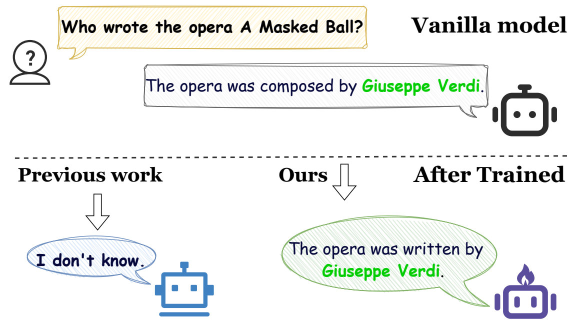
Confidence Introspection: A Self-reflection Method for Reliable and Helpful Large Language Models

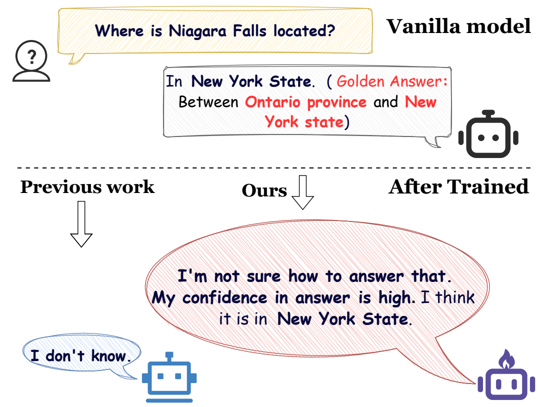
Anonymous ACL submission

Abstract

Large Language Models (LLMs) suffer from factual hallucinations, meaning the LLMs confidently provide responses that are inconsistent with reality. Previous studies explored fine-tuning-based verbalized confidence calibration to mitigate hallucination, yet these approaches often resulted in overly conservative models, compromising their ability to provide relevant knowledge. Inspired by human introspection processes, we propose Confidence Introspection Training, a novel approach that enables LLMs to accurately express their confidence while maintaining helpfulness. This method follows a two-stage framework: first, it estimates the confidence through question paraphrasing and sampling. Subsequently, through self-generated training data, the model develops the ability to classify questions as *known*, *uncertain*, or *unknown* while providing appropriate responses or relevant knowledge for each class. Experimental results demonstrate that our method effectively enhances the reliability of LLMs by accurately expressing confidence levels while preserving the model's ability to provide informative responses. [1]



(a) Our method retains accurate responses from the vanilla model while previous method tends to prevent answering.



(b) Vanilla model provides incomplete yet closely related knowledge to the golden answer. Previous method refuses to answer when the response of vanilla model lack golden answers. Our method preserves the original response with confidence assessment.

Figure 1: Our method compared to previous work.

1 Introduction

Reliability is the ability of LLMs to accurately express the confidence level assigned to a prediction (Mahaut et al., 2024). However, factual hallucinations, which lead LLMs to provide responses that contains nonfactual facts, greatly undermine the LLM's reliability.(Huang et al., 2023; Cheng et al., 2023; Zhang et al., 2024). Verbalized confidence calibration, which expresses confidence in natural language, offers a user-friendly approach to enhance LLM reliability and mitigate hallucination. (Tian et al., 2023b; Zhou et al., 2023).

Among various verbalized confidence calibration methods (Cole et al., 2023; Mahaut et al., 2024), fine-tuning-based methods with binary classification schemes effectively calibrate the confidence of LLMs. Specifically, models are trained to respond "*I don't know*" to potentially incorrect answers(Cheng et al., 2024; Zhang et al., 2024). However, these methods employ strict binary classification criteria, with many questions marked as *unknown*, leading models to favor brief rejecting responses (Section 2.2).

As Figure 1(a) demonstrates, while the vanilla model provides correct answers, the model trained

[1] Our model and data will be made public after the paper is accepted.

by previous method refuse to answer. Moreover, as illustrated in Figure 1(b), these methods ignore that the model likely possesses incomplete though relevant knowledge that could still be helpful to the users. A reasonable trained model should retain the vanilla model's ability to provide question-relevant knowledge while expressing confidence. Previous methods prioritize reliability by avoiding misinformation, but sacrifice **helpfulness** (Bai et al., 2022), the core LLM capability. This motivates us to seek a confidence calibration method that maintains model helpfulness while ensuring reliability.

