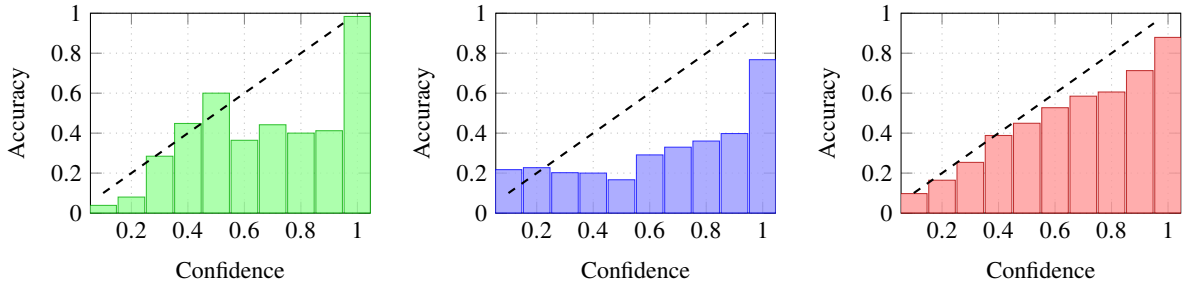

Figure 2: Calibration curves of Qwen3-4B-Instruct on MMLU-Pro when using Verbalized Confidence (Left), Verbalized Top-k (Mid), and Verbalized Probability Distribution (Right).

0324 (Liu et al., 2024). Due to the high cost of querying the GPT-4.1 and DeepSeek-V3-0324 APIs, we evaluated these two models on a random sample of up to 1,000 instances per dataset.

3 Experiments

3.1 Training-Free Evaluation

Given the black-box nature of current LLMs and the high cost of training large-scale models, we first evaluate various methods in a training-free setup, as shown in Tables 1 and 2. Our analysis yields the following observations.

Verbalized Distribution achieves better performance on more challenging tasks Conventional approaches based on token logits exhibit inconsistent performance across datasets and LLMs. For example, while p(True) achieves the best AUROC on MedQA using Qwen3-30B-A3B-Instruct, it performs worst on MMLU-Pro and MedXpertQA. These methods also suffer from over-confidence, often producing confidence scores above 95% even when their ability to distinguish correct predictions (high AUROC) remains intact. Among verbalization-based approaches, Verbalized Confidence displays a similarly high degree of over-confidence as logit-based methods. However, this issue is mitigated when using more capable LLMs, a finding that aligns with prior research. Verbalized Top-k accelerates the mitigation of over-confidence and consistently improves calibration performance. Nonetheless, its effectiveness remains limited for smaller LLMs and more challenging tasks. In contrast, Verbalized Distribution demonstrates significantly better performance on MMLU-Pro and MedXpertQA across all LLMs. On the less challenging MedQA, it remains competitive, achieving top-tier results, suggesting that simpler tasks may not require its more complex reasoning. Despite a slight decrease in some accuracy results, most demonstrate competitive performance. Crucially, **smaller LLMs gain more benefits from**

	HotpotQA				MedCaseReasoning			
	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓
Qwen3-4B-Instruct	0.303	–	–	–	0.298	–	–	–
↪ + Logit	0.303	0.586	0.639	0.647	0.298	0.492	0.519	0.513
↪ + p(True)	0.303	**0.733**	0.487	0.490	0.298	**0.651**	0.661	0.665
↪ + Verbalized Conf.	0.290	0.642	0.632	0.647	**0.315**	0.647	0.585	0.612
↪ + Verbalized Top-k	0.319	0.656	0.550	0.572	0.301	0.623	0.527	0.568
↪ + Verbalized Distrib.	**0.336**	0.714	**0.336**	**0.322**	0.311	0.648	**0.330**	**0.340**
Qwen3-30B-A3B-Instruct	0.437	–	–	–	0.404	–	–	–
↪ + Logit	0.437	0.596	0.485	0.486	0.404	0.567	0.381	0.338
↪ + p(True)	0.437	**0.702**	0.460	0.462	0.404	0.595	0.561	0.563
↪ + Verbalized Conf.	0.426	0.664	0.518	0.526	0.443	0.587	0.506	0.512
↪ + Verbalized Top-k	**0.439**	0.684	0.463	0.474	**0.449**	0.583	0.472	0.480
↪ + Verbalized Distrib.	0.438	0.689	**0.318**	**0.295**	0.429	**0.673**	**0.289**	**0.252**
GPT-4.1	**0.686**	–	–	–	0.644	–	–	–
↪ + Verbalized Conf.	0.679	0.728	0.291	0.290	**0.665**	0.629	0.284	0.263
↪ + Verbalized Top-k	0.673	0.755	0.232	0.206	0.623	0.654	0.248	0.165
↪ + Verbalized Distrib.	0.682	**0.772**	**0.208**	**0.170**	0.655	**0.663**	**0.211**	**0.030**
DeepSeek-V3	0.583	–	–	–	0.568	–	–	–
↪ + Verbalized Conf.	0.591	0.750	0.293	0.287	**0.575**	0.673	0.308	0.283
↪ + Verbalized Top-k	0.572	0.698	0.276	0.241	0.543	0.644	0.292	0.242
↪ + Verbalized Distrib.	**0.594**	**0.777**	**0.203**	**0.116**	0.551	**0.684**	**0.227**	**0.071**

Table 2: Comparison of training-free approaches on HotpotQA and MedCaseReasoning test set.

