CritiCal: Can Natural Language Critiques Help LLMs' Uncertainty or Confidence Calibration?

Anonymous ACL submission

Abstract

Confidence calibration is critical for safe use of Large Language Models (LLMs), where clear verbalized confidence enhances user trust. Traditional methods that mimic reference confidence expressions often fail to utilize the logic in model's original reasoning chain. We propose natural language critique as a solution, which is ideal for confidence calibration, as precise gold confidence labels are hard to obtain and often require multiple generations but assessing whether the confidence is appropriate is easy by analyzing its internal logic and answer correctness. This paper studies how natural language critiques can enhance verbalized confidence: (1) *What to critique*: uncertainty (question-focused) or confidence (answer-specific)? Analysis shows confidence suits multiple-choice tasks, while uncertainty excels in open-ended scenarios. (2) *How to critique*: self-critique or critique calibration training? We propose **Self-Critique**, enabling LLMs to critique and optimize their confidence, and **CritiCal**, using natural language **Criti**ques to train confidence **Cal**ibration. Experiments show that CritiCal significantly outperforms Self-Critique and other competitive baselines on reasoning-intensive questions, even surpassing its teacher, GPT-4o, in one case. CritiCal also shows robust generalization ability, proving its reliability. [1]

1 Introduction

Confidence calibration is crucial in ensuring the trustworthiness of LLMs in high-stakes applications (Vashurin et al., 2025; Xia et al., 2025). As LLMs increasingly interact with humans, verbalized confidence (e.g., "I am 80% confident") clarifies response certainty, fostering trust and effective collaboration (Lin et al., 2022; Xiong et al., 2024).

Critique Fine-Tuning (CFT) (Wang et al., 2025) has proven highly effective in improving LLM's accuracy. It makes LLMs learn from natural language critiques which clarify why answers are correct or incorrect, enabling more reasonable and fine-grained refinements rather than direct imitation of gold responses. We find this characteristic ideal for confidence calibration, as precise gold confidence labels are hard to obtain, but assessing whether the confidence is appropriate is easy based on answer correctness and the logic shown in reasoning. So, this paper investigates whether natural language critiques can help uncertainty or confidence calibration, addressing the following two questions.

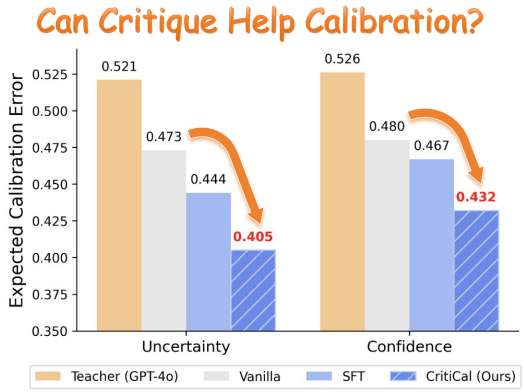

Figure 1: Comparisons between CritiCal and traditional SFT by DS-Qwen-7B on MATH-Perturb, showing CritiCal's huge potential in improving LLM's confidence calibration, since student can even outperform its teacher.

What to critique: uncertainty or confidence? Previous studies often treat uncertainty and confidence as antonyms, overlooking their distinction (Liu et al., 2025b): While confidence evaluates the reliability of a specific response, uncertainty captures the inherent difficulty or ambiguity of the question itself, which is invariant to the choice of answer for a given question. Although Lin et al. (2024b) explored this difference using a consistency-based method, it was limited to the diversity of model output and did not notice the difference between question types. We advance this by conducting a comprehensive study of LLMs' direct outputs and their verbalized uncertainty and confi-

[1] Data and code will be released upon acceptance.

dence. For brevity, we use "confidence" to broadly encompass both concepts, distinguishing them only when comparing their specific roles. *Extensive experiments show that verbalized confidence excels in multiple-choice questions, while uncertainty is better suited for open-ended tasks.*

How to critique: self-critique or critique calibration training? In contrast to prior methods (Huang et al., 2025c; Yang et al., 2025) that focus on self-improving accuracy in the absence of external references, our **Self-Critique** approach also refines LLM's confidence expressions by systematically analyzing the question, the reasoning process, and the final answer. But we find that such zero-shot self-correction often yields unsatisfactory performance. Thus, we further introduce **CritiCal**, a supervised fine-tuning (SFT) method that uses natural language **Criti**ques to train **Cal**ibration. During training, the model is prompted with the question and its own initial response; the target output is a GPT-4o-generated critique of the model's confidence, based on the comparison between model's reasoning process and a golden solution. Extensive experiments, both in-distribution and out-of-distribution, demonstrate that CritiCal significantly enhances confidence calibration for reasoning-intensive questions, *even achieving a better ECE score than its teacher, GPT-4o, in one case*, as is shown in Figure 1. *This suggests that a teacher model's calibration ability does not serve as an upper bound for its student, as long as the teacher possesses sufficient critique ability.* Finally, CritiCal also demonstrates strong generalization in out-of-distribution (OOD) settings.

2 Related Works

Confidence Calibration. Existing calibration approaches for LLMs include white-box methods using internal states or token probabilities and black-box methods based on output consistency or verbalized confidence (Malinin and Gales, 2021; Zong et al., 2025; Lin et al., 2024b; Li et al., 2025; Xu et al., 2024). A key advantage of CritiCal is that it remains black-box and sampling-free at inference time while still improving calibration quality based on natural language critiques.

Critique Learning. Recent critique and self-correction methods mainly improve answer correctness (Madaan et al., 2023; Welleck et al., 2023; Wang et al., 2025; Zhang et al., 2025). In contrast, CritiCal uses natural language critiques specifically for confidence calibration: the teacher critiques the student's confidence by comparing the student's reasoning with a golden reference, enabling the student to learn calibration from its own confidence errors in natural language.

Detailed discussions are in Appendix A.

3 What to Critique

Following prior calibration studies (Lin et al., 2024b), uncertainty and confidence are fundamentally different concepts.

Uncertainty $U(x)$ is a property of the *input* x. It characterizes the dispersion or diversity of all potential answers the model might generate for a given question, reflecting the model's lack of knowledge in the relevant domain. Crucially, uncertainty is *invariant* to the particular answer generated. It is a question-level quantity.

Confidence $C(x, s)$ is a property of a *specific output* s given input x. It measures how strongly the model commits to one particular answer. For the same high-uncertainty question, the model may assign high confidence to a mainstream answer while assigning lower confidence to alternative ones. Confidence is thus an answer-level quantity.

Put simply, uncertainty asks *"Is this question difficult?"*, while confidence asks *"Is this specific answer reliable?"* Despite this distinction, many existing works use these two concepts interchangeably in their prompts (Liu et al., 2025b). We empirically investigate how this choice affects calibration performance across different task formats, with prompt details provided in Appendix B.

4 How to Critique

To investigate how to use critique to enhance confidence calibration, we propose two methods: **Self-Critique**, a prompting-based approach, and **CritiCal**, a supervised fine-tuning (SFT) framework.

