

Stimulating In-Depth Confidence Estimation for LLMs through Reasoning over the Answer Space

Anonymous ACL submission

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 it promotes richer reasoning patterns, leading to better estimation.

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 by considering all candidate answers, thereby mitigating the short-sightedness that leads to overconfidence.

We evaluate our approach on datasets encom-

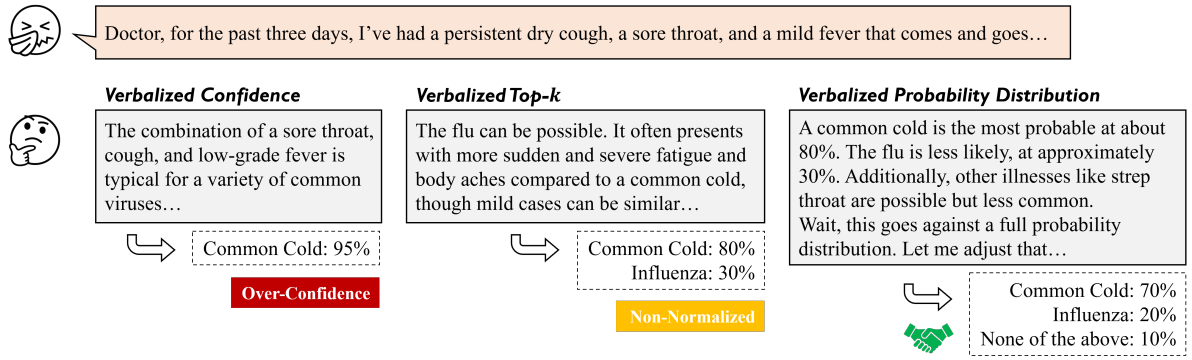


Figure 1: An example for illustrating the difference in the three verbalization-based methods.

084 passing both multiple-choice tasks (with a closed-
 085 set answer space) and open-ended tasks (with
 086 an open-set answer space) across various LLMs.
 087 For multiple-choice, we use MedQA (Jin et al.,
 088 2021), MMLU-Pro (Wang et al., 2024c), and
 089 MedXpertQA (Zuo et al.). For open-ended gen-
 090 eration, we use HotpotQA (Yang et al., 2018)
 091 and MedCaseReasoning (Wu et al., 2025). Re-
 092 sults indicate that verbalized probability distribu-
 093 tion achieves competitive performance on simpler
 094 datasets and significantly outperforms baselines
 095 on more challenging ones, using only training-
 096 free prompting. We further enhance the reasoning
 097 process via reinforcement learning (RL), building
 098 on prior work (Damani et al., 2025). These ex-
 099 periments show that our method converges faster
 100 than baselines and demonstrates robust perfor-
 101 mance on out-of-domain datasets. Finally, we an-
 102 alyze the reasoning patterns produced by different
 103 verbalization-based methods and discuss the limi-
 104 tations of our approach.

105 2 Preliminary

106 In this section, we first introduce three
 107 verbalization-based methods (§2.1), then discuss
 108 the prevalent metrics for confidence estimation
 109 (§2.2). Finally, we present the datasets and models
 110 evaluated in this study (§2.3).

111 2.1 Methods

112 Verbalization-based approaches use natural lan-
 113 guage reasoning to articulate confidence values.
 114 This paradigm effectively harnesses the reasoning
 115 capabilities of the latest LLMs and is applicable
 116 to black-box models where token probabilities are
 117 inaccessible. Following previous work (Tian et al.,
 118 2023), we introduce Verbalized Confidence and
 119 Verbalized Top- k .

120 **Verbalized Confidence** directly prompts an
 121 LLM to reason about both an answer and its con-
 122 fidence score. **Verbalized Top- k** , in contrast,
 123 prompts the model to produce k guesses alongside
 124 a probability for each. The highest-probability pre-
 125 diction is then selected as the final output. Follow-
 126 ing Tian et al. (2023) and Tao et al. (2024), we set
 127 $k = 2$ across all experiments. Our empirical re-
 128 sults (Table 5) demonstrate that increasing k does
 129 not lead to substantial improvement.

130 In this work, we propose **Verbalized Probabil-**
 131 **ity Distribution**, a method that prompts LLMs
 132 to reason about a probability distribution over
 133 the entire space of possible answers, thereby en-
 134 couraging a comprehensive confidence quantifica-
 135 tion. While it shares the conceptual similarity
 136 of considering alternatives with Verbalized Top-
 137 k , it is distinguished by its emphasis on reason-
 138 ing over the complete answer space instead of a
 139 fixed number of guesses. A comparison of differ-
 140 ent verbalization-based methods is shown in Fig-
 141 ure 1. For open-ended questions where enumerat-
 142 ing all answers is infeasible, we include a “None
 143 of the above” option to aggregate the probability
 144 of all low-likelihood candidates. We provide used
 145 prompts in the §C. We also include an introduction
 146 to conventional Logit and $p(\text{True})$ methods in §B,
 147 though they are not the main focus of this research.

148 It is important to note that although the tech-
 149 niques above do not necessarily employ chain-
 150 of-thought (CoT) reasoning, our study focuses
 151 uniquely on challenging reasoning tasks, applying
 152 CoT prompting consistently across all methods.