Drawing from introspection theory (Myers, 1986), a process of internal self-examination that yields clear insights and guides future actions, we propose **Confidence Introspection Training (CIT)**. This approach enables LLMs to learn from self-generated data (CIs dataset), develop accurate confidence awareness across different questions, while preserving inherent helpfulness. We estimate confidence using question sampling accuracy (CRes) and train the model on a five-class confidence classification task (Section 2.5) on self-generated data. During training, we encourage the model to generate the outputs of vanilla model to maintain its question-answering capabilities and prevent overly conservative behavior.

Experimental results demonstrate that, without injecting any external knowledge, the LLM can be tuned to accurately express its confidence level and provide helpful information. Compared to previous work, our method enhances the LLM's reliability and maintains the helpfulness of vanilla models.

In this paper, our contributions are:

- We propose a five-class confidence calibration method, CIT, which tunes the model on self-generated data to accurately express confidence while providing helpful information for answering questions.

- Experiments demonstrate the effectiveness and generalization abilities of CIT. Furthermore, CIT achieves outstanding performance across multiple models on both in-domain and out-of-domain data.

- Analysis and human evaluation demonstrates that CIT can effectively ensure answer generation accuracy and maintaining higher helpfulness compared to existing methods.

2 Confidence Introspection Training

2.1 Confidence Estimation of LLMs

To assess LLM confidence, inspired by Kadavath et al. (2022), we first paraphrase each question and generate five responses each for the original and paraphrased versions. We calculate the sampling accuracy by counting correct responses among these ten outputs using Lexical Matching, which marks responses as correct if they contain the golden answer. This matching method achieves approximately 90% consistency with human evaluations (Wang et al., 2023; Cheng et al., 2024).

We define **CRes** as the total number of correct responses, representing the confidence estimation for each question. Figure 2 shows the distribution of questions across different CRes values for Mistral-7B (Jiang et al., 2023) on TriviaQA (TQA) (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019). For detailed percentages and other model results, see Table 7.

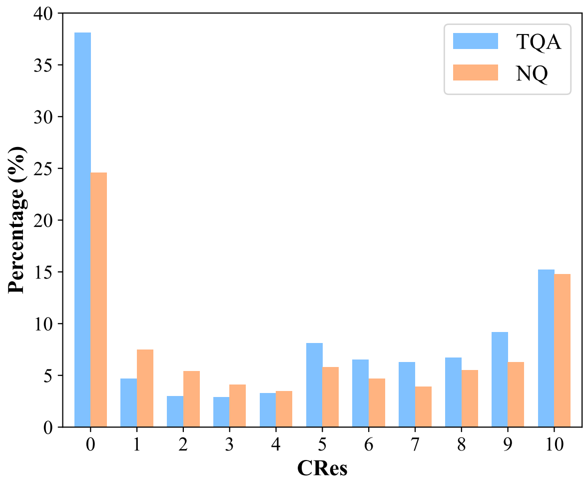


Figure 2: The distribution of CRes values (sampling accuracy) for questions sampled by Mistral-7B on TQA and NQ datasets

To address errors from paraphrasing and sampling, we cannot directly calibrate LLMs using raw CRes values. Instead, we classify questions into groups based on their CRes values and assign each class a unified confidence level for LLM calibration.

2.2 Weaknesses of Binary Classification

Binary classification is a common approach in previous works (Cheng et al., 2024; Xu et al., 2024), where questions answered entirely correctly in sampling are labeled as *known*, while others are considered *unknown*. In other words, they mark all