4.1 Self-Critique

While prior studies on self-improvement (Huang et al., 2025c; Yang et al., 2025) focuses on refining reasoning processes and improving answer accuracy, **Self-Critique** targets confidence calibration, which aligns model's verbalized confidence score with both the answer correctness and the uncertainty demonstrated in the reasoning chain. The model is prompted to reassess the question, its initial reasoning, and potential ambiguities or logical gaps, refining both the answer and confidence score

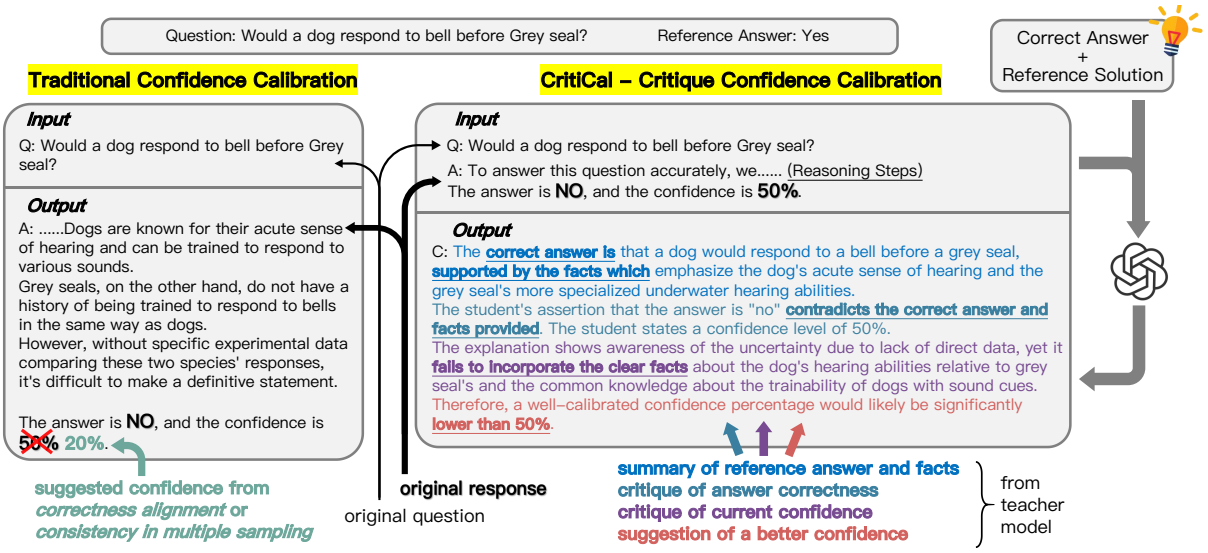

Figure 2: The training data of CritiCal and traditional calibration method. CritiCal introduces a teacher model to critique the student's confidence score based on the original reasoning steps, correct answer, and reference solution.

to improve calibration. The detailed prompt is provided in Appendix B.

4.2 CritiCal

To further enable LLMs to express well-calibrated confidence aligned with their reasoning, we propose **CritiCal**, an SFT method that guides LLMs to refine their confidence expressions using critiques of their initial confidence expressions.

As illustrated in Figure 2, CritiCal differs from traditional methods (Zhang et al., 2024; Xu et al., 2024) in both supervision target and training data structure. Prior approaches usually take the question as input and train the model to output an answer paired with a target confidence score, which is often derived from answer correctness (e.g., 100% if correct) or generation frequency (e.g., 70% for 7/10 appearances). By contrast, CritiCal uses natural language critiques as supervision: a teacher model (GPT-4o) evaluates the student's confidence by identifying hesitations or contradictions in the student's reasoning and comparing the response with a golden reference solution. Because the golden solution is provided, generating critiques for student responses is straightforward and does not require the teacher to rely on its own knowledge. Accordingly, each CritiCal training instance uses the question plus the student's original response as input and the teacher's confidence-calibration critique as target output. This sampling-free design encourages confidence calibration through reasoning-grounded feedback rather than scalar-label imitation.

In practice, we sample 2K questions for each training set and first prompt the student model to generate answers with confidence scores. We then pair these responses with benchmark reference solutions and query the teacher model to produce confidence-calibration critiques. The student model is finally fine-tuned on these critiques. For large reasoning models (LRMs), we instruct the teacher to use special "</think>" tokens to separate explanation from final judgment, which helps mitigate knowledge shift and improve learning effectiveness. The detailed prompt for critique generation is provided in Appendix B.

5 Experiments

In this section, we answer the two questions: what to critique (§5.2) and how to critique (§5.3, §5.4).

5.1 Experimental Setup

Datasets. All the experiments involved a total of 7 datasets: TriviaQA (Joshi et al., 2017) with open-ended, single-hop factuality questions; ComparisonQA (Zong et al., 2025) with multiple-choice, single-hop factuality questions; StrategyQA (Geva et al., 2021) with yes/no, multi-hop factuality reasoning questions; HotpotQA (Yang et al., 2018) with open-ended, multi-hop factuality reasoning questions; MATH (Hendrycks et al., 2021) with open-ended, mathematical reasoning questions; MATH-500 (Lightman et al., 2024a) with harder ones selected from MATH test set; and MATH-Perturb (Huang et al., 2025b) with selected perturbed ones from MATH. (Appendix E)

Models. Our test involves LLMs: Llama-3.1-

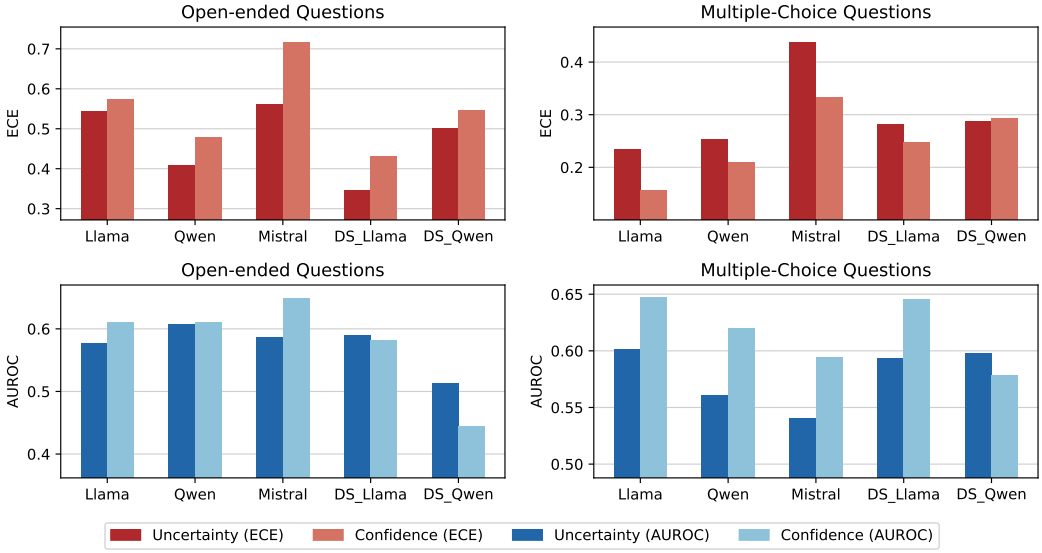

Figure 3: Mean ECE and AUROC values for each model across the same category of benchmarks. The dark bars are the result under uncertainty prompt, and the light ones are of confidence. Further analysis under the setting of multi-turn Self-Critique can be found in Appendix C.

8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024a), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), LRMs: DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025), and an API: GPT-4o (OpenAI, 2024), for their diverse architectures. **Metrics.** Following previous works (Xu et al., 2024; Huang et al., 2025c; Li et al., 2025), we use accuracy (via exact match for open-ended questions) for response correctness measurement, expected calibration error (ECE) with 10 bins for confidence-accuracy alignment, and area under the receiver operating characteristic curve (AUROC) for confidence-based discrimination of correct and incorrect responses. For both accuracy and AUROC, the higher the better, but ECE is the opposite.