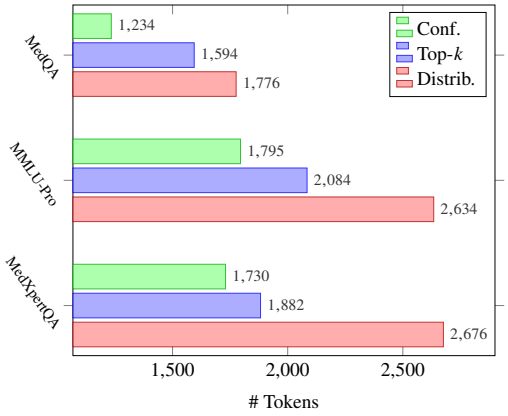

Figure 3: Averaged token consumption of different verbalization-based approaches across MedQA, MMLU-Pro, and MedXpertQA test set when using Qwen3-4B-Instruct.

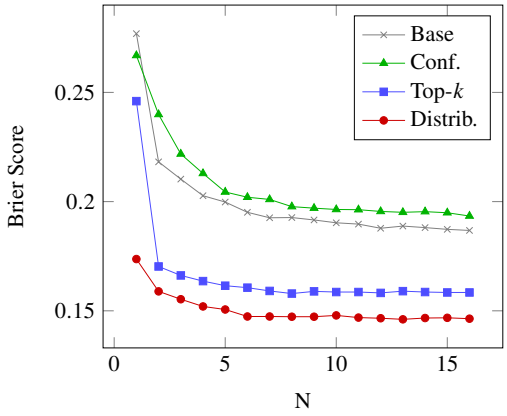

Figure 4: Brier scores of different methods combined with answer consistency on MMLU-Pro test set.

reasoning verbalized distribution, highlighting the practical utility of this method. These conclusions are consistent for tasks without a known search space, as shown in Table 2. The calibration curve in Figure 2 further illustrates the strong alignment between predicted confidence and empirical accuracy.

Verbalized Distribution stimulated deeper reasoning in confidence estimation Figure 3 plots the token consumption of different methods. We observe that Verbalized Top-k consistently consumes more tokens than Verbalized Confidence, while Verbalized Distribution has the highest cost of the three. This validates that the reasoning process is significantly influenced by user prompts and that deeper thinking can improve performance. Using a verbalized distribution appears to stimulate deeper reasoning more effectively. This finding raises an interesting question: can a longer prompt with more fine-grained guidance lead to even better performance? To investigate this, we implemented the long prompt provided in (Damani et al., 2025). The results, shown in Table 11, demonstrate that **the performance gains stem from the requirement to predict a distribution, rather than from more complex instructions.**

Verbalized Distribution can be further improved through repetitive sampling Leveraging answer consistency across multiple reasoning trajectories is an effective way to obtain a more accurate confidence score, albeit at a significantly higher computational cost (Wang et al., 2024a;

	MedQA				MMLU-Pro				MedXpertQA			
	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
RL from the Qwen3-4B-Instruct Model												
RLVR	0.782	–	–	–	**0.719**	–	–	–	0.192	–	–	–
RLCR + Conf.	0.788	0.672	0.160	0.091	0.713	0.668	0.214	0.174	**0.207**	0.532	0.646	0.688
RLCR + Top-k	0.781	0.762	0.159	0.046	0.713	**0.801**	0.191	0.155	0.194	0.534	0.499	0.581
RLCR + Distrib.	**0.795**	**0.763**	**0.141**	**0.038**	0.715	0.800	**0.159**	**0.052**	0.204	**0.551**	**0.384**	**0.451**
Zero-RL from the Qwen3-Base Model												
RLVR	0.765	–	–	–	0.646	–	–	–	**0.170**	–	–	–
RLCR + Conf.	**0.775**	**0.725**	**0.155**	0.047	**0.650**	0.746	0.225	0.174	0.162	0.544	0.508	0.607
RLCR + Top-k	0.747	0.722	0.169	0.058	0.639	0.750	0.199	0.084	0.160	0.528	0.353	0.458
RLCR + Distrib.	0.758	0.712	0.165	**0.041**	0.636	**0.757**	**0.188**	**0.027**	0.166	**0.569**	**0.322**	**0.407**

Table 3: Comparison of RL training-based approaches on MedQA, MMLU-Pro, and MedXpertQA test set.

	HotpotQA				MedCaseReasoning			
	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
RL using HotpotQA training set								
RLVR	0.355	–	–	–	**0.324**	–	–	–
RLCR + Conf.	0.345	0.838	0.156	0.054	0.322	0.646	0.421	0.454
RLCR + Top-k	**0.362**	0.827	0.160	0.045	0.312	**0.677**	0.265	0.260
RLCR + Distrib.	0.358	**0.851**	**0.147**	**0.029**	0.318	0.644	**0.230**	**0.156**
RL using MedCaseReasoning training set								
RLVR	0.299	–	–	–	0.359	–	–	–
RLCR + Conf.	0.282	0.616	0.558	0.582	**0.370**	0.619	0.359	0.355
RLCR + Top-k	**0.337**	0.705	0.436	0.475	0.345	0.646	0.279	0.257
RLCR + Distrib.	0.334	**0.728**	**0.279**	**0.244**	0.357	**0.658**	**0.227**	**0.104**

Table 4: Comparison of RL training-based approaches on HotpotQA and MedCaseReasoning test set.