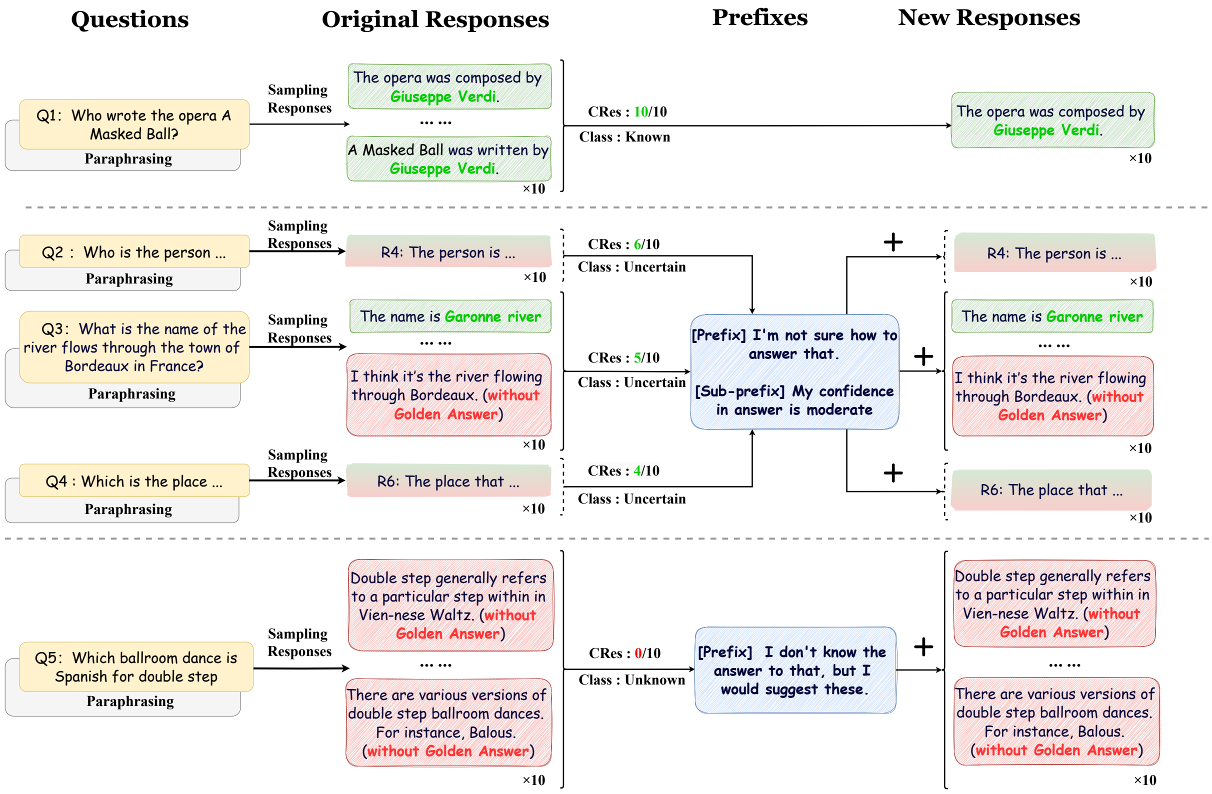


Figure 3: We classify the question based on the CRes as *known*, three confidence level of *uncertain* and *unknown*. We add certain prefixes to responses for *uncertain* and *unknown* questions, forming the expected output data of CIs dataset. In the figure, we use questions with a CRes between $4 \sim 6$ as an example of *uncertain* class. The data construction process for CRes in $1 \sim 3$ and $7 \sim 9$ is the same.

questions with CRes below 10 as *unknown*.

However, this binary classification overlooks the possibility that models may possess partial knowledge of certain questions. The incomplete knowledge leads to inconsistent performance, with questions yielding CRes values between 1 and 9, where the model sometimes answers correctly and other times incorrectly. Moreover, question with CRes between 1∼9 represents a substantial portion (TQA: 46.7%, NQ: 60.7%). Such questions differ from questions where model has no knowledge of the answer. Classifying all these questions as *unknown* is unreasonable, especially for high CRes values like 8 where correct answers are highly probable.

What makes matters worse is that previous works trained the model to respond only with a brief response like *I don't know* when encountering *unknown* questions. The large proportion of training data with simple refusal responses causes training bias, leading the model to habitually refuse to answer most questions, becoming conservative and unhelpful.

2.3 Confidence level and Prefixes

Considering the limitations of binary classification, we classify questions by CRes into three classes, each corresponding to a distinct confidence level: **known**, **uncertain** and **unknown**, as shown in Equation 1.

$$Class = \begin{cases} \textbf{Unknown}, & \text{CRes} = 0 \\ \textbf{Uncertain}, & 1 \le \text{CRes} \le 9 \\ \textbf{Known}, & \text{CRes} = 10 \end{cases} \quad (1)$$

Prefixes of confidence level Building on the three confidence levels, we make the LLM to express confidence level of question through **specific prefixes** while outputting answers. For example, the expected output format for an *unknown* question is:

$$R_{nk} : \{\textbf{Prefix}\} + \{\textbf{Response}\}$$

The "**Prefix**" refers to statements like "*I don't know the answer to that, but I would suggest these*". The "**Response**" refers to the vanilla model's original response to the question. Both *unknown* and *uncertain* questions have corresponding prefixes, as detailed in Appendix B. For *known* questions, no prefix is added and the LLM responds as usual.