5.2 Uncertainty vs. Confidence

Confidence measures the reliability of a specific response, while uncertainty reflects the inherent difficulty or ambiguity of the question itself. Despite their distinction, previous works often treat them casually in prompt (Liu et al., 2025b). Thus, we investigate the impact of their difference by explicitly distinguishing them within the prompt itself across various scenarios, as detailed in Appendix B. We evaluate 5 models across 6 benchmarks (TriviaQA, ComparisonQA, StrategyQA, HotpotQA, MATH-500, MATH-Perturb) grouped into open-ended and multiple-choice question types.

Figure 3 presents the mean ECE and AUROC for each model across benchmark categories. The results reveal that even minor definition distinctions in prompt can elicit statistically different behaviors from models based on question types. **Open-ended Questions:** Uncertainty consistently achieves a lower ECE across both LLMs and LRMs. Open-ended tasks have a near-infinite answer space, making confidence estimates overly sensitive to superficial response phrasing. Uncertainty, being question-focused rather than answer-focused, captures the model's general domain mastery and thus yields more stable calibration. **Multiple-Choice Questions:** The trend reverses, with confidence significantly outperforming uncertainty in both ECE and AUROC. The constrained option set enables elimination strategies (i.e., comparing logical contradictions across choices to derive precise confidence scores), even when residual question ambiguity remains.

Thus, uncertainty and confidence should not be used interchangeably in prompts. Their definitions should be aligned with the task format to ensure optimal model calibration.

5.3 Self-Critique Analysis

This section investigates the impact of Self-Critique on confidence calibration and average confidence scores across multiple iterations.

5.3.1 Multi-Turn Self-Critique

To comprehensively evaluate Self-Critique performance, we conduct multi-turn Self-Critique experiments with 4 models, those have both reasoning and non-reasoning ones, across the same 6 benchmarks. In each iteration, the model receives the

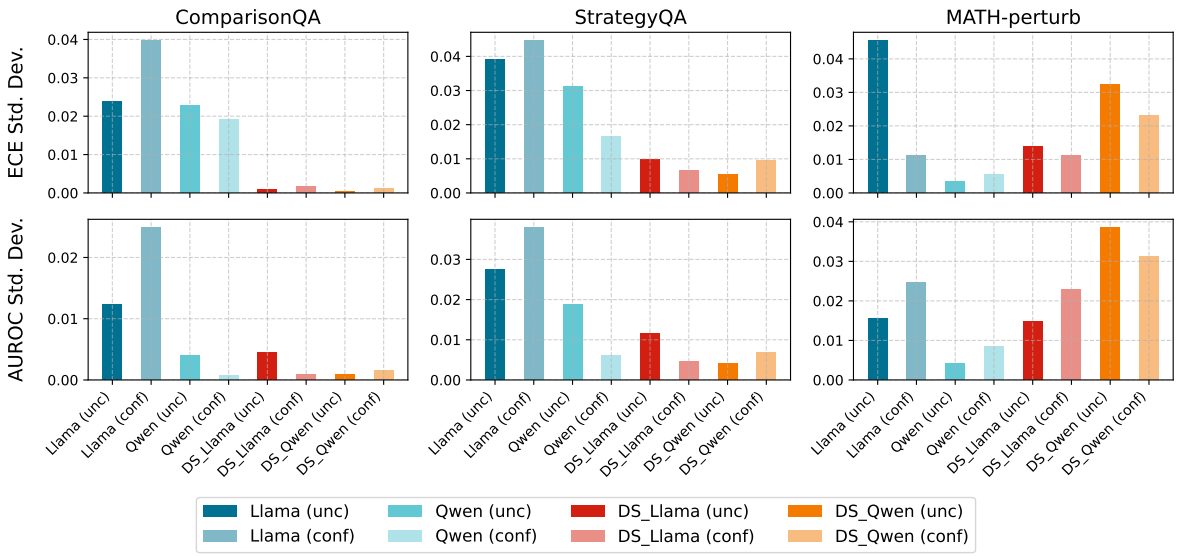

Figure 4: Standard deviation of multi-turn Self-Critique for ECE and AUROC across three benchmarks. Each bar represents the standard deviation of a model's performance (uncertainty or confidence) across 6 iterations, where iteration 0 denotes the original response and iterations 1–5 indicate Self-Critique. Benchmarks are selected as representative of their task category due to question similarity under the same type. Full results are in Appendix C.

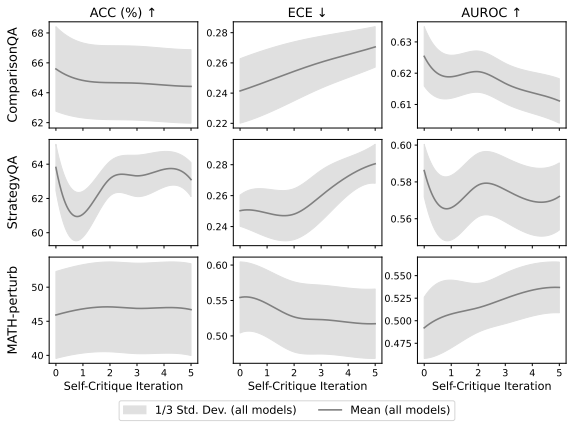

Figure 5: Multi-turn Self-Critique results on ComparisonQA, StrategyQA, and MATH-perturb benchmarks. Each plot shows the smoothed mean performance (solid line) and the corresponding 1/3 standard deviation range (shaded area) for ACC, ECE, and AUROC. Iteration 0 represents the original response without Self-Critique.

results of **all** previous iterations as context. Detailed prompt is provided in Appendix B.

Figure 4 illustrates the standard deviation of multi-turn Self-Critique for each model, focusing on ECE and AUROC to evaluate confidence calibration stability. Figure 5 presents the average performance of all models across three benchmarks, highlighting the impact of Self-Critique on different tasks. The specific variation curves of each model are shown in Figure 12 in Appendix C.

Task Analysis. The six benchmarks are categorized into three tasks: one-hop factuality (ComparisonQA and TriviaQA), multi-hop factuality reasoning (StrategyQA and HotpotQA), and math reasoning (MATH500 and MATH-Perturb). As shown in Figure 5, the semi-transparent light gray area represents the average performance of all models with a one-third standard deviation. For accuracy, models exhibit greater stability on one-hop factuality and math reasoning tasks compared to multi-hop factuality reasoning. However, unlike prior self-improvement studies (Madaan et al., 2023; Welleck et al., 2023), Self-Critique shows no notable accuracy improvements, as it primarily targets confidence calibration rather than answer correctness. For ECE and AUROC, Self-Critique exhibits relatively stable performance with slight improvements in math reasoning tasks. In contrast, factuality-related benchmarks experience negative impacts, with increased average ECE and decreased average AUROC. Without access to external knowledge, models tend to generate hallucinatory reasoning to rationalize their initial answers, creating a closed-loop error amplification where Self-Critique reinforces the mistake rather than correcting it. These findings suggest that Self-Critique has limited effectiveness, significantly worsening calibration for factuality-related tasks while only marginally enhancing it for math reasoning. Thus, prompting-based Self-Critique alone is inadequate for robust confidence calibration.