Xiong et al.). We investigate how verbalization-based approaches can be combined with this technique when more computational resources are allowed. As shown in Figure 4, we sample at most 16 reasoning trajectories using a temperature of 0.8. We follow Wang et al. (2024a) in using answer frequency as the confidence score for our baseline. For the verbalization-based approach, the predicted confidence scores are used as weights in a weighted aggregation (Xiong et al.). Specifically, for each unique answer, we sum its corresponding confidence scores across all predictions. The answer with the highest aggregate score is selected, and its final confidence score is obtained by normalizing this sum. Results show that Verbalized Confidence only has competitive performance with the Baseline, while both Verbalized Top-k and Verbalized Distribution benefit further from multiple sampling. However, Verbalized Distribution demonstrates better sample efficiency, achieving a lower Brier Score with a much smaller sample size N.

3.2 RL Training-Based Evaluation

Recent advances have demonstrated the efficacy of RL in refining reasoning processes, including improved confidence estimation (Tao et al., 2024; Damani et al., 2025). In this section, we investigate whether RL can further enhance confidence estimation by optimizing a model's verbalized probability distribution.

In this work, we follow RLCR (Damani et al., 2025) to implement RL for all verbalization-based methods, utilizing a reward function that incorporates both answer accuracy and the Brier Score:

$$r = y - (\hat{p} - y)^2,$$

where $\hat{p}$ is the predicted confidence and y is the answer correctness.

Additionally, if the output has formatting issues—for example, if it does not adhere to the defined structure—we assign a score of $r = -1$. This is particularly relevant for verbalized distribution predictions, where the confidence scores for all candidates must sum to 1 to constitute a valid probability distribution.

We adopt GRPO (Shao et al., 2024) as our RL algorithm. Following (Damani et al., 2025; Bereket and Leskovec), we remove the division by the standard deviation in the advantage calculation, as this has been shown to mitigate poor calibration. Our training is based on the prevalent veRL framework (Sheng et al., 2024). During the rollout phase,

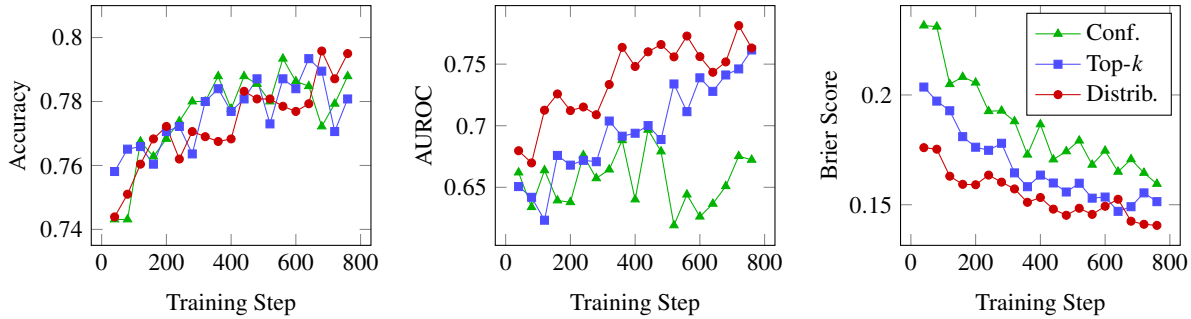

Figure 5: Comparison of different methods on MedQA test set across different training steps when using Qwen3-4B-Instruct during RL training.

	In-Domain (Medical)				**Out-of-Domain**									
	Hlth	Bio	Chem	Psych	Bus	CS	Econ	Eng	Hist	Law	Math	Oth	Phil	Phys
Prob.	0.318	0.130	0.127	0.302	0.172	0.215	0.200	0.240	0.447	0.580	0.072	0.387	0.400	0.145
RL + Prob.	0.260	0.136	**0.111**	0.239	0.154	0.207	0.186	0.199	0.341	0.512	0.071	0.321	0.299	0.124
Top-k	0.289	0.142	0.127	0.252	0.175	0.210	0.196	0.256	0.393	0.559	0.074	0.343	0.339	0.134
RL + Top-k	0.230	0.121	0.114	0.199	0.143	**0.166**	0.164	0.218	0.316	0.408	**0.064**	0.270	0.288	0.118
Distrib.	0.211	0.122	0.123	0.195	0.131	0.192	0.160	0.190	0.271	0.322	0.066	0.240	0.240	0.118
RL + Distrib.	**0.190**	**0.116**	0.117	**0.168**	**0.128**	0.171	**0.141**	**0.172**	**0.231**	**0.297**	0.067	**0.211**	**0.224**	**0.110**

Table 5: Brier Score performance on In-Domain (Medical) vs Out-of-Domain tasks from the MMLU-Pro test set.

we sample 512 instances, generating 8 rollouts for each with a temperature of 1. For the policy update phase, we use 128 instances and their corresponding rollouts, with a learning rate of 1×10^{-6}. This configuration results in 4 training steps per rollout step, balancing the efficiency of the rollouts with the preservation of the on-policy nature of the algorithm.