The prefixes are generated by the LLM itself, with the model being prompted to produce statements that convey uncertainty or the unknown. Since both the "**Response**" and "**Prefix**" are generated by the model itself, this minimizes the impact on the LLM when changing the output format.

Sub-prefix of uncertain questions In addition to prefixes, we hope the LLM can express confidence levels more precisely for *uncertain* questions, since this class encompasses questions with CRes values from 1 to 9. Considering sampling errors, we further divide uncertain questions into three sub-classes based on CRes, with each sub-class having a specific **Sub-prefix** to indicate its confidence level (CRes range), as shown in Table 1.

CRes	Sub-Prefixes
$1 \sim 3$	My confidence in answer is **low**.
$4 \sim 6$	My confidence in answer is **moderate**.
$7 \sim 9$	My confidence in answer is **high**.

Table 1: Sub-prefixes for three confidence levels of *uncertain* questions with different CRes.

Therefore, for an *uncertain* question, the LLM is expected to response like:

$$R_{uc} : \{\textbf{Prefix}\} + \{\textbf{Sub-Prefix}\} + \{\textbf{Response}\}$$

Overall, our method's confidence levels consist of **five classes**: **known**, **unknown**, and **three confidence levels** for **uncertain** questions. Introducing the *uncertain* class enables the LLM to provide different responses from *unknown* questions, preventing it from defaulting to *I don't know* and refusing to answer most questions. Additionally, this ensures higher reliability for *unknown* questions. A question must be answered incorrectly all five times before and after paraphrasing to be classified as *unknown*.

2.4 CIs Dataset Construction

Based on our five-class confidence classification, we annotated questions from TriviaQA (TQA), Natural Questions (NQ), and GSM8K (Cobbe et al., 2021) for Mistra and LLaMA2 (Jiang et al., 2023; Touvron et al., 2023) to create the Confidence Introspection Dataset (CIs Dataset). We calculate CRes (Section 2.1) and append prefixes (or sub-prefixes) to responses to generate expected outputs. Figure 3 illustrates the data construction process and expected response format.

The details of CIs dataset are shown in Table 9. Our train and dev sets are from TQA training set, while the test set is from TQA, NQ and GSM8K. NQ and GSM8K data serve as out-of-domain tests for method generalization. All sets contain as close to an equal number of questions from each class as possible.

2.5 Training Strategy

Based on the CIs dataset, we perform Confidence Introspection Training, **CIT**, to train the LLM to distinguish between five classes questions: **known**, **unknown**, **three confidence levels** of **uncertain**.The training objective of CIT is shown in Equation 2.

$$L = -\frac{1}{N}\sum_{1}^{N}\begin{cases}\log \mathrm{P}(R|Q_{\textbf{kn}}) \\ \log \mathrm{P}(P_{\textbf{uk}} + R|Q_{\textbf{uk}}) \\ \log \mathrm{P}(P_{\textbf{uc}} + C_{\textbf{uc}} + R|Q_{\textbf{uc}})\end{cases} \quad (2)$$

Here, P denotes the prefix, Q are questions with different confidence level, R represent the original response, C represents the sub-prefix in Table 1. For *known* (**kn**) questions, the LLM generates only the response R. For *uncertain* (**uc**) or *unknown* (**uk**) questions, it generates both the prefix P (sub-prefix C) and response R. This output format reflects the model's confidence while maintaining consistency with its original responses.

In summary, compared to the binary classification in previous methods, we introduce an *uncertain* class that enables detailed confidence assessment. Furthermore, we preserved the original responses of the vanilla model to maintain its ability to answer uncertain and unknown questions, rather than providing no knowledge about the questions. Importantly, for uncertain questions, the original responses likely contain correct answers.

3 Experiments

We treat the model's confidence expression (*known*, *unknown*, three level of *uncertain*) as a **five-classification** task.