Model Analysis. In Figure 4 and Figure 12, LLMs are represented in cool colors, while LRMs are depicted in warm colors. For ECE and AUROC, LRMs demonstrate greater stability on factuality-

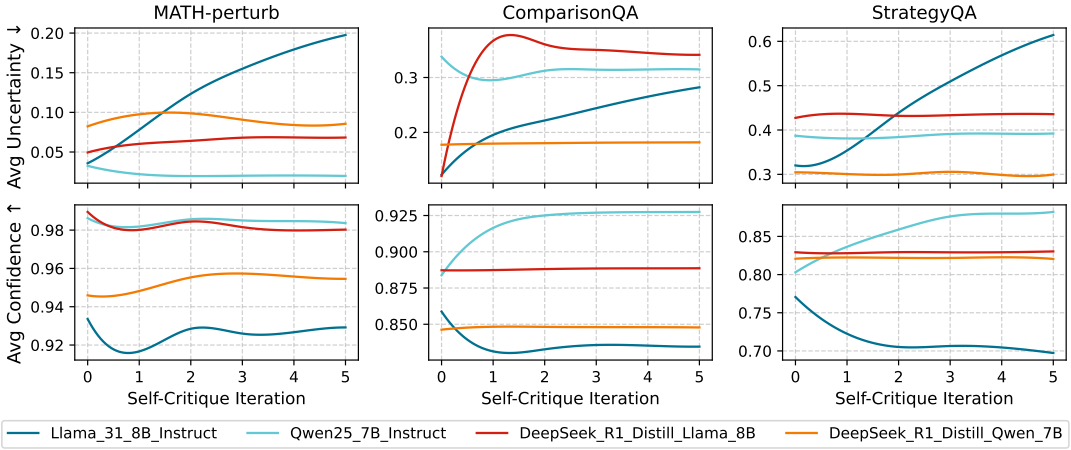

Figure 6: Curves of average uncertainty and confidence scores during multi-turn Self-Critique across 3 benchmarks.

related benchmarks, with significantly lower standard deviation than LLMs, whose calibration varies widely. This stability in LRMs probably arises from their extended reasoning processes, enabling deeper reflection on initial responses and preventing erratic confidence shifts. Although LRMs show an increase in standard deviation in math-related questions, Figure 12 reveals that this stems from their progressively refined confidence calibration. Overall, LRMs exhibit more consistent and reliable confidence calibration compared to LLMs.

5.3.2 Average Confidence Change during Self-Critique

Different from prior works (Huang et al., 2025c) finding models become more confident despite incorrect answers during self-improvement, Self-Critique focuses on refining confidence calibration, leading to more complex outcomes as it prioritizes confidence expression over answer correctness.

Results, shown in Figure 6, vary significantly by model. Llama consistently increases in uncertainty and decreases in confidence across all benchmarks, while Qwen shows the opposite trend, becoming more confident and less uncertain. This suggests that multi-turn Self-Critique amplifies these model-specific tendencies. The two DeepSeek distilled models generally maintain more consistent uncertainty and confidence scores compared to non-reasoning models, except for an increase in uncertainty of the distilled Llama on ComparisonQA. This indicates that extended reasoning processes enhance the robustness of confidence expressions.

5.4 CritiCal Analysis

We evaluate the performance of CritiCal in both in-distribution and out-of-distribution settings across multiple benchmarks.

We select one representative benchmark from each of the 3 tasks outlined in §5.3.1: one-hop factuality (ComparisonQA), multi-hop factuality reasoning (StrategyQA), and math reasoning (MATH-Perturb). For fair comparison, we randomly sample 2K questions from the training set to construct training data each time, using the method described in §4.2. For MATH-Perturb, where questions are perturbations of a subset of MATH, we sample from the original MATH training set to build training data and test only on perturbed questions from the original test set to prevent data leakage. Training is conducted using LlamaFactory (Zheng et al., 2024) with a batch size of 64 and other default hyper-parameters, taking approximately half an hour per each dataset on a 45G single GPU.

5.4.1 In Distribution

We first test the in-distribution performance of models fine-tuned with CritiCal, using Qwen and DeepSeek-Distill-Qwen as examples. Results are presented in Table 1.

For fair comparison, we include several sampling-free baselines: (1) **Vanilla**, a zero-shot prompt that directly asks model's verbalized confidence (Xiong et al., 2024). (2) **Self-Critique**, the non-training method described in §4.1. (3) **SFT_Hard**, a SFT approach using a suggested confidence score based on model's original response (0% for incorrect answers, 100% for correct) for calibration (Zhang et al., 2024), with uncertainty as the inverse. (4) **SFT_Soft**, a smoother SFT variant with confidence scores of 20% and 80%. (5) The performance of the teacher model, GPT-4o, is also included for reference. In addition, we provide an extra comparison with stronger baselines involv-

Model	Type	Method	Train	None Reasoning			Require Reasoning					
				ComparisonQA			StrategyQA			MATH-Perturb		
				ACC (↑)	ECE (↓)	AUROC (↑)	ACC (↑)	ECE (↓)	AUROC (↑)	ACC (↑)	ECE (↓)	AUROC (↑)
GPT-4o	Unc	Vanilla	N	90.91	0.089	0.772	78.60	0.079	0.740	42.36	0.521	0.695
	Conf	Vanilla	N	91.97	0.036	0.787	79.48	0.103	0.716	44.54	0.526	0.683
Qwen-2.5-7B-Instruct (LLM)	Unc	Vanilla	N	69.65	**0.224**	0.615	64.63	0.283	0.507	39.57	0.587	0.525
		Self_Critique	N	68.24	0.268	0.605	**67.25**	0.308	0.464	40.00	0.583	0.542
		SFT_Hard	Y	69.49	0.229	0.616	65.07	0.288	0.537	36.52	0.605	0.554
		SFT_Soft	Y	69.68	0.228	0.615	64.19	0.245	0.564	38.70	0.593	0.558
		CritiCal	Y	**69.76**	**0.224**	**0.619**	**67.25**	**0.221**	**0.597**	**40.87**	**0.558**	**0.586**
	Conf	Vanilla	N	69.67	0.195	0.628	65.07	0.226	0.612	37.83	0.609	0.571
		Self_Critique	N	68.39	0.238	**0.630**	62.88	0.238	0.603	37.39	0.610	0.578
		SFT_Hard	Y	69.90	0.194	0.629	66.38	0.216	0.616	**41.30**	0.617	0.558
		SFT_Soft	Y	69.90	**0.193**	**0.630**	66.38	0.193	0.629	38.70	0.611	0.562
		CritiCal	Y	**69.97**	0.194	**0.630**	**69.00**	**0.179**	**0.644**	40.00	**0.588**	**0.593**
DeepSeek-R1-Distill-Qwen-7B (LRM)	Unc	Vanilla	N	52.18	0.331	0.586	58.52	0.247	**0.609**	62.54	0.473	0.380
		Self_Critique	N	52.12	0.330	**0.588**	59.83	0.242	0.604	64.87	0.491	0.383
		SFT_Hard	Y	**52.36**	**0.325**	0.578	59.83	0.281	0.558	65.65	0.446	0.413
		SFT_Soft	Y	52.51	**0.325**	0.580	62.45	0.272	0.516	66.09	0.444	0.437
		CritiCal	Y	52.30	0.326	0.579	**65.07**	**0.223**	0.572	**67.83**	**0.405**	**0.457**
	Conf	Vanilla	N	52.35	**0.326**	0.598	58.52	0.261	0.559	65.05	0.480	0.274
		Self_Critique	N	52.31	0.327	**0.602**	57.64	0.278	0.541	65.05	0.516	0.271
		SFT_Hard	Y	52.62	0.332	0.577	61.14	0.242	0.509	66.52	0.487	0.270
		SFT_Soft	Y	**52.66**	0.333	0.578	61.14	0.235	0.597	65.65	0.467	0.301
		CritiCal	Y	52.55	0.333	0.580	**66.81**	**0.176**	**0.630**	**69.13**	**0.432**	**0.328**

Table 1: Performance of various LLMs and LRMs on ComparisonQA, StrategyQA, and MATH-Perturb. The "Train" column indicates whether the method needs additional training, providing a fair comparison. The best performances among all methods are **bold-faced**. CritiCal fails for none-reasoning benchmarks (the gray columns), but shows effectiveness for those requiring reasoning.

ing white-box, multi-sample, and reflection-based methods in Appendix D.