We mainly experiment on Qwen3-4B-Instruct and also attempt Qwen3-8B-Base which has not gone through instructing tuning. Intuitively, Qwen3-4B-Instruct has grasp reasoning skills for answer prediction, which can benefit the confidence estimation. In contrast, Qwen3-8B-Base might explore relatively more blindly by depending on knowledge learnt from pretraining. Results are shown in Table 3 and 4. We observe the following conclusions.

The advantage of predicting verbalized distribution is preserved after RL training Through extensive experience gained during RL, an LLM may learn an optimal strategy that works for various prompts. We observe that the performance improves significantly for all methods, and the gap between different verbalization-based methods narrows. However, the overall trends remain consistent with the training-free evaluation. The Verbalized Distribution method maintains its advantage across different tasks, especially the more challenging ones. This is likely due to its better initialization, which results in faster and smoother convergence, as shown in Figure 11. This demonstrates that **the choice of prompt used for training confidence estimation is crucial**. It appears difficult for an LLM to independently discover the optimal reasoning strategy through RL alone. An interesting question for future work is whether larger LLMs with stronger exploration capabilities could mitigate this issue, though testing this would require computational resources beyond our current means.

Training with RL improves performance across different domains Research (Chu et al.) has demonstrated that RL is more effective than supervised fine-tuning at preserving a model's cross-domain capabilities. Echoing previous work (Damani et al., 2025), we again validate RL's strong cross-domain performance on confidence estimation tasks, regardless of the prompts used. This is most apparent when comparing results in Table 2 and 4: although RL does not significantly improve out-of-domain task accuracy, it consistently enhances the quality of confidence estimation. Using the broader range of topics in MMLU-Pro, we further examine performance across diverse domains in Table 5. Results show unique advantages of Verbalized Distribution on both in-domain and out-of-domain tasks, which again supports the conclusion. Interestingly, **the selection of the training**

	ES	UA	LC	Avg.
Verbalized Conf.	4.25	2.37	2.26	3.08
Verbalized Top-k	4.00	2.92	1.67	2.86
Verbalized Distrib.	**4.57**	**3.38**	**3.39**	**3.78**
Verbalized Conf.	2.73	3.35	2.71	2.93
Verbalized Top-k	2.59	3.25	2.16	2.67
Verbalized Distrib.	**3.47**	**4.23**	**3.58**	**3.76**

Table 6: Evaluation results across different methods on 200 cases sampled from the MMLU-Pro (Up) and HotpotQA (Down) test set. Metric abbreviations: ES (Evidential Strength), UA (Uncertainty Awareness), LC (Logical Calibration).

set is crucial for out-of-domain performance. As shown in Table 4, training on HotpotQA effectively improves results on MedCaseReasoning, but the reverse is not true. This may be due to the different reasoning complexities of these tasks, which results in varying transfer capabilities across domains.

3.3 Analyses & Discussion

This section provides a deeper analysis of how Verbalized Distribution incentives foster better reasoning strategies for confidence estimation and discusses the limitations of this technique.

Reasoning patterns for confidence estimation While we have demonstrated that Verbalized Distribution elicit deeper reasoning and improve confidence scores, the specific mechanisms behind this improvement remain unclear. To address this, we analyze the reasoning traces to determine whether the LLMs exhibit effective reasoning behaviors that align with human expectations. We summarize three criteria indicative of high-quality reasoning:

- *Evidential Strength*: The clarity and directness of the facts and logic used to support the answer.
- *Uncertainty Awareness*: The active identification of unknowns, assumptions, potential flaws, and alternative answers within the reasoning process.
- *Logical Calibration*: The alignment between the confidence score and the underlying balance of evidence and limitations, leading to reliable estimates.

Based on these criteria, we employed GPT-5o-mini to rate reasoning traces generated by different verbalization-based methods on a scale of 1 to 5. The evaluation prompt is shown in Figure 10. We activated the model's thinking mode to obtain more reliable ratings. The results, presented in Table 6, indicate that while all methods enumerate evidence for their determinations, the Verbalized Confidence approach sometimes fails to acknowledge its own limitations, leading to lower scores in uncertainty awareness. In contrast, the Verbalized Distribution method produces more logically sound reasoning for deriving confidence scores, earning it the highest rating for logical calibration. A case study in Figure 12 clearly illustrates that predicting verbalized distributions activates these in-depth reasoning behaviors.

	ACC↑	AUROC↑	Brier↓	ECE↓
Verbalized Conf.	0.734	0.741	0.240	0.239
↪ + RLCR	0.737	0.776	0.225	0.229
Verbalized Top-k	0.684	0.774	0.271	0.281
↪ + RLCR	0.729	**0.786**	**0.196**	**0.201**
Verbalized Distrib.	0.670	0.653	0.298	0.305
↪ + RLCR	**0.740**	0.775	0.201	0.205

Table 7: Test results of Qwen3-4B-Instruct on OlympiadBench.