153 2.2 Metrics

154 Following established literature, we evaluate our
 155 method using the Area Under the Receiver Operat-
 156 ing Characteristic curve (AUROC), the Expected

157 Calibration Error (ECE), and the Brier Score. 200
158 These metrics provide complementary views on 201
159 model performance. 202

160 The AUROC measures a model’s ability to dis- 203
161 criminate between positive and negative classes, 204
162 independent of the classification threshold. It is 205
163 calculated as the area under the plot of the True 206
164 Positive Rate (TPR) against the False Positive Rate 207
165 (FPR). For a set of n predictions, it can be com- 208
166 puted as: 209

$$167 \text{AUROC} = \frac{\sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \mathbb{I}(s_i^+ > s_j^-)}{n_+ \cdot n_-},$$

168 where s_i^+ and s_j^- are the confidence scores for the i - 214
169 th positive and j -th negative instance, respectively, 215
170 and n_+ and n_- are the total number of positive and 216
171 negative instances. 217

172 The ECE quantifies calibration, which is the 218
173 alignment between predicted confidence and em- 219
174 pirical accuracy. It is computed by grouping pre- 220
175 dictions into M bins (B_1, \dots, B_M) based on their 221
176 confidence scores:

$$177 \text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

178 where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ denote the accuracy 222
179 and average confidence within bin B_m . A key lim- 223
180 itation of ECE is that it does not assess a model’s 224
181 discriminative power; a model can be perfectly cal- 225
182 ibrated yet fail to separate the classes. 226

183 To provide a holistic assessment, we also report 227
184 the **Brier Score**, which simultaneously evaluates 228
185 both calibration and discrimination. It is defined 229
186 as the mean squared error between the predicted 230
187 probabilities and the true labels: 231

$$188 \text{Brier Score} = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2,$$

189 where \hat{p}_i is the predicted probability for instance 232
190 i and $y_i \in \{0, 1\}$ is the corresponding true label. 233
191 A lower Brier score indicates better overall per- 234
192 formance. 235

193 2.3 Evaluation Setup

194 We focus on tasks that require reasoning to predict 236
195 the best answer from multiple alternatives, rather 237
196 than on simple classification or true/false verifica- 238
197 tion. To encompass various scenarios, the tasks 239
198 in this work are categorized based on whether the 240
199 answer space is closed-set or open-set. 241

200 For closed-set tasks, we evaluate three multiple- 201
202 choice tasks requiring reasoning: MedQA (Jin 203
204 et al., 2021), MMLU-Pro (Wang et al., 2024c), and 205
206 MedXpertQA (Zuo et al.). For open-set tasks, we 207
208 evaluate two open-ended generation tasks: Hot- 209
210 potQA (Yang et al., 2018) and MedCaseReason- 211
212 ing (Wu et al., 2025). The medical domain is se- 213
214 lected as a suitable and important case study due 215
216 to its practical relevance and reasoning-intensive 217
218 nature. For evaluation, we use exact match 219
220 for closed-set tasks and LLM-as-a-Judge (Zheng 221
222 et al., 2023) for open-set tasks to ensure more pre- 223
224 cise assessment. Evaluation prompts are provided 225
226 in §D. 227

228 Our experiments include a diverse set of LLMs 229
230 at various scales, encompassing both open-source 231
232 models, such as the latest Qwen3 series (Yang 233
234 et al., 2025), and more advanced closed-source 235
236 models, such as GPT-4.1 (Achiam et al., 2023) and 237
238 DeepSeek-V3-0324 (Liu et al., 2024). A detailed 239
240 introduction to the datasets and models used in this 241
242 study is provided in §A. 243

244 3 Experiments

245 In this section, we first compare various 246
247 verbalization-based methods using only prompt- 248
249 ing, without fine-tuning the LLMs (§3.1). Next, 249
250 we investigate how reinforcement learning (RL) 251
252 can improve the reasoning process for confidence 253
254 estimation (§3.2). Finally, we examine the mech- 255
256 anisms behind predicting a verbalized probability 256
257 distribution and discuss the limitations of this ap- 257
258 proach (§3.3). 258

259 3.1 Training-Free Evaluation

260 Given the black-box nature of current LLMs and 261
262 the high cost of training large-scale models, we 262
263 first evaluate various methods in a training-free 263
264 setup, as shown in Tables 1. Our analysis yields 264
265 the following observations. 265

266 **Verbalized Probability Distribution improves** 266
267 **confidence estimation while maintaining an-** 267
268 **swer quality** Conventional approaches based 268
269 on token logits exhibit inconsistent performance 269
270 across datasets and LLMs. For example, while 270
271 $p(\text{True})$ achieves the best AUROC on open-set 271
272 tasks using Qwen3-4B-Instruct, it performs worst 272
273 on closed-set tasks when using Qwen3-30B-A3B- 273
274 Instruct. These methods also suffer from over- 274
275 confidence, often producing confidence scores 275
276 above 95% even when their ability to distinguish 276
277 278

	Closed-Set				Open-Set			
	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow
Qwen3-4B-Instruct	0.553	–	–	–	0.301	–	–	–
\hookrightarrow + Logit	0.553	0.641	0.431	0.429	0.301	0.539	0.579	0.580
\hookrightarrow + p(True)	0.553	0.558	<u>0.361</u>	<u>0.365</u>	0.301	0.692	0.574	0.578
\hookrightarrow + Verbalized Conf.	0.535	<u>0.658</u>	0.426	0.425	0.303	0.645	0.609	0.630
\hookrightarrow + Verbalized Top- k	0.545	<u>0.655</u>	0.392	0.393	<u>0.310</u>	0.640	<u>0.539</u>	<u>0.570</u>
\hookrightarrow + Verbalized Distrib.	0.545	0.670	0.247	0.202	0.324	<u>0.681</u>	0.333	0.331
Qwen3-30B-A3B-Instruct	<u>0.638</u>	–	–	–	0.421	–	–	–
\hookrightarrow + Logit	<u>0.638</u>	0.642	0.349	0.348	0.421	0.582	0.433	0.412
\hookrightarrow + p(True)	<u>0.638</u>	0.636	0.351	0.351	0.421	<u>0.649</u>	0.511	0.513
\hookrightarrow + Verbalized Conf.	0.639	<u>0.656</u>	0.323	0.313	<u>0.435</u>	0.626	0.512	0.519
\hookrightarrow + Verbalized Top- k	0.637	0.645	0.314	<u>0.302</u>	0.444	0.634	0.468	0.477
\hookrightarrow + Verbalized Distrib.	0.635	0.680	0.199	0.171	0.434	0.681	0.304	0.274
GPT-4.1	0.714	–	–	–	0.665	–	–	–
\hookrightarrow + Verbalized Conf.	0.720	0.648	0.263	0.263	0.672	0.679	0.288	0.277
\hookrightarrow + Verbalized Top- k	0.713	<u>0.734</u>	<u>0.197</u>	<u>0.178</u>	0.648	<u>0.705</u>	<u>0.240</u>	0.186
\hookrightarrow + Verbalized Distrib.	<u>0.714</u>	0.760	0.163	0.136	<u>0.669</u>	0.718	0.210	0.100
DeepSeek-V3	0.652	–	–	–	<u>0.576</u>	–	–	–
\hookrightarrow + Verbalized Conf.	0.663	0.689	0.264	0.241	0.583	<u>0.712</u>	0.301	0.285
\hookrightarrow + Verbalized Top- k	0.654	<u>0.707</u>	<u>0.225</u>	<u>0.195</u>	0.558	0.671	<u>0.284</u>	<u>0.242</u>
\hookrightarrow + Verbalized Distrib.	<u>0.656</u>	0.733	0.178	0.143	0.573	0.731	0.215	0.094