3.1 Baselines

We categorize baselines into three types (in Table 2): **Prompt-based**, **Sampling-based**, and **Fine-**

Type	Methods	TriviaQA			Natural Questions			GSM8K		
		ACC ↑	F1 ↑	ECE↓	ACC ↑	F1 ↑	ECE↓	ACC ↑	F1 ↑	ECE↓
Mistral-7B-Instruct										
Prompt	I-Prompt	22.72	13.33	0.35	21.28	11.11	0.42	6.02	3.71	0.65
	UEP	20.02	14.55	0.29	20.52	16.10	0.30	23.61	11.00	0.69
	S-Token	21.62	13.78	0.31	21.40	13.59	0.32	20.76	8.04	0.49
Sampling	TOPK-S	21.06	16.25	0.27	22.60	20.09	0.28	21.71	10.84	0.63
	CEF	21.32	11.05	0.33	20.80	8.82	0.41	4.91	1.89	0.49
Fine-tune	R-Tuning	21.02	13.25	0.28	21.64	14.91	0.28	20.60	11.67	0.56
	Pf-SFT	26.86	22.06	0.25	25.64	22.85	0.29	6.50	4.76	0.57
	IDK-DPO	27.12	23.97	-	22.52	19.54	-	22.03	14.79	-
	CIT(Ours)	**29.52**	**28.59**	**0.14**	**25.96**	**24.25**	**0.22**	**25.20**	**14.94**	**0.34**
LLaMA2-7B-Chat										
Prompt	I-Prompt	20.28	7.37	0.55	20.56	8.20	0.67	1.64	0.65	0.68
	UEP	20.68	12.12	0.26	21.76	15.84	0.42	1.64	0.66	0.92
	S-Token	20.40	14.81	0.42	21.76	13.80	0.32	17.21	6.92	0.40
Sampling	TOPK-S	23.34	19.74	0.15	23.28	19.83	0.27	9.02	6.72	0.61
	CEF	22.18	14.86	0.21	21.92	16.52	0.35	24.59	14.93	0.75
Fine-tune	R-Tuning	22.46	15.99	0.19	22.8	17.32	0.29	19.67	8.07	0.64
	Pf-SFT	22.52	15.25	0.23	21.68	13.62	0.38	18.85	13.70	0.48
	IDK-DPO	21.40	13.76	-	23.20	17.79	-	23.77	7.73	-
	CIT(Ours)	**24.32**	**19.40**	**0.14**	**25.60**	**21.54**	**0.17**	**24.59**	**16.49**	**0.38**

Table 2: The experimental results of ACC, F1 and ECE across three datasets and two models. Since DK-DPO only provides responses for the *known* question, we can't compute its ECE. Compared to other baselines, especially those requiring fine-tuning, our CIT method performs the best and exhibits stronger robustness.

tuning-based methods. For detailed implementation and prompts of baselines, see Appendix C.

Introspection Prompting (I-Prompt) We use natural language instructions to prompt the vanilla model to provide answer and confidence. This can be considered a zero-shot method.

Uncertainty Expression Prompting (UEP) Zhou et al. (2023) used different uncertainty-level prompts to guide model generation, expressing varying levels of confidence.

Surrogate-Token (S-Token) Mahaut et al. (2024) prompted the LLM to provide specific tokens, and use the probabilities assigned to those tokens to predict the confidence.

Top-K Sampling (TOPK-S) Tian et al. (2023a) prompt the model to generate k answers with their estimated confidence scores. Xiong et al. (2024) extended this framework with a consistency-based sampling strategy.

Confidence Elicitation Framework (CEF) Xiong et al. (2024) proposed a prompt-sampling-aggregation framework for eliciting verbalized confidence. We employed vanilla prompts, self-random sampling combined with consistency aggregation.

R-Tuning Zhang et al. (2024) trained the model to predict binary confidence estimates and answers, while their training relied on golden answer of the question.

Pf-SFT Liu et al. (2023) enhanced the SFT method by labeling correct outputs as [GOOD] and incorrect ones as [BAD]. This enabled the LLM to learn human preferences by recognizing that [GOOD]-labeled responses are more suitable answers than [BAD]-labeled ones.

IDK-DPO Cheng et al. (2024) employs Direct Preference Optimization (Rafailov et al., 2023) to train the LLM to say *I don't know* while maintaining normal outputs for known questions.