Limitations of the Verbalized Distribution method The premise of predicting a distribution is that the LLM is aware of multiple plausible candidates, and uncertainty can arise from their indistinguishability. However, for certain tasks, predicting multiple answers is inherently difficult. Math reasoning is a typical example: LLMs reason step-by-step to reach a final answer, where the correctness of each step is usually absolute, leaving little room for alternative outcomes. This contrasts sharply with more open-ended tasks like question answering, where multiple answers can be plausible for an LLM (e.g., "What was Michael Jackson's first album?"). To validate this, we conducted an experiment on the textual modality portion of the OlympiadBench (He et al., 2024) using Qwen3-4B-Instruct. The correctness of an answer is verified with `Math-Verify`[1]. Following prior work (Zeng et al., 2025), we adopt the MATH dataset (Hendrycks et al.) for RL training. Due to the model's already strong performance on MATH, to ensure training efficacy on challenging problems, we filter the dataset to include only instances where the model fails in at least one of four sampled re-

[1] https://github.com/huggingface/Math-Verify

sponses (Team et al., 2025). This yields a final set of 2,359 training instances. Results in Table 7 show that Verbalized Distribution suffer a significant performance drop compared to Verbalized Confidence. However, we demonstrate that RL can largely recover this performance. Another promising approach is to automatically select the proper prompting strategy for different tasks, which we leave for future work.

4 Related Work

Early research on model calibration primarily focused on traditional classification tasks, often by adjusting the logits from the classification head (Gupta et al.). Pioneering techniques such as Histogram Binning (Zadrozny and Elkan, 2001) and Platt Scaling (Platt et al., 1999) were developed for binary classification. A widely adopted extension for multi-class settings is Temperature Scaling (Guo et al., 2017), which uses a single scalar parameter to smoothly adjust the softmax distribution.

With the advent of LLMs, the research paradigm has shifted toward confidence estimation through prompting. For instance, Kadavath et al. (2022) demonstrated that LLMs can produce well-calibrated confidence scores by prompting them for a true/false determination. In response to the black-box nature of many LLM services, where token probabilities and internal representations are inaccessible, verbalization-based approaches have emerged as an effective solution. For example, Tian et al. (2023) found that having the model generate its confidence in natural language yields better calibration. A similar technique involves generating linguistic expressions (e.g., "highly likely") and has shown comparable performance (Lin et al.; Tian et al., 2023). Xiong et al. further enhanced this method by aggregating scores from repetitive sampling. The study most closely related to our work is (Wang et al., 2024b), which was the first to investigate verbalized probability distributions. However, its performance relative to other verbalization-based methods remains unclear.

The aforementioned studies have been predominantly evaluated on simple classification or question-answering tasks. Recent work has extended this inquiry to more complex domains, finding that chain-of-thought reasoning can improve confidence scores (Yoon et al., 2025; Devic et al., 2025; Mei et al., 2025). Unlike previous work (Azaria and Mitchell, 2023) that used an external classifier to predict confidence scores from intermediate states, this approach is also transparent, clearly showing how the score is estimated. Notably, Damani et al. (2025) trained LLMs using RL to generate reasoning processes specifically for confidence estimation. Despite these advances, the characteristics of an effective rationale remain undefined. Our work partially addresses this gap by comparing verbalization-based methods, highlighting the critical role of probability distribution prediction.

Several concurrent studies are also remotely relevant: Zhang et al. (2025) proposed verbalized sampling to generate diverse responses, while Li et al. (2025) used verbalized probability distributions for preference rating. In contrast, our work focuses specifically on confidence estimation across a variety of tasks, positioning it as a complementary contribution to this field.

5 Conclusion

This work investigates confidence estimation through reasoning using verbalized probability distribution prompting. Our results show that this method incentivizes more in-depth analysis and produces better confidence scores, which is particularly beneficial for smaller LLMs and challenging tasks. These gains are maintained under RL training. Further analysis indicates that predicting verbalized probability distributions produces reasoning patterns more closely aligned with human expectations. However, we also find that prompt effectiveness varies by task, suggesting that future research should focus on automatic prompt selection or optimization for robust and generalizable confidence estimation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. In *LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration*.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, and 1 others. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Siddartha Devic, Charlotte Peale, Arwen Bradley, Sinead Williamson, Preetum Nakkiran, and Aravind Gollakota. 2025. Trace length is a simple uncertainty signal in reasoning models. *arXiv preprint arXiv:2510.10409*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. 2012. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *CoRR*.

Zhuohang Li, Xiaowei Li, Chengyu Huang, Guowang Li, Katayoon Goshvadi, Bo Dai, Dale Schuurmans, Paul Zhou, Hamid Palangi, Yiwen Song, and 1 others. 2025. Judging with confidence: Calibrating autoraters to preference distributions. *arXiv preprint arXiv:2510.00263*.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. 2025. Reasoning about uncertainty: Do reasoning models know when they don't know? *arXiv preprint arXiv:2506.18183*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

John Platt and 1 others. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5984–5996.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D 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.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. 2024a. Self-consistency boosts calibration for math reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6023–6029.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. 2024b. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. 2025. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. 2025. Reasoning models better express their confidence. *arXiv preprint arXiv:2505.14489*.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Boxuan Zhang and Ruqi Zhang. 2025. Cot-uq: Improving response-wise uncertainty quantification in llms with chain-of-thought. *arXiv preprint arXiv:2502.17214*.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R Tomz, Christopher D Manning, and Weiyan Shi. 2025. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *arXiv preprint arXiv:2510.01171*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*.