Table 1: Comparison of confidence estimation methods across various LLMs. The best result for each metric is highlighted in **bold**, and the runner-up is underlined. We report detailed results on more models for each specific task in §G.

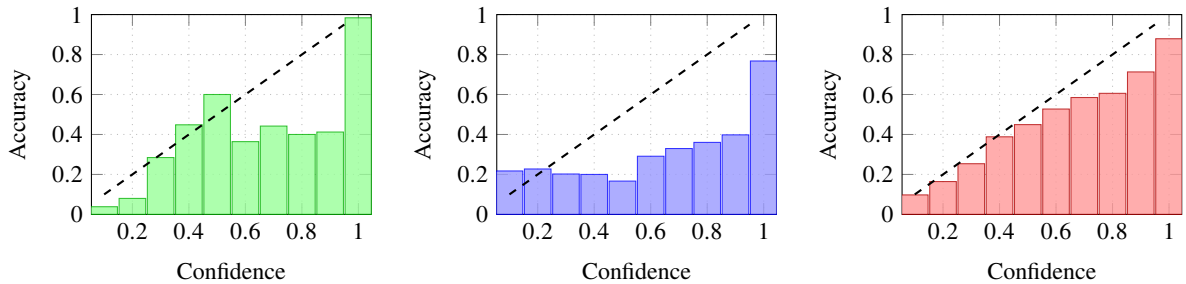


Figure 2: Calibration curves of Qwen3-4B-Instruct on MMLU-Pro when using Verbalized Confidence (Left), Verbalized Top- k (Mid), and Verbalized Probability Distribution (Right).

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 Probability Distribution demonstrates significantly better performance on both closed-set and open-set tasks across all LLMs, while maintaining the competitive performance. Crucially, **smaller LLMs gain more benefits from reasoning verbalized probability distribution**, highlighting the practical utility of this method. The calibration curve in Fig-

ure 2 further illustrates the strong alignment between predicted confidence and empirical accuracy.

Verbalized Probability 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 Probability 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 probability 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 imple-

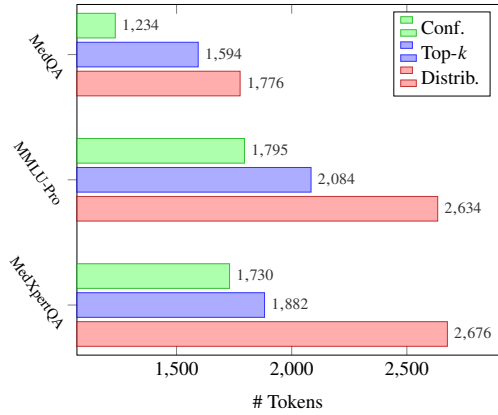


Figure 3: Averaged token consumption of different verbalization-based approaches across MedQA, MMLU-Pro, and MedXpertQA test set when using Qwen3-4B-Instruct.

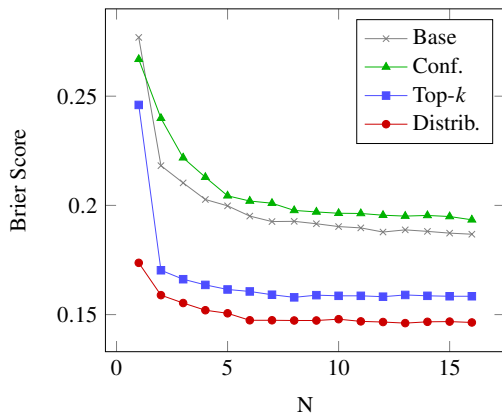


Figure 4: Brier scores of different methods combined with answer consistency on MMLU-Pro test set.

mented the long prompt provided in (Damani et al., 2025). The results, shown in Table 12, demonstrate that **the performance gains stem from the requirement to predict a distribution, rather than from more complex instructions.**

Verbalized Probability 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; 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 ap-

proach, 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 Probability Distribution benefit further from multiple sampling. However, Verbalized Probability 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. The detailed training configuration is described in §A. Our experiments use Qwen3-4B-Instruct, Llama-3.2-3B-Instruct, and Qwen3-8B-Base. Intuitively, the extensively post-trained Qwen3-4B-Instruct has acquired stronger reasoning skills for answer prediction, which may also benefit confidence estimation. In contrast, Qwen3-8B-Base has not undergone instruction tuning, thus may ex-

	Closed-Set				Open-Set			
	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow
<i>RL from the Llama-3.2-3B-Instruct</i>								
RLVR	0.371	–	–	–	0.279	–	–	–
RLCR + Conf.	0.379	0.600	0.320	0.308	0.275	0.678	0.289	0.299
RLCR + Top- k	0.371	0.579	0.340	0.340	0.277	0.709	0.228	0.206
RLCR + Distrib.	0.375	0.602	0.302	0.290	0.286	0.715	0.170	0.094
<i>RL from the Qwen3-4B-Instruct</i>								
RLVR	0.564	–	–	–	0.340	–	–	–
RLCR + Conf.	0.569	0.624	0.340	0.318	0.334	0.742	0.289	0.254
RLCR + Top- k	0.563	0.699	0.283	0.261	0.337	0.752	0.213	0.153
RLCR + Distrib.	0.571	0.705	0.228	0.180	0.338	0.748	0.189	0.093
<i>Zero-RL from the Qwen3-8B-Base</i>								
RLVR	0.527	–	–	–	0.342	–	–	–
RLCR + Conf.	0.529	0.672	0.296	0.276	0.343	0.718	0.271	0.232
RLCR + Top- k	0.515	0.667	0.240	0.200	0.326	0.717	0.202	0.137
RLCR + Distrib.	0.525	0.680	0.219	0.152	0.330	0.725	0.181	0.078

Table 2: Performance of different methods when combined with RL training. Detailed results are reported in §H.