Type	Methods	TriviaQA		Natural Questions		TriviaQA		Natural Questions	
		LEM-K ↑	LEM-U↑	LEM-K↑	LEM-U↑	LEM-K↑	LEM-U↑	LEM-K↑	LEM-U↑
		Mistral-7B-Instruct				*LLaMA2-7B-Chat*			
Prompt	I-Prompt	61.18	66.29	55.51	39.00	44.21	0.00	31.39	0.00
	UEP	62.87	55.97	50.81	44.76	30.66	60.12	30.65	46.49
	S-Token	65.66	51.17	62.26	49.56	59.02	61.37	57.14	58.14
Sampling	TOPK-S	56.80	58.82	40.10	52.50	53.85	59.85	13.38	48.39
	CEF	66.39	63.20	58.85	57.92	61.73	66.28	54.52	50.74
Fine-tune	IDK-DPO	69.23	-	27.44	-	71.64	-	69.21	-
	Pf-SFT	65.63	60.00	57.86	53.00	67.89	56.28	59.52	38.74
	R-Tuning	70.85	28.26	59.81	48.17	27.67	67.56	49.87	57.42
	CIT-GD	77.58	47.03	16.30	26.96	55.05	51.48	49.59	27.59
	CIT(Ours)	**86.78**	**66.86**	**63.57**	**60.81**	**76.77**	**67.92**	**72.95**	**58.51**

Table 3: The LEM-K and LEM-U of Mistral-7B and LlaMA2-7B. IDK-DPO method only generates responses for *known* class questions. CIT achieves the best generation accuracy across all models and datasets.

3.2 Evaluation Metrics

Introspection Accuracy (ACC) The accuracy of five-class classification task. The class of question is determined by the presence of certain prefixes or sub-prefixes in the response. Responses without any prefix are classified as *known*. [2].

Introspection F1 Score (F1) Macro-F1 scores across the five classes.

Expected Calibration Error (ECE) ECE (Guo et al., 2017) measures the difference between generation accuracy and confidence within specified confidence intervals. A smaller ECE value indicates better calibration performance.

3.3 Implementation Details

We use the Mistral-7B-Instruct, LLaMA2-7B-Chat for our CIT method and other baselines. The learning rate is set to 2×10^{-5}, with 3 training epochs. We conduct all experiments on Nvidia A800 GPUs. For more details, please refer to the Appendix E.

3.4 Main Results

Reliability As shown in Table 2, CIT achieves the best ACC and F1 across all model types and datasets, demonstrating both strong classification ability and effective confidence calibration. Most baselines show significant performance degradation on GSM8K, the mathematical reasoning dataset that differs significantly from the training data. In contrast, CIT demonstrates strong robustness, supporting the reliability of our method. Additionally, fine-tuning-based methods like IDK-DPO, Pf-SFT or sampling-based method like CEF show considerable performance variation across different models, whereas CIT maintains consistent strong performance on both model architectures.

Helpfulness Providing highly accurate responses with high confidence is an essential aspect of LLM helpfulness. Our CIT method achieved the best ECE, demonstrating better alignment between model confidence and generation accuracy. This demonstrates the ability of our CIT approach to maintain or even enhance the helpfulness of the vanilla model.

4 Analysis

4.1 Accuracy of high-confidence question

Although ECE reflects calibration performance, high-confidence responses better serve user needs by indicating *known* or minimally *uncertain* (predicting CRes 7∼9) answers, especially for knowledge-intensive tasks like TQA and NQ. These responses are more likely to provide correct answers or helpful domain knowledge. In Table 3, we introduced Lexical-Matching metrics: **LEM-K** for *known* responses and **LEM-U** for *uncertain* responses. Our CIT method achieves the highest LEM-K and LEM-U scores, demonstrating it maintains the vanilla model's helpfulness while improving calibration [3].

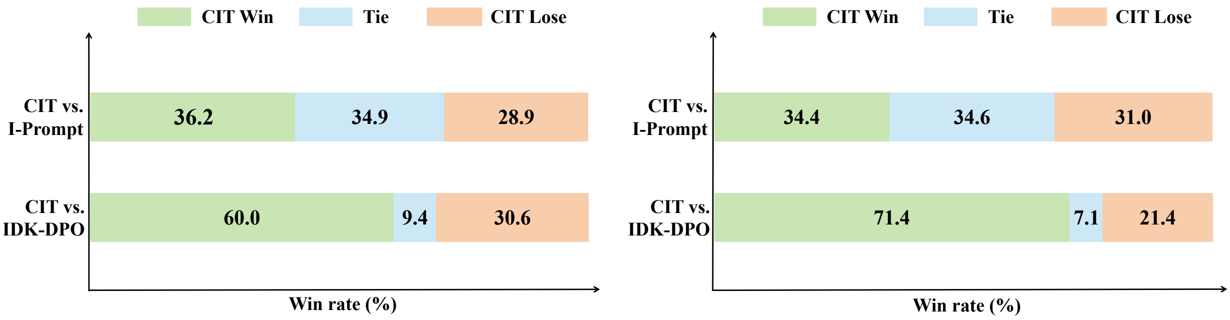
Effect of Self-generated Data To validate the effectiveness of training with model self-generated data, we trained our method with the golden answers (denoted as **CIT-GD** in Table 3). The results

[2] For prompt-based method, we relax the evaluation criteria by adding more prefixes, shown in Appendix C.2.