Prompt for Verbalized Confidence

Question: `[Question]`

Reason step-by-step to formulate your final answer. Your answer must be a single entity, a short phrase, or a yes/no. Then, reason about the confidence in your answer. Conclude by providing a JSON object that states the final answer and your estimated confidence in it:

```
{
 "final_answer": "Your final answer",
 "confidence": "0-1"
}
```

Figure 6: The prompt for the Verbalized Confidence method for HotpotQA.

Prompt for Verbalized Top-k

Question: `[Question]`

Reason step-by-step to formulate 2 best guesses and probability that each is correct. Each answer must be a single entity, a short phrase, or a yes/no. Your final output must be a JSON array:

```
[
{
 "candidate": "first most likely answer",
 "confidence": "0-1"
},
{
 "candidate": "second most likely answer",
 "confidence": "0-1"
}
]
```

Figure 7: The prompt for the Verbalized Top-k method for HotpotQA.

A Prompts for Verbalization-Based Methods

We summarize the instructions for verbalization-based methods: Verbalized Confidence (Figure 6), Verbalized Top-k (Figure 7), and Verbalized Probability Distribution (Figure 8).

B Evaluation Prompts

This section summarizes the evaluation prompts, which comprise a prompt for judging the correctness of a prediction against the gold answer (Figure 9) and a prompt for scoring reasoning patterns (Figure 10).

Prompt for Verbalized Probability Distribution

Question: `[Question]`

Reason step-by-step to formulate your answer. You may propose multiple possible answers (fewer than five). Each answer must be a single entity, a short phrase, or a yes/no. Always include "None of the above" as a possible answer. Reason about the confidence in each possible answer. Your final output must be a JSON array where the confidence scores form a probability distribution (they must sum to 1.0):

```
[
  {
    "candidate": "Candidate 1",
    "confidence": "0-1"
  },
  {
    "candidate": "Candidate 2",
    "confidence": "0-1"
  },
  ...
  {
    "candidate": "None of the above",
    "confidence": "0-1"
  }
]
```

Figure 8: The prompt for the Verbalized Probability Distribution method for HotpotQA.

Evaluation Prompt for Verification

Question: `[Question]`
Ground Truth: `[Gold_Answer]`
Prediction: `[Predicted_Answer]`
Is the prediction correct? You must answer "yes" or "no" only.

Figure 9: The prompt for evaluating the correctness of a predicted answer for HotpotQA.

C The Choice of k for Verbalized Top-k

Following previous work (Tian et al., 2023; Tao et al., 2024), we adopt $k = 2$ in the main text. Tian et al. (2023) experimented with $k = 2$ and 4, showing that increasing k does not consistently improve performance. To further validate this finding, we experiment with different values of k on MMLU-Pro, as shown in Table 8. Our results demonstrate that different values of k yield competitive performance.

Evaluation Prompt for Reasoning Patterns

You will be provided with three reasoning processes (Method A, B, and C) that each culminate in a final answer and a **confidence score**. Your task is to **evaluate the quality of the confidence estimation process itself**, independent of whether the final answer is correct.

Scoring Instructions:

For each method, you must assign a score from 1 to 5 (where 1 is Poor and 5 is Excellent) for the following three criteria:

- **Evidential Strength**: The clarity and directness of the facts and logic used to support the answer.
- **Uncertainty Awareness**: The active identification of unknowns, assumptions, potential flaws, and alternative answers within the reasoning process.
- **Logical Calibration**: The alignment between the confidence score and the underlying balance of evidence and limitations, leading to reliable estimates.

Output Format:

First analyse each reasoning process carefully and finally provide a single, valid JSON dictionary. Use the following structure:

```
{
  "Method A": {
    "Evidential Strength": <score>,
    "Uncertainty Awareness": <score>,
    "Logical Calibration": <score>
  },
  "Method B": {
    "Evidential Strength": <score>,
    "Uncertainty Awareness": <score>,
    "Logical Calibration": <score>
  },
  "Method C": {
    "Evidential Strength": <score>,
    "Uncertainty Awareness": <score>,
    "Logical Calibration": <score>
  }
}
```

Remember to base your scores solely on the process of estimating confidence, not on the accuracy of the final answer.

Figure 10: The prompt for evaluating the quality of a reasoning process from different perspectives.

	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
Qwen3-4B-Instruct	0.720	–	–	–
↪ + Verbalized Top-2	0.716	0.772	0.236	0.230
↪ + Verbalized Top-4	0.712	0.743	0.240	0.233
↪ + Verbalized Top-x	0.695	0.749	0.231	0.216

Table 8: Results of Verbalized Top-k when using different k on MMLU test set. x is the exact option number of the corresponding question.

	MedQA	MMLU-Pro	MedXpertQA
Verbalized Distrib.	0.387	0.452	1.190
↪ + RLCR	0.347	0.435	1.124

Table 9: Multi-Label Brier Scores of Qwen3-4B-Instruct on MedQA, MMLU-Pro, MedXpertQA test sets.