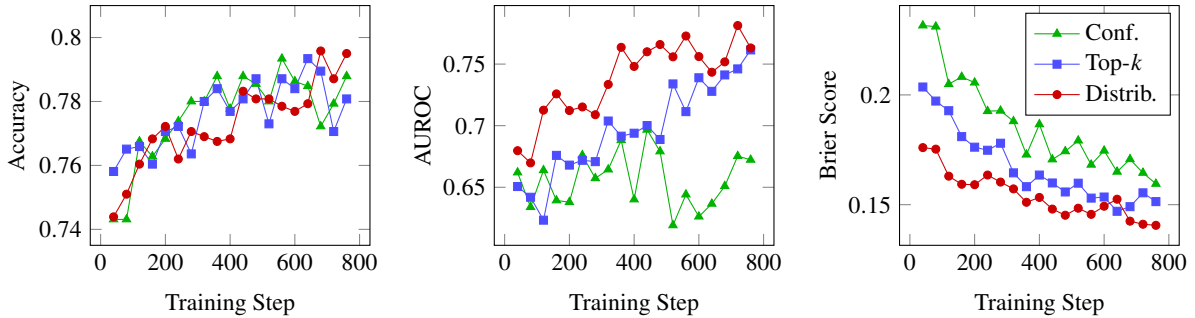


Figure 5: Comparison of different methods on MedQA test set across different training steps when using Qwen3-4B-Instruct. Out-of-distribution results on MMLU-Pro are shown in Figure 11.

353 explore answers in a comparatively less guided manner, relying primarily on knowledge acquired during pretraining.

354
355
356 Results are shown in Table 2. We observe the following conclusions.

357
358 **The advantage of predicting verbalized probability 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 Probability 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 5. This demonstrates that the

373 **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.

374
375
376
377
378 **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 8 and 11: although RL does not significantly improve out-of-domain task accuracy, it consis-

tently enhances the quality of confidence estimation. Using the broader range of topics in MMLU-Pro, we further examine performance across diverse domains in Table 13. Results show unique advantages of Verbalized Probability Distribution on both in-domain and out-of-domain tasks, which again supports the conclusion. Interestingly, **the selection of the training set is crucial for out-of-domain performance**. As shown in Table 11, 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 Probability 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 Probability 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-5-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 3, indicate that while all methods enumerate evidence for their determinations, the Verbalized

	ES	UA	LC	Avg.
Verbalized Conf.	4.25	2.37	2.26	3.08
Verbalized Top- <i>k</i>	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- <i>k</i>	2.59	3.25	2.16	2.67
Verbalized Distrib.	3.47	4.23	3.58	3.76

Table 3: 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).

	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
Verbalized Conf.	0.734	0.741	0.240	0.239
↔ + RLCR	<u>0.737</u>	<u>0.776</u>	0.225	0.229
Verbalized Top- <i>k</i>	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	<u>0.201</u>	<u>0.205</u>

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

Confidence approach sometimes fails to acknowledge its own limitations, leading to lower scores in uncertainty awareness. In contrast, the Verbalized Probability 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 probability distributions activates these in-depth reasoning behaviors.

It is important to note that the presence of these reasoning behaviors does not guarantee better confidence estimation. Performance is also influenced by the quality of the detailed reasoning content, which is model-dependent and therefore difficult to evaluate. For instance, although Verbalized Top-*k* slightly outperforms Verbalized Confidence on HotpotQA (Table 11), it receives lower statistical scores across these behavioral metrics. Nevertheless, we believe the clear advantage of Verbalized Probability Distribution over the other methods helps explain, at least in part, its final stronger performance.

Limitations of the Verbalized Probability Distribution 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 training and evaluation details are shown in §A. Results in Table 4 show that both Verbalized Top- k and Verbalized Probability Distribution suffer 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 confidence 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. An extension for multi-class settings is Temperature Scaling (Guo et al., 2017), which uses a single scalar parameter to 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 aggregat-

ing 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, it is limited to closed-set tasks and 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.

568 Limitations

569 As discussed in §3.3, the effectiveness of
570 verbalization-based methods may vary across
571 tasks. Although we demonstrate that RL can ad-
572 dress this issue, it cannot be directly applied via
573 simple prompting to existing LLMs that lack spe-
574 cific RL training. Flexibly selecting prompting
575 strategies based on task features presents a promis-
576 ing alternative, but it is non-trivial to automati-
577 cally identify these features and understand their
578 influence on confidence estimation strategies. An-
579 other limitation is that predicting verbalized prob-
580 ability distributions consumes more tokens during
581 reasoning. This study does not explore efficient
582 reasoning methods for confidence estimation. Re-
583 lated work on mathematical reasoning has success-
584 fully used RL techniques, such as length penalty
585 rewards (Team et al., 2025; Yeo et al.), to improve
586 efficiency. While a similar approach could be ap-
587 plied here, its effectiveness remains to be verified.

588 References

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

594 Amos Azaria and Tom Mitchell. 2023. The internal
595 state of an llm knows when its lying. In *Findings*
596 *of the Association for Computational Linguistics:*
597 *EMNLP 2023*, pages 967–976.

598 Michael Bereket and Jure Leskovec. Uncalibrated rea-
599 soning: Grpo induces overconfidence for stochastic
600 outcomes. In *LLM for Scientific Discovery: Reason-*
601 *ing, Assistance, and Collaboration*.