[3] LEM-U of I-prompt of LlAMa2-7B is 0 because no questions are classified

Types	Methods	ACC$_3\uparrow$	F1$_3\uparrow$	ECE$_3\downarrow$	ACC$_5\uparrow$	F1$_5\uparrow$	ECE$_5\downarrow$	ACC$_{11}\uparrow$	F1$_{11}\uparrow$	ECE$_{11}\downarrow$
Prompt	I-Prompt	35.70	25.22	0.34	22.72	13.33	0.35	10.24	3.71	0.33
	UEP	31.63	6.73	0.27	20.02	14.55	0.29	9.67	3.59	0.48
	S-Token	37.17	31.84	0.17	21.62	13.78	0.31	9.82	3.43	0.29
Sampling	CEF	36.30	24.95	0.31	21.32	11.05	0.33	9.45	3.03	0.31
	TOPK-S	34.47	27.16	0.14	21.06	16.25	0.27	9.15	4.37	0.32
Fine-tune	R-Tuning	33.50	18.87	0.40	21.02	13.25	0.28	10.00	5.73	0.23
	Pf-SFT	40.73	33.88	0.28	26.86	22.06	0.25	10.67	3.73	0.41
	IDK-DPO	40.20	32.85	-	27.12	23.97	-	12.30	11.69	-
	CIT(Ours)	**48.03**	**48.10**	**0.14**	**29.52**	**28.59**	**0.14**	**12.67**	**12.80**	**0.16**

Table 4: Evaluation of CIT and baselines with 3,5,11 classification classes. CIT achieves the best results across different class numbers.



(a) **Win rate on TriviaQA**

(b) **Win rate on Natural Questions**

Figure 4: "Win" indicates that elevators consider the response from CIT to be more helpful. "Lose" indicates that the response of baselines are more helpful. "Tie" means that two responses convey similar meanings.

show that CIT-Golden exhibits substantially decreased LEM-K and LEM-U scores. This indicates that training with self-generated responses, which better align with the model's inherent expression style while avoiding disruption. And this is one reason why R-Tuning adopts similar loss functions yet performs worse than ours.

4.2 Ablation Study

Class Number of Confidence level We conducted experiments of 3-class and 11-class classifications on TriviaQA and compared the results with the 5-class results, as shown in Table 4. When the entire uncertain class is treated as one class, the task becomes a 3-class classification (*known*, *unknown*, *uncertain*). And if each CRes value (1∼9) in uncertain class is treated as one single class, the task becomes an 11-class classification.

The results of 11-class indicate that a smaller gap between different confidence level increases the impact of sampling errors and makes it more challenging for learning of LLM. This is why we did not design CIT based on an eleven-class classification. Moreover, although the 3-class ACC is higher, it lacks specific confidence information for the uncertain question, offering limited assistance to users. Our CIT method achieves the best results across different numbers of class.

Model Size and Type We introduce LLaMA2-13B-Chat, and conduct our CIT method on LLaMA2-7B-Chat and LLaMA2-13B-Chat using three, five and eleven confidence level classes, as shown in Table 5.

We find that LLaMA2-13B-Chat outperforms LLaMA2-7B-Chat across all classification tasks, demonstrating the scaling law. Meanwhile, ACC of Mistral-7B even exceeds that of the 13B model, which may be attributed to its inherently stronger semantic understanding apabilities(Jiang et al., 2023) .