D The Quality of the Full Probability Distribution

In the main text, we follow common practice by evaluating models based on a single final answer derived from the Verbalized Probability Distribution. Our RL reward function is also designed to align with this objective. In this section, we further assess the quality of the entire predicted distribution using the Multi-Brier score (ranging from 0 to 2). A lower score denotes better performance. As shown in Table 9, RL improves the calibration of the full distribution, not just the probability assigned to the final answer.

E Training-Free Experiments on More Models

Table 10 presents additional experiments on Mistral-3.2-24B-Instruct and Baichuan-M2-32B-GPTQ-INT4. The results align with the findings in the main text, showing that Verbalized Probability Distribution method usually ranks as the best or runner-up across confidence estimation metrics.

F Experiment Results with a More Complex Prompt

Table 11 presents the results obtained using a more complex prompt from (Damani et al., 2025). This prompt incorporates fine-grained guidance, such as considering alternative approaches and enumerating plausible uncertainties. Contrary to what might be expected, this complexity did not yield a performance improvement. This finding reinforces that our performance gain is attributable to the requirement for probability distribution prediction.

G Case Study

Figure 12 presents a case from the MMLU-Pro test set, demonstrating that the Verbalized Probability Distribution provides more in-depth confidence analyses. This method carefully evaluates the confidence of each option and assigns probabilities through computation and verification. This rigorous process likely leads to the improved final confidence scores.

	MedQA				MMLU-Pro				MedXpertQA			
	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓
Mistral-3.2-24B-Instruct	**0.789**	–	–	–	**0.666**	–	–	–	**0.200**	–	–	–
↪ + Logit	**0.789**	0.697	0.194	0.196	**0.666**	0.658	0.299	0.294	**0.200**	0.526	0.753	0.763
↪ + p(True)	**0.789**	**0.777**	**0.147**	0.083	**0.666**	**0.735**	**0.196**	**0.092**	**0.200**	0.539	0.520	0.576
↪ + Verbalized Conf.	0.780	0.666	0.165	0.099	0.648	0.700	0.263	0.242	0.181	0.536	0.640	0.695
↪ + Verbalized Top-k	0.780	0.705	0.158	**0.047**	0.657	0.721	0.229	0.180	0.186	**0.556**	0.514	0.600
↪ + Verbalized Distrib.	0.785	0.699	0.155	0.052	0.659	0.723	0.209	0.129	**0.200**	0.549	**0.390**	**0.444**
Baichuan-M2-32B-GPTQ-INT4	0.823	–	–	–	**0.709**	–	–	–	0.207	–	–	–
↪ + Logit	0.823	0.665	0.176	0.176	**0.709**	0.560	0.287	0.288	0.207	0.481	0.788	0.789
↪ + p(True)	0.823	0.644	0.140	0.078	**0.709**	0.520	0.248	0.151	0.207	0.459	0.570	0.617
↪ + Verbalized Conf.	**0.829**	0.656	**0.130**	0.043	0.701	**0.739**	0.203	0.147	**0.211**	0.516	0.608	0.660
↪ + Verbalized Top-k	0.791	**0.704**	0.151	**0.029**	0.700	0.720	0.210	0.156	0.187	0.509	0.526	0.607
↪ + Verbalized Distrib.	0.814	0.700	0.161	0.148	0.698	0.734	**0.185**	**0.070**	0.207	**0.539**	**0.303**	**0.345**

Table 10: Comparison of training-free approaches on MedQA, MMLU-Pro, and MedXpertQA test set.

	HotpotQA				MedCaseReasoning			
	ACC↑	AUROC↑	Brier↓	ECE↓	ACC↑	AUROC↑	Brier↓	ECE↓
Qwen3-4B-Instruct	0.297	0.739	0.433	0.455	0.283	0.656	0.544	0.588
↪ + RLCR	0.332	0.799	0.168	0.051	0.310	0.665	0.319	0.339
Qwen3-30B-A3B-Instruct	0.433	0.659	0.455	0.464	0.396	0.612	0.489	0.498
GPT-4.1	0.665	0.745	0.297	0.301	0.596	0.647	0.354	0.349
DeepSeek-V3-0324	0.558	0.744	0.285	0.268	0.523	0.624	0.358	0.343

Table 11: Results on HotpotQA and MedCaseReasoning test set using the long prompt from (Damani et al., 2025).

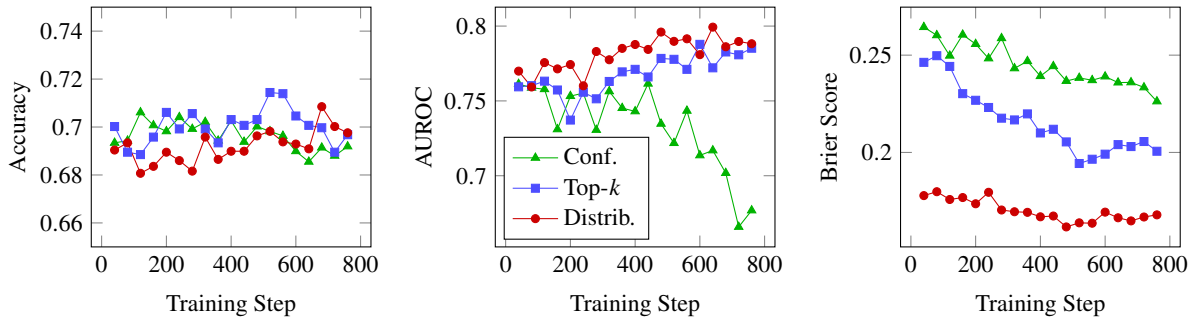

Figure 11: Comparison of different metrics on randomly sampled 2,048 MMLU-Pro test set across different training steps when using Qwen3-4B-Instruct during RL training.