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

608 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang
609 Tong, Saining Xie, Dale Schuurmans, Quoc V Le,
610 Sergey Levine, and Yi Ma. Sft memorizes, rl gen-
611 eralizes: A comparative study of foundation model
612 post-training. In *Forty-second International Confer-*
613 *ence on Machine Learning*.

614 Mehul Damani, Isha Puri, Stewart Slocum, Idan Shen-
615 feld, Leshem Choshen, Yoon Kim, and Jacob An-
616 dreas. 2025. Beyond binary rewards: Training lms
617 to reason about their uncertainty. *arXiv preprint*
618 *arXiv:2507.16806*.

Siddartha Devic, Charlotte Peale, Arwen Bradley,
Sinead Williamson, Preetum Nakkiran, and Aravind
Gollakota. 2025. Trace length is a simple uncer-
tainty 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 Wein-
berger. 2017. On calibration of modern neural net-
works. 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 Jitkrit-
tum, Ankit Singh Rawat, Aditya Krishna Menon,
and Sanjiv Kumar. Language model cascades:
Token-level uncertainty and beyond. In *The Twelfth*
International Conference on Learning Representa-
tions.

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 pro-
moting agi with olympiad-level bilingual multi-
modal scientific problems. In *Proceedings of the*
62nd Annual Meeting of the Association for Compu-
tational 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 Con-*
ference on Neural Information Processing Systems
Datasets and Benchmarks Track (Round 2).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-
sch, Chris Bamford, Devendra Singh Chaplot, Diego
de las Casas, Florian Bressand, Gianna Lengyel,
Guillaume Lample, Lucile Saulnier, L elio Ren-
nard Lavaud, Marie-Anne Lachaux, Pierre Stock,
Teven Le Scao, Thibaut Lavril, Thomas Wang, Tim-
oth ee 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. *Jour-*
nal of the American Medical Informatics Associa-
tion, 19(2):263–274.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng,
Hanyi Fang, and Peter Szolovits. 2021. What dis-
ease does this patient have? a large-scale open do-

676	main question answering dataset from medical exams. <i>Applied Sciences</i> , 11(14):6421.	732
677		733
678	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. <i>arXiv preprint arXiv:2207.05221</i> .	734
679		735
680		736
681		737
682		738
683		739
684	Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. <i>Advances in neural information processing systems</i> , 32.	740
685		741
686		742
687	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. <i>CoRR</i> .	743
688		744
689		745
690		746
691		747
692	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. <i>arXiv preprint arXiv:2510.00263</i> .	748
693		749
694		750
695		751
696		752
697		753
698	Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. <i>Transactions on Machine Learning Research</i> .	754
699		755
700		756
701	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. <i>arXiv preprint arXiv:2412.19437</i> .	757
702		758
703		759
704		760
705		761
706	Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lillard, Ola Shorinwa, and Anirudha Majumdar. 2025. Reasoning about uncertainty: Do reasoning models know when they don't know? <i>arXiv preprint arXiv:2506.18183</i> .	762
707		763
708		764
709		765
710		766
711	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. <i>Advances in neural information processing systems</i> , 35:27730–27744.	767
712		768
713		769
714		770
715		771
716		772
717	John Platt and 1 others. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. <i>Advances in large margin classifiers</i> , 10(3):61–74.	773
718		774
719		775
720		776
721	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. <i>arXiv preprint arXiv:2402.03300</i> .	777
722		778
723		779
724		780
725		781
726		782
727	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. <i>arXiv preprint arXiv:2409.19256</i> .	783
728		784
729		785
730		786
731		787
		788
		789
	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 <i>Findings of the Association for Computational Linguistics ACL 2024</i> , pages 5984–5996.	790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807
		808
		809
		810
		811
		812
		813
		814
		815
		816
		817
		818
		819
		820
		821
		822
		823
		824
		825
		826
		827
		828
		829
		830
		831
		832
		833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
		871
		872
		873
		874
		875
		876
		877
		878
		879
		880
		881
		882
		883
		884
		885
		886
		887
		888
		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900

790 Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Ya- 844
791 mada. On verbalized confidence scores for llms. In 845
792 *ICLR Workshop: Quantify Uncertainty and Halluci- 846*
793 *nation in Foundation Models: The Next Frontier in 847*
794 *Reliable AI*. 848

795 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, 849
796 William Cohen, Ruslan Salakhutdinov, and Christo- 850
797 pher D Manning. 2018. Hotpotqa: A dataset for 851
798 diverse, explainable multi-hop question answering. 852
799 In *Proceedings of the 2018 conference on empiri- 853*
800 *cal methods in natural language processing*, pages 854
801 2369–2380. 855

802 Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neu- 856
803 big, and Xiang Yue. Demystifying long chain-of- 857
804 thought reasoning in llms. In *ICLR 2025 Workshop 858*
805 *on Deep Generative Model in Machine Learning: 859*
806 *Theory, Principle and Efficacy*. 860

807 Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunky- 861
808 oung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, 862
809 Yireun Kim, and Minjoon Seo. 2025. Reason- 863
810 ing models better express their confidence. *arXiv 864*
811 *preprint arXiv:2505.14489*. 865

812 Bianca Zadrozny and Charles Elkan. 2001. Obtaining 866
813 calibrated probability estimates from decision trees 867
814 and naive bayesian classifiers. In *Icml*, volume 1. 868

815 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Ke- 869
816 qing He, Zejun Ma, and Junxian He. 2025. Simpl- 870
817 erl-zoo: Investigating and taming zero reinforcement 871
818 learning for open base models in the wild. *arXiv 872*
819 *preprint arXiv:2503.18892*. 873

820 Boxuan Zhang and Ruqi Zhang. 2025. Cot-uj: 874
821 Improving response-wise uncertainty quantification 875
822 in llms with chain-of-thought. *arXiv preprint 876*
823 *arXiv:2502.17214*. 877

824 Jiayi Zhang, Simon Yu, Derek Chong, Anthony Si- 878
825 cilia, Michael R Tomz, Christopher D Manning, 879
826 and Weiyang Shi. 2025. Verbalized sampling: How 880
827 to mitigate mode collapse and unlock llm diversity. 881
828 *arXiv preprint arXiv:2510.01171*. 882

829 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan 883
830 Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, 884
831 Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 885
832 2023. Judging llm-as-a-judge with mt-bench and 886
833 chatbot arena. *Advances in neural information pro- 887*
834 *cessing systems*, 36:46595–46623. 888

835 Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, 889
836 Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, 890
837 and Bowen Zhou. Medxpertqa: Benchmarking 891
838 expert-level medical reasoning and understanding. 892
839 In *Forty-second International Conference on Ma- 893*
840 *chine Learning*.