Our key observations are as follows: **(1) CritiCal significantly improves calibration on reasoning-intensive tasks.** CritiCal shows large decreases in ECE and increases in AUROC on StrategyQA and MATH-Perturb compared to all baselines. This advantage stems from the richer supervision of natural language critiques: rather than discrete correctness labels (0/100% or 20/80%), critiques pinpoint which specific reasoning step is flawed, enabling the model to map reasoning quality to confidence values instead of memorizing binary correctness patterns. **(2) CritiCal shows limited improvement on ComparisonQA due to lacking reasoning steps in training data.** CritiCal yields little gain on ComparisonQA. Through inspection of the training data, we find that on such single-hop multiple-choice questions, the model tends to produce direct answers without explicit reasoning steps, leaving the critique with little basis to assess confidence calibration. In contrast, multi-hop reasoning tasks naturally elicit extended reasoning chains, providing rich cues for critique and higher-quality training data. This suggests that models should be encouraged to produce explicit reasoning steps when constructing CritiCal training data. **(3) A teacher's calibration ability does not serve as an upper bound for its student.** On MATH-Perturb, DeepSeek-Distill-Qwen fine-tuned with CritiCal even achieves a lower ECE than GPT-4o itself, suggesting that the student learns calibration from its own errors rather than imitating the teacher's confidence. This is because the teacher's critique is grounded in the student's revealed hesitations or contradictions and the comparison between the student's answer and the golden reference, instead of the teacher's own confidence or private knowledge. **(4) Uncertainty and confidence distinctions persist in CritiCal.** Models trained with CritiCal maintain the pattern from §5.2: uncertainty favors open-ended tasks while confidence favors multiple-choice, as evidenced by superior ECE and AUROC.

5.4.2 Out of Distribution

We test OOD performance on StrategyQA and MATH-Perturb, with results presented in Table 2.

While both SFT_Hard and SFT_Soft exhibit degraded calibration on OOD data, indicating limited generalization, CritiCal consistently outperforms these baselines in OOD settings. Notably, when

Model	Method	Uncertainty						Confidence					
		In-distribution			Out-of-distribution			In-distribution			Out-of-distribution		
		ACC (↑)	ECE (↓)	AUROC (↑)	ACC (↑)	ECE (↓)	AUROC (↑)	ACC (↑)	ECE (↓)	AUROC (↑)	ACC (↑)	ECE (↓)	AUROC (↑)
StrategyQA to MATH-Perturb													
Qwen-2.5-7B-Instruct (LLM)	SFT_Hard	36.52	0.605	0.554	37.83	0.603	0.531	**41.30**	0.617	0.558	37.83	0.610	0.542
	SFT_Soft	38.70	0.593	0.558	39.13	0.610	0.543	38.70	0.611	0.562	36.09	0.625	0.578
	CritiCal	**40.87**	**0.558**	**0.586**	**39.57**	**0.574**	**0.595**	40.00	**0.588**	**0.593**	**42.17**	**0.571**	**0.593**
DS-Distill-Qwen-7B (LRM)	SFT_Hard	65.65	0.446	0.413	66.52	0.444	0.424	66.52	0.487	0.270	64.78	0.493	0.266
	SFT_Soft	66.09	0.444	0.437	64.35	0.450	0.423	65.65	0.467	0.301	65.22	0.476	0.276
	CritiCal	**67.83**	**0.405**	**0.457**	**67.83**	**0.375**	**0.465**	**69.13**	**0.432**	**0.328**	**69.13**	**0.434**	**0.350**
MATH-Perturb to StrategyQA													
Qwen-2.5-7B-Instruct (LLM)	SFT_Hard	65.07	0.288	0.537	63.76	0.272	0.550	66.38	0.216	0.616	65.50	0.203	0.599
	SFT_Soft	64.19	0.245	0.564	64.63	0.288	0.531	66.38	0.193	0.629	65.94	0.216	0.591
	CritiCal	**67.25**	**0.221**	**0.597**	**66.81**	**0.243**	**0.567**	**69.00**	**0.179**	**0.644**	**67.25**	**0.189**	**0.629**
DS-Distill-Qwen-7B (LRM)	SFT_Hard	59.83	0.281	0.558	59.39	0.293	0.544	61.14	0.242	0.509	59.39	0.255	0.544
	SFT_Soft	62.45	0.272	0.516	60.26	0.281	0.535	61.14	0.235	0.597	61.14	0.234	0.568
	CritiCal	**65.07**	**0.223**	**0.572**	**62.88**	**0.230**	**0.586**	**66.81**	**0.176**	**0.630**	**64.19**	**0.197**	**0.620**

Table 2: Comparisons of CritiCal's in-distribution and out-of-distribution performances. OOD Models are all trained on StrategyQA and tested on MATH-Perturb. The best performances among all methods are **bold-faced**.

Type	Method	ACC	ECE	AUROC
StrategyQA (Multi-hop)				
Uncertainty	CritiCal-SFT	67.25	0.221	0.597
	CritiCal-DPO	69.61	0.227	0.614
Confidence	CritiCal-SFT	69.00	0.179	0.644
	CritiCal-DPO	66.81	0.181	0.634
ComparisonQA (One-hop)				
Uncertainty	CritiCal-SFT	69.76	0.224	0.619
	CritiCal-DPO	69.61	0.227	0.614
Confidence	CritiCal-SFT	69.97	0.194	0.630
	CritiCal-DPO	69.94	0.192	0.630

Table 3: Comparisons of using SFT and DPO as the training method respectively for CritiCal.

trained on higher-quality critique training data, CritiCal sometimes achieves even better calibration on OOD questions than on in-distribution ones. This strong generalization likely stems from the nature of natural language critiques on multi-hop reasoning data, which encourage models to build robust, reasoning-grounded confidence strategies that transfer across domains rather than overfitting to surface patterns of a specific dataset.

5.4.3 Analysis of Training Method

We also explore another popular optimization method, DPO (Rafailov et al., 2023), for the training of CritiCal. While the input structure remains the same, DPO differs in its output, consisting of a chosen response and a rejected response. We use teacher critiques (the same as SFT) as the chosen one, and model's Self-Critique as the rejected one due to its suboptimal critique performance.

Since StrategyQA (multi-hop factuality reasoning) and MATH-Perturb (math reasoning) show similar performance trends in Table 1, we test only StrategyQA for multi-hop reasoning due to limited computing resources. For clarity, we denote SFT-based CritiCal as CritiCal-SFT and DPO-based CritiCal as CritiCal-DPO, with results shown in Table 3. We can see that in both multi-hop and one-hop reasoning, the results of CritiCal-SFT and CritiCal-DPO differ very little compared to the huge improvement in Table 1. Given DPO's higher computational cost, SFT remains a more sufficient and efficient training method for CritiCal.