Case study on MMLU-Pro

Question: Typical advertising regulatory bodies suggest, for example that adverts must not: encourage _____, cause unnecessary _____ or _____, and must not cause _____ offence.
Options: A. Safe practices, Fear, Jealousy, Trivial B. Unsafe practices, Distress, Joy, Trivial C. Safe practices, Wants, Jealousy, Trivial D. Safe practices, Distress, Fear, Trivial E. Unsafe practices, Wants, Jealousy, Serious F. Safe practices, Distress, Jealousy, Serious G. Safe practices, Wants, Fear, Serious H. Unsafe practices, Wants, Fear, Trivial I. Unsafe practices, Distress, Fear, Serious

---

Verbalized Confidence:
Let's solve this step by step. ... The reasoning is based on standard advertising regulation principles. The combination of "unsafe practices", "distress", "fear", and "serious offence" is well-documented in regulatory guidelines. No other option fits all parts correctly. Therefore, I am highly confident.
Final Answer:

```
{"final_answer": "I", "confidence": 0.98}
```

Verbalized Top-k:
Let's analyze the question step by step: ... Final decision: Best: I Second: H (with lower confidence due to "wants") But "wants" is not typically regulated. After re-evaluation, the only option that makes sense is I. Thus, second best might be B — but "joy" is not a negative. No. Therefore, I conclude: Best: I Second: None strong — but if forced, perhaps H. But let's go with: - I: 0.95 - H: 0.35 (due to "wants" being out of place) Alternatively, perhaps the original intended answer is I.
Final Answer:

```
[
  {"option": "I", "confidence": 0.95},
  {"option": "H", "confidence": 0.35}
]
```

Verbalized Probability Distribution:
Let's analyze the question step by step: ... 1. I → Strongly correct → High confidence 2. E → Close, but "wants" is not a standard concern → Medium confidence 3. H → "wants" is problematic → Medium-low confidence 4. B → "joy" is wrong → Low confidence 5. D, F, G, C, A → contain "safe practices" → Wrong direction → Very low confidence 6. All others → flawed in key points → Very low
Step 5: Assign Confidence Scores (must sum to 1.0) I: 0.60 E: 0.15 H: 0.10 B: 0.05 Others (A, C, D, F, G): 0.00 (too wrong) Wait — we must assign to all options. Total: 1.0 Better distribution: I: 0.60 E: 0.15 H: 0.10 B: 0.05 A, C, D, F, G: 0.05 each → sum = $0.05 \times 5 = 0.25$ → total = $0.60 + 0.15 + 0.10 + 0.05 + 0.25 = 1.15$ → too high. Adjust: I: 0.60 E: 0.15 H: 0.10 B: 0.05 A, C, D, F, G: 0.05 each → $0.05 \times 5 = 0.25$ → total = $0.60 + 0.15 + 0.10 + 0.05 + 0.25 = 1.15$ → still over. So reduce the others: Better: I: 0.60 E: 0.15 H: 0.10 B: 0.05 A: 0.02 C: 0.02 D: 0.02 F: 0.02 G: 0.02 Sum: $0.60 + 0.15 = 0.75$; $+0.10 = 0.85$; $+0.05 = 0.90$; $+0.02 \times 5 = 0.10$ → total = 1.00. Yes.
Final Confidence Scores: A: 0.02 B: 0.05 C: 0.02 D: 0.02 E: 0.15 F: 0.02 G: 0.02 H: 0.10 I: 0.60 Wait — H is better than B. Let's reassess. Actually, H has "unsafe practices, wants, fear, trivial" — "wants" is weak, "trivial" is weaker than "serious". E has "serious offence" — better than trivial. So I is best. H: "fear" is good, "wants" is bad → lower than E. So better to give: I: 0.60 E: 0.15 H: 0.10 B: 0.05 A: 0.02 C: 0.02 D: 0.02 F: 0.02 G: 0.02 Sum: $0.60 + 0.15 + 0.10 + 0.05 + 0.02 \times 5 = 0.60 + 0.15 + 0.10 + 0.05 + 0.10 = 1.00$ Yes.
Final Answer:

```
[
  {"option": "A", "confidence": "0.02"},
  {"option": "B", "confidence": "0.05"},
  {"option": "C", "confidence": "0.02"},
  {"option": "D", "confidence": "0.02"},
  {"option": "E", "confidence": "0.15"},
  {"option": "F", "confidence": "0.02"},
  {"option": "G", "confidence": "0.02"},
  {"option": "H", "confidence": "0.10"},
  {"option": "I", "confidence": "0.60"}
]
```

Figure 12: A test case from the MMLU-Pro test set, comparing different verbalization-based methods.