841 A Detailed Setup

842 **Datasets.** The tasks used in this work are catego-
843 rized based on whether the answer space is closed-

set or open-set. For the closed-set tasks, we eval-
uate three multiple-choice tasks requiring reason-
ing: 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 in-
stances, respectively. MedQA and MedXpertQA
are both medical benchmarks, with the latter be-
ing notably more difficult. MMLU-Pro offers a
comprehensive evaluation across diverse domains,
such as physical, business, and law. For the open-
set tasks, we employ HotpotQA (Yang et al., 2018)
and MedCaseReasoning (Wu et al., 2025). Hot-
potQA focuses on multi-hop question answering,
while MedCaseReasoning comprises clinical diag-
nosis cases. We utilize the distractor subset of
HotpotQA, from which we sample 1,000 test in-
stances. No supporting facts are provided, because
we notice this is too simple for recent LLMs. Re-
garding 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 closed-set tasks, we compute ac-
curacy using an exact match. For open-set tasks,
we employ LLM-as-a-Judge (Zheng et al., 2023)
to compare the prediction against the ground truth.
As the ground truth is given, this task can be
easily handled even for open-source small LLMs.
We consistently adopt Qwen3-30B-A3B-Instruct
for evaluation. Evaluation prompts are presented
in §D. For the math reasoning task involved in
§3.3, the correctness of an answer is verified with
Math-Verify¹.

Models. Our evaluation encompasses a diverse
set of LLMs of varying scales. This in-
cludes the latest models from the Qwen3 se-
ries (Qwen3-4B-Instruct-2507 and Qwen3-30B-
A3B-Instruct-2507, Yang et al. 2025), the es-
tablished Mistral-3.2-24B-Instruct (Jiang et al.,
2023), and the medical-specific Baichuan-M2-
32B-GPTQ-INT4 (Dou et al., 2025). We also
tested Llama-3.2-3B-Instruct. However, due to its
limited capability, this model often fails to follow
output format instructions. Therefore, we only
evaluated it within the RL training setting. Fur-
thermore, we tested more advanced closed-source
models, including GPT-4.1 (Achiam et al., 2023)
and DeepSeek-V3-0324 (Liu et al., 2024). We
also evaluated recent LLMs capable of performing

¹<https://github.com/huggingface/Math-Verify>

894 long chain-of-thought (CoT) reasoning, such as
 895 the open-sourced Qwen3-4B-Thinking-2507 and
 896 the closed-sourced GPT-5-Mini. Due to the high
 897 cost associated with querying APIs for closed-
 898 source LLMs, we evaluated these models on a ran-
 899 dom sample of up to 1,000 instances per dataset.

900 **Training.** Our training is based on the prevalent
 901 veRL framework (Sheng et al., 2024). During the
 902 rollout phase, we sample 512 instances, generat-
 903 ing 8 rollouts for each with a temperature of 1. For
 904 the policy update phase, we use 128 instances and
 905 their corresponding rollouts, with a learning rate
 906 of 1×10^{-6} . This configuration results in 4 training
 907 steps per rollout step, balancing the efficiency of
 908 the rollouts with the preservation of the on-policy
 909 nature of the algorithm. For the math reasoning
 910 task, following prior work (Zeng et al., 2025), we
 911 adopt the MATH dataset (Hendrycks et al.) for
 912 RL training. Due to the already strong perform-
 913 ance of Qwen3-4B-Instruct on MATH, to en-
 914 sure training efficacy on challenging problems, we
 915 filter the dataset to include only instances where
 916 the model fails in at least one of four sampled re-
 917 sponses (Team et al., 2025). This yields a final set
 918 of 2,359 training instances.

919 B Conventional Methods

920 We introduce two conventional yet representa-
 921 tive methods that use the model’s generated to-
 922 ken probabilities as confidence scores: Logit and
 923 $p(\text{True})$ (Kadavath et al., 2022).

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

$$932 \text{Conf}_{\text{Logit}} = \prod_{t=1}^L P(\hat{a}_t | x, r, \hat{a}_{<t}).$$

933 The **$p(\text{True})$** method prompts the model to eval-
 934 uate its own generated response, \hat{y} . The confi-
 935 dence score is the probability the model assigns
 936 to generating the token “True” in response to the
 937 judgment prompt p :

$$938 \text{Conf}_{p(\text{True})} = P(\text{“True”} | x, \hat{y}, p).$$

Prompt for Verbalized Confidence

939 **Question:** [Question]

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

```
945 {
  946   "final_answer": "Your final answer",
  947   "confidence": "0-1"
  948 }
```

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

Prompt for Verbalized Top- k

949 **Question:** [Question]

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

```
953 [
  954   {
  955     "candidate": "first most likely answer",
  956     "confidence": "0-1"
  957   },
  958   {
  959     "candidate": "second most likely answer",
  960     "confidence": "0-1"
  961   }
  962 ]
```

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

963 C Prompts for Verbalization-Based Methods 940

941 We summarize the instructions for verbalization-
 942 based methods: Verbalized Confidence (Figure 6),
 943 Verbalized Top- k (Figure 7), and Verbalized Prob-
 944 ability Distribution (Figure 8).

945 D Evaluation Prompts

946 This section summarizes the evaluation prompts,
 947 which comprise a prompt for judging the correct-
 948 ness of a prediction against the gold answer (Fig-
 949 ure 9) and a prompt for scoring reasoning patterns
 950 (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.