6 Conclusions

This study investigates whether natural language critiques can improve verbalized confidence calibration in LLMs from two perspectives: what to critique and how to critique. For what to critique, we find that confidence works better for multiple-choice tasks, while uncertainty is more suitable for open-ended tasks, providing clear guidance for calibration strategies. For how to critique, we study Self-Critique and propose CritiCal, which uses teacher-generated natural language critiques for calibration training. Experiments show that CritiCal consistently outperforms Self-Critique and other baselines when training data contains reasoning processes, and generalizes well in both in-distribution and out-of-distribution settings. Compared with DPO, SFT is sufficient and more efficient for CritiCal training. These findings highlight the potential of natural language critiques for improving LLM confidence calibration.

Limitations

While CritiCal demonstrates significant improvements in confidence calibration for LLMs, several limitations still exist that cannot be covered in this single work.

The generalizability of CritiCal's performance is potentially constrained by the specific benchmarks used in our experiments. Although we select diverse tasks (one-hop factuality, multi-hop factuality reasoning, and math reasoning), these benchmarks may not fully represent the broad range of real-world scenarios where LLMs are deployed, such as creative writing or multi-modality tasks. Further evaluation on a wider array of datasets could strengthen claims about CritiCal's robustness.

Additionally, computational constraints restrict our ability to evaluate all benchmarks in the comparison of training methods, SFT and DPO, where only ComparisonQA and StrategyQA are tested. Although these benchmarks are carefully chosen to represent one-hop and multi-hop factuality reasoning tasks, this limitation may obscure potential variations in CritiCal's effectiveness across other task types. Future work could leverage greater computational resources to conduct a more comprehensive analysis, incorporating additional benchmarks and training configurations.

Ethics Statement

This paper utilizes several publicly available datasets, including ComparisonQA, TriviaQA, StrategyQA, HotpotQA, MATH, MATH-Perturb, and MATH-500, which are accessible to the research community under CC, Apache 2.0, MIT, Apache 2.0, MIT, MIT, and MIT licenses, respectively. The data is anonymized, ensuring our work does not raise privacy concerns regarding specific entities.

Our experiments involve the use of LLaMA, Qwen, Mistral, DeepSeek Distill Llama, DeepSeek Distill Qwen, and GPT-4o, so the same risks from LLMs research are also applicable to this work.

While CritiCal seeks to increase trust in AI by training on natural language critiques, there is a risk of users overly relying on its confidence estimates. These estimates may occasionally be inaccurate. Therefore, users are advised to treat these confidence expressions as a reference only.

References

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 967–976. Association for Computational Linguistics.

Umesh Bodhwani, Yuan Ling, Shujing Dong, Yarong Feng, Hongfei Li, and Ayush Goyal. 2025. A calibrated reflection approach for enhancing confidence estimation in LLMs. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 399–411, Albuquerque, New Mexico. Association for Computational Linguistics.

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. *CoRR*, abs/2507.16806.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 81 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025a. Efficient test-time scaling via self-calibration. *CoRR*, abs/2503.00031.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. 2025b. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *CoRR*, abs/2502.06453.

Liangjie Huang, Dawei Li, Huan Liu, and Lu Cheng. 2025c. Beyond accuracy: The role of calibration in self-improving large language models. *CoRR*, abs/2504.02902.

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. *CoRR*, abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. Conftuner: Training large language models to express their confidence verbally. *CoRR*, abs/2508.18847.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024a. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024b. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024a. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10351–10368. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024b. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025a. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 206–221. Association for Computational Linguistics.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025b. Uncertainty quantification and confidence calibration in large language models: A survey. *CoRR*, abs/2503.15850.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Arjun S. Nair and Kristina P. Sinaga. 2025. Disproving the feasibility of learned confidence calibration under binary supervision: An information-theoretic impossibility. *CoRR*, abs/2509.14386.

OpenAI. 2024. Hello gpt-4o. *OpenAI*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in*

Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 979–995. Association for Computational Linguistics.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *CoRR*, abs/2310.08118.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Trans. Assoc. Comput. Linguistics*, 13:220–248.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9426–9439. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yubo Wang, Xiang Yue, and Wenhu Chen. 2025. Critique fine-tuning: Learning to critique is more effective than learning to imitate. *CoRR*, abs/2501.17703.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024b. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 6886–6898. Association for Computational Linguistics.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21381–21396. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5985–5998. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2025. Confidence v.s. critique: A decomposition of self-correction capability for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3998–4014. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*

Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 7113–7139. Association for Computational Linguistics.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025. Critique-grpo: Advancing LLM reasoning with natural language and numerical feedback. *CoRR*, abs/2506.03106.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

Qing Zong, Zhaowei Wang, Tianshi Zheng, Xiyu Ren, and Yangqiu Song. 2025. Comparisonqa: Evaluating factuality robustness of llms through knowledge frequency control and uncertainty. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4101–4117. Association for Computational Linguistics.

Appendices

A Detailed Related Works

A.1 Confidence Calibration

Confidence calibration methods for LLMs are divided into white-box and black-box approaches. White-box methods use internal information, such as attention mechanisms (Lin et al., 2024a; Li et al., 2023), hidden layers (Azaria and Mitchell, 2023), or token probabilities (Malinin and Gales, 2021; Zong et al., 2025) for precise confidence estimates. Conversely, black-box ones rely on model outputs without accessing internal structure. Consistency-based methods (Lin et al., 2024b; Huang et al., 2025a; Wang et al., 2024b; Su et al., 2024) assess confidence by sampling multiple outputs and measuring their similarity, assuming consistent responses indicate higher certainty. Verbalization-based approaches (Li et al., 2025; Liu et al., 2025a; Zhang et al., 2024) train LLMs to explicitly express confidence through scores or epistemic markers. SaySelf (Xu et al., 2024) uses a teacher model to generate reflective rationales and confidence scores by analyzing inconsistencies across numerous sampled reasoning chains. However, it mainly trains the student to imitate the teacher's reasoning steps and confidence expressions, while our method asks the teacher to critique the student's confidence by comparing the student's own reasoning with a golden reference, so that the student learns calibration from its own errors. These methods are also computationally inefficient due to reliance on diverse outputs.

A.2 Critique Learning

Self-correction has recently emerged as a promising approach to enhance LLMs' performance. Studies such as Madaan et al. (2023) and Welleck et al. (2023) utilize model's own feedback to refine outputs, though Huang et al. (2024) and Valmeekam et al. (2023) note limitations in its reliability for reasoning tasks. Alternatively, critique learning uses specialized models to provide feedback. Zhang et al. (2025) and Yang et al. (2024b) develop outcome-based reward, while Wang et al. (2024a) and Lightman et al. (2024b) focus on process reward to improve reasoning by evaluating intermediate steps. Further work by Wang et al. (2025) explicitly leverages natural language critiques as a training objective to encourage deeper understanding and reasoning. However, they all focus on

using critique to improve LLM accuracy, whereas our work explores natural language critiques to improve confidence calibration. Damani et al. (2025) uses critique to train LLMs to reason about their uncertainty but is still limited to numerical critiques. Nair and Sinaga (2025) proves that learning calibrated confidence is impossible with only binary supervision. This motivates our exploration of natural language critiques in confidence calibration.