E 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 5. Our results demonstrate that different values of k yield competitive performance.

F The Quality of the Full Probability Distribution

In the main text, we follow common practice by evaluating models based on a single final answer

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

Table 5: 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
\hookrightarrow + RLCR	0.347	0.435	1.124

Table 6: Multi-Label Brier Scores of Qwen3-4B-Instruct on MedQA, MMLU-Pro, MedXpertQA test sets.

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 6, RL improves the calibration of the full distribution, not just the probability assigned to the final answer.

G Detailed Results of Training-Free Experiments

Table 7 and Table 8 present the detailed results for the training-free evaluation of various LLMs across specific tasks. These results are consistent with the findings in the main text, showing that the Verbalized Probability Distribution method typically ranks as the best or runner-up across confidence estimation metrics. Results for LLMs capable of long chain-of-thought (CoT) reasoning are shown separately in Table 9. We observe that, in most cases, directly using long CoT reasoning hurts confidence estimation performance. This underscores the importance of the specialized training explored in our RL experiments. Nevertheless, these results still demonstrate the overall effectiveness of the Verbalized Probability Distribution approach.

H Detailed Results of RL Training Experiments

Table 10 and 11 present the training results of various LLMs on each task. The findings largely validate the effectiveness of the Verbalized Probability Distribution approach, achieving compet-

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.

	MedQA				MMLU-Pro				MedXpertQA			
	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow
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	<u>0.534</u>	<u>0.572</u>
↪ + Verbalized Conf.	0.735	0.663	0.238	0.224	0.711	0.764	0.250	0.243	0.160	<u>0.548</u>	<u>0.789</u>	<u>0.808</u>
↪ + Verbalized Top- <i>k</i>	0.749	0.675	<u>0.207</u>	<u>0.181</u>	0.716	<u>0.772</u>	<u>0.236</u>	<u>0.230</u>	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	<u>0.565</u>	<u>0.672</u>	<u>0.689</u>
↪ + Verbalized Top- <i>k</i>	0.870	0.655	0.112	0.072	0.790	<u>0.755</u>	<u>0.172</u>	<u>0.152</u>	<u>0.252</u>	<u>0.525</u>	<u>0.658</u>	<u>0.683</u>
↪ + 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
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	<u>0.695</u>
↪ + Verbalized Top- <i>k</i>	0.780	0.705	0.158	0.047	0.657	0.721	0.229	0.180	0.186	0.556	<u>0.514</u>	<u>0.600</u>
↪ + Verbalized Distrib.	0.785	0.699	<u>0.155</u>	<u>0.052</u>	0.659	<u>0.723</u>	<u>0.209</u>	<u>0.129</u>	0.200	<u>0.549</u>	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	<u>0.207</u>	0.481	0.788	0.789
↪ + p(True)	0.823	0.644	<u>0.140</u>	0.078	0.709	0.520	0.248	0.151	<u>0.207</u>	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- <i>k</i>	0.791	0.704	0.151	0.029	0.700	0.720	0.210	0.156	0.187	0.509	<u>0.526</u>	<u>0.607</u>
↪ + Verbalized Distrib.	0.814	<u>0.700</u>	<u>0.161</u>	0.148	0.698	<u>0.734</u>	0.185	0.070	<u>0.207</u>	0.539	0.303	0.345
GPT-4.1	0.936	–	–	–	0.808	–	–	–	0.397	–	–	–
↪ + Verbalized Conf.	0.936	0.679	0.059	<u>0.053</u>	<u>0.806</u>	0.715	0.172	0.172	0.417	0.551	0.559	0.564
↪ + Verbalized Top- <i>k</i>	0.939	0.822	0.052	0.041	0.796	0.798	0.141	<u>0.092</u>	<u>0.404</u>	<u>0.583</u>	<u>0.399</u>	<u>0.402</u>
↪ + Verbalized Distrib.	0.940	<u>0.812</u>	<u>0.056</u>	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- <i>k</i>	0.882	0.796	0.089	0.044	0.771	<u>0.776</u>	<u>0.151</u>	<u>0.074</u>	0.309	<u>0.550</u>	<u>0.434</u>	<u>0.467</u>
↪ + Verbalized Distrib.	<u>0.883</u>	0.807	0.097	0.096	0.788	0.817	0.135	0.059	0.297	0.574	0.302	0.275

Table 7: Comparison of training-free approaches on MedQA, MMLU-Pro, and MedXpertQA test sets.

itive accuracy while yielding superior scores in confidence estimation metrics. One exception is the Qwen3-8B-Base model on out-of-distribution MedCaseReasoning, which shows relatively lower accuracy. This may be attributed to the fact that this model has not undergone post-training, coupled with a discrepancy between learning to predict a probability distribution and learning to select a single best answer. For instance, we observe that multiple candidates within a given distribution often receive the same confidence score, indicating that the model has not learned to adequately distinguish the most likely option, which affects the final evaluation. This suggests that more diverse training domains or adjusted training objectives could be important for better balancing answer prediction and confidence estimation. Despite this, our method maintains strong overall performance in most cases, which aligns with the conclusions drawn in the main text.

I Experiment Results with a More Complex Prompt

Table 12 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 enumer-

ating 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.

J 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.

	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- <i>k</i>	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- <i>k</i>	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- <i>k</i>	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- <i>k</i>	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 8: Comparison of training-free approaches on HotpotQA and MedCaseReasoning test set.

	MedQA				MMLU-Pro				MedXpertQA			
	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
Qwen3-4B-Thinking	0.762	–	–	–	0.722	–	–	–	0.168	–	–	–
↪ + Verbalized Conf.	0.765	0.615	0.197	0.165	0.723	0.685	0.228	0.207	0.170	0.511	0.713	0.755
↪ + Verbalized Top- <i>k</i>	0.777	0.718	0.174	0.121	0.711	0.780	0.209	0.188	0.176	0.513	0.636	0.698
↪ + Verbalized Distrib.	0.770	0.698	0.163	0.036	0.710	0.774	0.181	0.128	0.168	0.539	0.497	0.584
GPT-5-mini	0.948	–	–	–	0.801	–	–	–	0.445	–	–	–
↪ + Verbalized Conf.	0.944	0.835	0.048	0.042	0.805	0.783	0.145	0.099	0.445	0.603	0.419	0.425
↪ + Verbalized Top- <i>k</i>	0.951	0.861	0.046	0.086	0.818	0.808	0.123	0.062	0.437	0.597	0.358	0.345
↪ + Verbalized Distrib.	0.941	0.862	0.051	0.081	0.812	0.820	0.120	0.043	0.433	0.624	0.280	0.211
	HotpotQA				MedCaseReasoning				Average			
	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
Qwen3-4B-Thinking	0.299	–	–	–	0.317	–	–	–	0.454	–	–	–
↪ + Verbalized Conf.	0.317	0.765	0.333	0.385	0.315	0.639	0.509	0.546	0.458	0.643	0.396	0.412
↪ + Verbalized Top- <i>k</i>	0.330	0.776	0.376	0.432	0.309	0.659	0.426	0.471	0.461	0.689	0.364	0.382
↪ + Verbalized Distrib.	0.317	0.781	0.203	0.169	0.290	0.662	0.356	0.394	0.451	0.691	0.280	0.262
GPT-5-mini	0.669	–	–	–	0.725	–	–	–	0.718	–	–	–
↪ + Verbalized Conf.	0.682	0.840	0.155	0.084	0.720	0.737	0.192	0.120	0.719	0.760	0.192	0.154
↪ + Verbalized Top- <i>k</i>	0.659	0.870	0.174	0.137	0.702	0.702	0.195	0.084	0.713	0.768	0.179	0.143
↪ + Verbalized Distrib.	0.668	0.862	0.151	0.082	0.691	0.723	0.196	0.103	0.709	0.778	0.160	0.104

Table 9: Results on LLMs capable of long chain-of-thought reasoning.

	MedQA				MMLU-Pro				MedXpertQA			
	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓	ACC ↑	AUROC ↑	Brier ↓	ECE ↓
<i>RL from the Llama-3.2-3B-Instruct Model</i>												
RLVR	0.631	–	–	–	0.354	–	–	–	0.127	–	–	–
RLCR + Conf.	0.640	0.664	0.221	0.100	0.362	0.633	0.315	0.289	0.134	0.502	0.423	0.536
RLCR + Top- <i>k</i>	0.630	0.626	0.238	0.131	0.353	0.622	0.335	0.323	0.129	0.488	0.448	0.566
RLCR + Distrib.	0.618	0.652	0.235	0.125	0.373	0.639	0.290	0.252	0.133	0.514	0.382	0.492
<i>RL from the Qwen3-4B-Instruct Model</i>												
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- <i>k</i>	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
<i>Zero-RL from the Qwen3-Base Model</i>												
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- <i>k</i>	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.	0.765	0.733	0.158	0.053	0.639	0.757	0.187	0.018	0.171	0.551	0.312	0.386

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

	HotpotQA				MedCaseReasoning			
	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow
<i>RL from the Llama-3.2-3B-Instruct Model</i>								
RLVR	0.380	–	–	–	0.177	–	–	–
RLCR + Conf.	0.368	0.779	0.196	0.125	0.182	0.577	0.381	0.473
RLCR + Top- k	0.370	0.773	0.187	0.063	0.183	0.644	0.268	0.348
RLCR + Distrib.	0.397	0.784	0.186	0.068	0.175	0.645	0.153	0.119
<i>RL from the Qwen3-4B-Instruct Model</i>								
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
<i>Zero-RL from the Qwen3-8B-Base Model</i>								
RLVR	0.407	–	–	–	0.276	–	–	–
RLCR + Conf.	0.422	0.794	0.177	0.040	0.264	0.642	0.364	0.423
RLCR + Top- k	0.413	0.806	0.176	0.047	0.239	0.627	0.228	0.226
RLCR + Distrib.	0.415	0.808	0.172	0.039	0.245	0.642	0.189	0.116
<i>RL from Qwen3-4B-Instruct using MedCaseReasoning training set</i>								
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 11: Comparison of RL training-based approaches on HotpotQA and MedCaseReasoning test set.

	HotpotQA				MedCaseReasoning			
	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow	ACC \uparrow	AUROC \uparrow	Brier \downarrow	ECE \downarrow
Qwen3-4B-Instruct	0.297	0.739	0.433	0.455	0.283	0.656	0.544	0.588
\hookrightarrow + 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 12: Results on HotpotQA and MedCaseReasoning test set using the long prompt from (Damani et al., 2025).

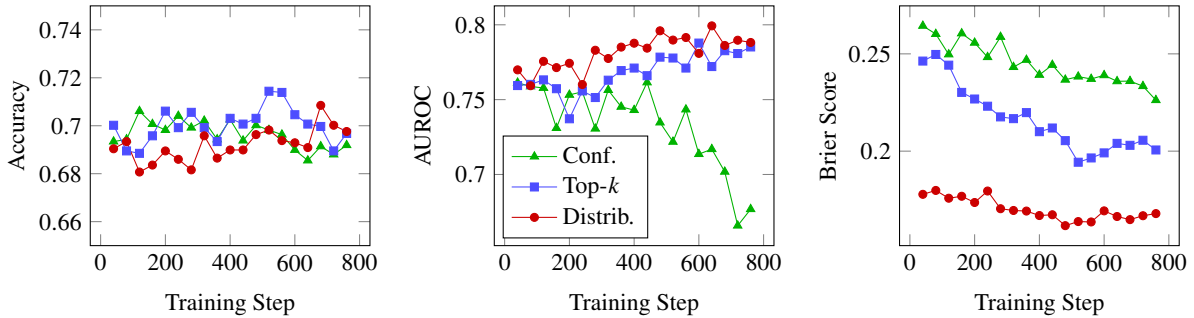


Figure 11: Comparison of different metrics on randomly sampled 2,048 MMLU-Pro test set across different training steps when using Qwen3-4B-Instruct during RL training.

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

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.