B Prompt

We design our prompt according to previous works (Xiong et al., 2024). Figure 7, 8, 9, 10 illustrate the prompt we use for vanilla uncertainty inquiry, vanilla confidence inquiry, Self-Critique with confidence, and critique generation, respectively, with StrategyQA as an example.

Vanilla prompt using uncertainty

Answer the following yes/no question and provide your uncertainty score. Your response should end with 'The answer is [your_answer], and the uncertainty is [uncertainty_percentage]%' where [your_answer] is yes or no, and the uncertainty percentage is a number between 0 and 100, indicating **how uncertain you are about the question**. If you are not sure, you should give a higher uncertainty percentage.
Question: *[Question]*

Figure 7: The vanilla prompt using uncertainty on StrategyQA. Placeholders *[Question]* will be replaced with the real one.

C Detailed Self-Critique Results

Figure 11 shows the difference between uncertainty and confidence after Self-Critique. The distinctions between these two concepts still remain evident after applying Self-Critique.

Figure 12 displays the performance trajectories of each model across all six benchmarks. Self-Critique demonstrates relatively stronger improvements on mathematical reasoning tasks but falls short on factuality-related tasks, highlighting its limitations and lack of robustness.

D Comparison with Additional Baselines

To further validate the effectiveness of CritiCal, we compare it with several additional baselines on the StrategyQA dataset.

Vanilla prompt using confidence

Answer the following yes/no question and provide your confidence score. Your response should end with 'The answer is [your_answer], and the confidence is [confidence_percentage]%' where [your_answer] is yes or no, and the confidence percentage is a number between 0 and 100, indicating **how sure you are about your answer**. If you are not sure, you should give a lower confidence percentage.
Question: *[Question]*

Figure 8: The vanilla prompt using confidence on StrategyQA. Placeholders *[Question]* will be replaced with the real one.

- **P(True)**: This white-box method (Kadavath et al., 2022) involves a two-step process. First, the model generates an answer. Second, we feed the concatenation of the "Question + Generated Answer" back into the model to query its correctness. The confidence score is defined as the probability assigned to the token "True".
- **Multi-Candidate Sampling with Reflection (MCS_R)**: Adapted from Bodhwani et al. (2025), the original method consists of three components: Multi-Candidate Sampling (MCS), Reflection-based Prompting, and calibration based on golden ordering (for ranking tasks). Since our work targets general generative tasks rather than ranking, we retained the MCS (generating N candidates via temperature sampling) and Reflection components to improve confidence estimation, while excluding the ranking-specific calibration module.
- **Self-Consistency (SC)**: Following Wang et al. (2023), this method employs: (1) multi-path sampling with temperature > 0 (e.g., 0.7), (2) majority voting to select the most frequent answer, and (3) confidence estimation as the frequency ratio of that most frequent answer among all sampled answers (e.g., 7/10 agreement yields 0.7 confidence).
- **Ensemble**: Following Li et al. (2025), this method prompts the LLM k times to gener-

Multi-turn Self-Critique prompt using confidence on StrategyQA

You previously answered the following yes/no question, and your responses may have gone through several rounds of refinement. Below is the question, your initial response, and all subsequent refined responses. Now, reassess the question and your previous reasoning, including the initial and all refined responses. Consider any potential ambiguities, logical steps, or overlooked aspects that could improve the accuracy of your response and the calibration of your confidence score. Answer the question and provide a new confidence score.
Question: *[Question]*
Initial response: *[Initial_Responses]*
All Previously Refined responses: *[All_Refined_Responses_in_Previously_iterations]*
Your response should end with 'The refined answer is [your_answer], and the confidence is [confidence_percentage]%' where [your_answer] is yes or no, and the confidence percentage is a number between 0 and 100, indicating how sure you are about your refined answer. If you are not sure about your refined answer, you should give a lower confidence percentage.

Figure 9: The prompt for multi-turn Self-Critique using confidence on StrategyQA. Placeholders *[Question]*, *[Initial_Responses]*, and *[Refined_Responses]* will be replaced with the real ones.

Critique generation prompt for LLMs on StrategyQA

Confidence indicates how how sure the student is about his answer. If he is not sure, he should give a lower confidence percentage. You are a teacher expert in confidence calibration. A student previously answered a question and provided his confidence score. Please evaluate the calibration of his confidence score for the question based on his response. If his response is incorrect, the confidence percentage should be low.
Question: *[Question]*
Correct Answer: *[Correct_Answer]*
Facts: *[Facts]* (this can be seen as a reference solution)
Student's Response: *[Student's_Response]*
Using the facts and the correct answer as a reference, assess whether the confidence percentage is well-calibrated, considering the clarity and strength of the reasoning provided in the student's response and also your own knowledge of the question. Is the confidence percentage appropriate, too high, or too low? Provide a brief explanation of your evaluation, focusing on how well his confidence aligns with the strength of his reasoning and the context of the question.

Figure 10: The prompt we use to generate confidence calibration critique for LLMs on StrategyQA. Placeholders *[Question]*, *[Correct_Answer]*, *[Facts]*, and *[Student's_Response]* will be replaced with the real ones. For LRMs, we will add the sentence "Use "</think>" to separate your thinking process and your final response" at the end of above prompt.

ate top-k answers with confidence. We group identical answers across runs and average confidence scores within each answer group. The final confidence is the averaged score corresponding to the selected final answer.

For a relatively fair comparison with CritiCal, which provides a direct answer and confidence score in a single inference pass, we set the sampling number to 3, a small number, for the sampling-based methods (MCS_R, SC, and Ensemble). The results are presented in Table 4.

As shown in Table 4, CritiCal demonstrates superior calibration performance (ECE and AUROC) compared to other methods. Notably, P(True) is a white-box method requiring access to internal model probabilities, while MCS_R, SC, and Ensemble are sampling-based methods that necessitate multiple generation passes. In contrast, at in-

Type	Method	Access Type	Gen. Count	Qwen2.5-7B-Instruct			DeepSeek-R1-Distill-Qwen-7B		
				ACC(↑)	ECE(↓)	AUROC(↑)	ACC(↑)	ECE(↓)	AUROC(↑)
-	P(True)	White-box	2	**70.74**	0.606	0.449	57.64	0.246	0.466
Uncertainty	MCS_R	Black-box	3	69.00	0.264	0.504	62.45	0.284	0.497
	SC	Black-box	3	70.31	0.236	0.568	60.70	0.262	0.572
	Ensemble	Black-box	3	70.31	0.336	0.487	60.26	0.237	0.492
	CritiCal	Black-box	1	67.25	**0.221**	**0.597**	**65.07**	**0.223**	**0.579**
Confidence	MCS_R	Black-box	3	67.69	0.251	0.590	64.63	0.207	0.537
	SC	Black-box	3	70.31	0.223	0.592	62.88	0.226	0.583
	Ensemble	Black-box	3	69.43	0.225	0.615	67.52	0.200	0.546
	CritiCal	Black-box	1	69.00	**0.179**	**0.644**	**66.81**	**0.176**	**0.630**

Table 4: Comparison with additional baselines on StrategyQA. **Access Type** indicates whether the method is white-box or black-box. **Gen. Count** denotes the number of model generation passes used per question.

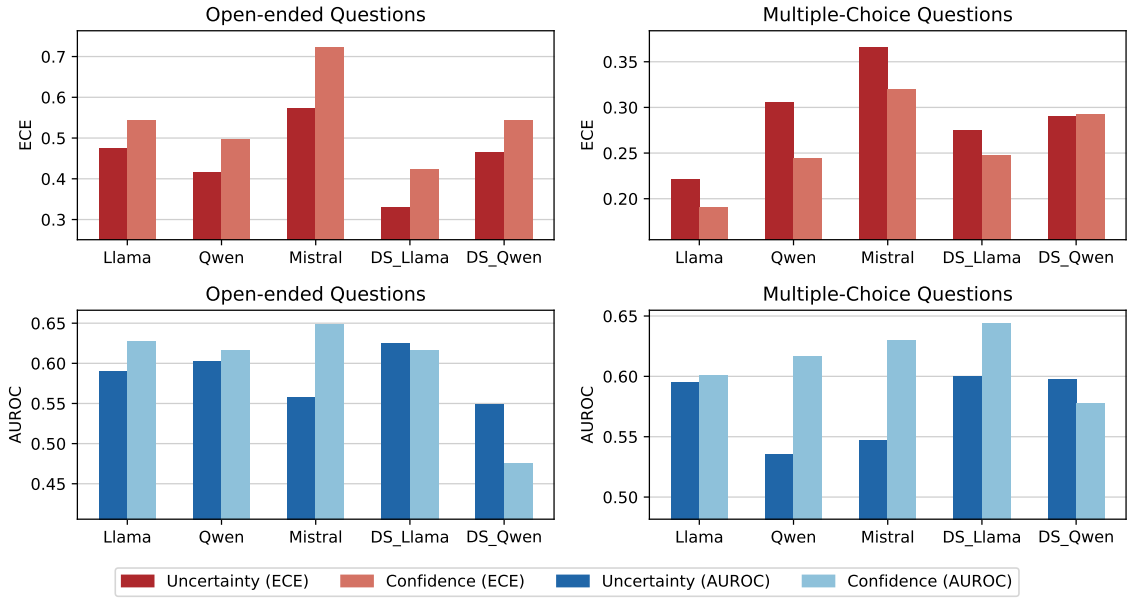

Figure 11: Mean ECE and AUROC values for each model across the same category of benchmarks, which are taken the average of across the 5 turns of Self-Critique. The dark bars are the result under uncertainty prompt, and the light ones are of confidence.

ference time, CritiCal is a black-box, single-pass method that achieves the best calibration performance, highlighting its efficiency and effectiveness. Furthermore, while CritiCal achieves the best ECE and AUROC, its accuracy is not always the highest. This indicates that the improvement in calibration is not merely a byproduct of improved accuracy, but rather a genuine enhancement of the model's ability to assess its own confidence, further proving the validity of our approach.

E Benchmark Statistics

For training, we randomly sampled 2K questions for each benchmark. For testing, we used the full sets of most benchmarks, except for MATH-Perturb. Since MATH-Perturb is constructed from the original MATH dataset, we excluded those derived from the MATH training set to avoid data leakage. Instead, we only evaluated on perturbations from the MATH test set. Note that the training regarding math reasoning relied on the original MATH training set. Table 5 details the question counts for each benchmark.

Benchmark	# Question
TriviaQA	11,313
ComparisonQA	56,696
StrategyQA	229
HotpotQA	7,405
MATH-500	500
MATH-Perturb	230

Table 5: Benchmark statistics: the number of questions we tested in each benchmark.

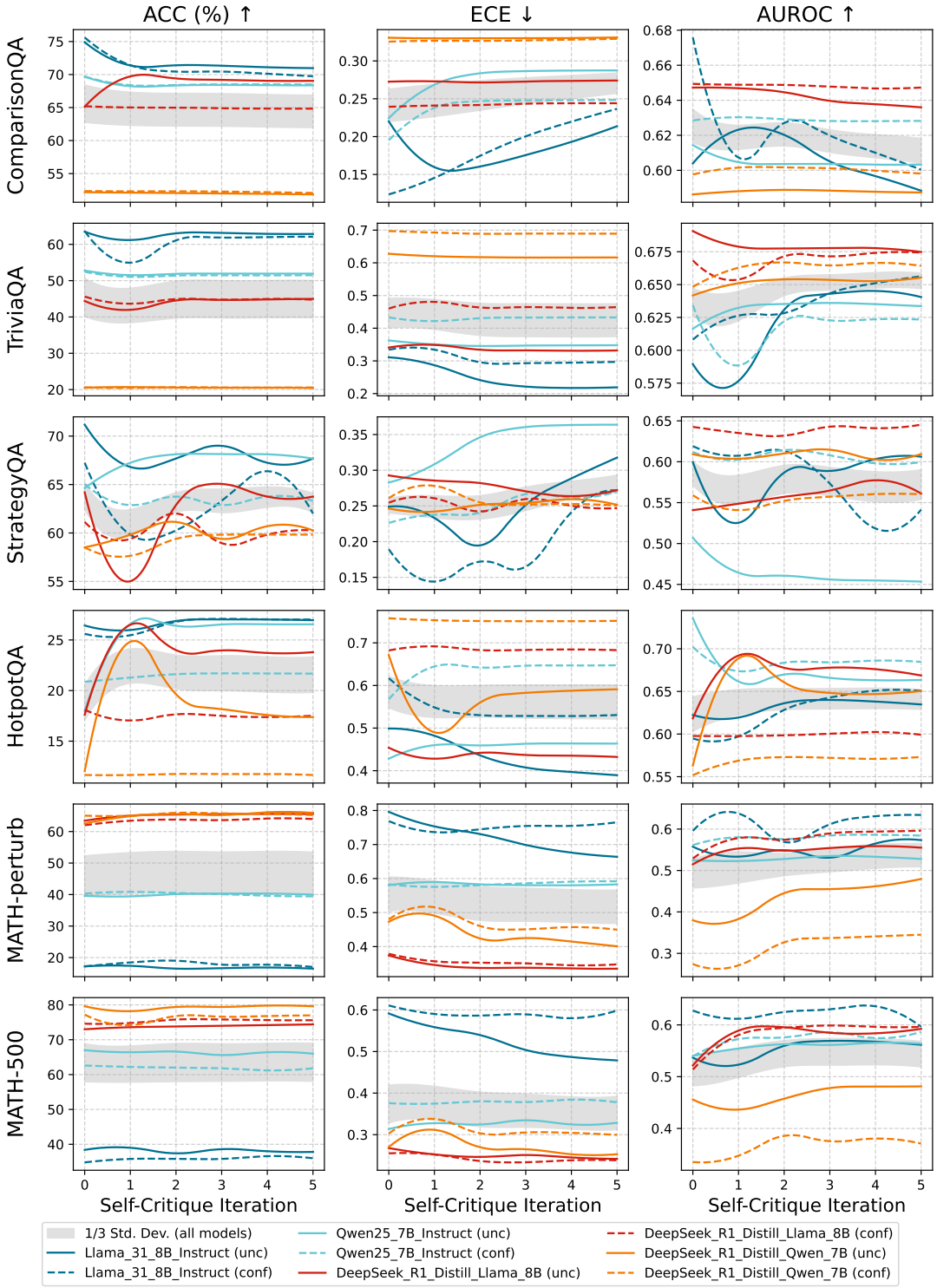

Figure 12: Multi-turn Self-Critique results on all the six benchmarks. The 0 iteration means the original response without Self-Critique. The semi-transparent light gray area represents the average performance of all models with a one-third standard deviation.