4.3 Human evaluation of Helpfulness

Helpfulness (Bai et al., 2022) stresses that LLMs should provide accurate, useful information related to the question. Even if the model lacks full knowledge of a question, it should provide incomplete but

Models	ACC$_3$↑	F1$_3$↑	ECE$_3$↓	ACC$_5$↑	F1$_5$↑	ECE$_5$↓	ACC$_{11}$↑	F1$_{11}$↑	ECE$_{11}$↓
LLaMA2-7B	39.70	37.29	0.24	24.32	19.40	0.14	9.36	2.03	0.35
LLaMA2-13B	42.60	39.57	0.17	25.60	20.62	0.11	11.55	6.33	0.22
Mistral-7B	48.03	48.10	0.14	29.52	28.59	0.14	12.67	12.80	0.16

Table 5: Evaluation of CIT on 3 models with 3,5,11 classification classes.

related knowledge. We conduct the human evaluation of CIT, IDK-DPO and I-Prompt method, aiming to aiming to analyze the helpfulness of models trained by different methods. Details and instructions can be found in the Appendix F.

As shown in Figure 4, compared to IDK-DPO, CIT demonstrates a stronger ability to provide helpful information (the rate of CIT wins greater than losses). Moreover, CIT method even outperforms the vanilla model (I-Prompt), with the proportion of wins exceeding losses. This demonstrates that CIT prevents the model from defaulting to *I don't know* responses while preserving and even enhancing the ability to offer helpful suggestions. More case studies are presented in Table 10.

4.4 Confidence and Correct Steps Percentage

CRes	$1 \sim 3$	$4 \sim 6$	$7 \sim 9$
Pct.	0.29	0.58	0.84
LeM	10.00	35.00	65.00

Table 6: Average percentage (Pct.) of correct reasoning steps of total steps and generation lexical matching (LeM) for three uncertain confidence level.

Confidence of question is derived from sampling accuracy. For multi-step reasoning in mathematical problems, we are concerned whether confidence can be reflected within the reasoning path. In Table 6, we extracted sixty *uncertain* questions with the responses of CIT from GSM8K and conduct human evaluation. The analysis revealed a statistically positive correlation between the accuracy of intermediate reasoning steps and the confidence levels. This finding suggests that higher confidence scores are associated with more accurate step-by-step reasoning processes. Furthermore, higher accuracy in intermediate reasoning steps corresponds to improved lexical matching in the generated outputs. Details are in the Appendix G.

5 Related Work

5.1 Factual hallucinations in LLMs

Factual hallucination occurs that the LLM's response is inconsistent with reality. LLMs may confidently generate responses while being unaware of their own errors. (Huang et al., 2023; Ji et al., 2023). This may arise from both misinformation in pretraining data (Li et al., 2023a,b) and preference-aligned tuning methods.(Perez et al., 2022; Wei et al., 2024).

5.2 Verbalized Confidence Calibration

Verbalized Confidence Calibration is an important approach to mitigate factual hallucinations by ensuring that models express their confidence through language that reflects the probability of factual correctness (Tian et al., 2023b; Xiong et al., 2024). Previous work can be categorized into prompt-based,sampling-based and fine-tuning-based calibration methods. Prompt-based methods (Zhou et al., 2023) suffer from low accuracy, while sampling-based(Cole et al., 2023; Tian et al., 2023a) approaches require multiple generations, making them impractical for real-time applications. Fine-tuning-based methods train LLMs to express uncertainty through responses like *I don't know* when they lack relevant knowledge (Lin et al., 2022; Cheng et al., 2024). While these methods offer faster responses and higher accuracy, they tend to make LLMs habitually decline questions without providing any informative responses. This limitation stems from their binary classification schemes and oversimplified response patterns.

6 Conclusions

We propose Confidence Introspection Training to enable the LLM to discern what question they know, what they don't know and what they are uncertain about. Experimental results demonstrate that our proposed method not only enhances the model's reliability by accurately expressing its confidence, but also maintains the helpfulness of the vanilla model.

Limitations

The limitations of our work can be summarized in two main aspects. First, we mainly focus on enabling the LLM to generate its own assessments of confidence. Whether this approach can be combined with classification models remains a research topic worth exploring. Second, our experiments were conducted mainly on models with size of 7B and 13B. Due to resource constraints, we did not perform experiments on larger models.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. Preprint, arXiv:2212.08073.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? Preprint, arXiv:2401.13275.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. Preprint, arXiv:2310.03368.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. Preprint, arXiv:2305.14613.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. Preprint, arXiv:1706.04599.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. Preprint, arXiv:2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38.

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

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada. 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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. Preprint, arXiv:2207.05221.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:452–466.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In Findings of the Association for Computational Linguistics: ACL 2023, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel Tetreault, Eduard Hovy, and Heng Ji. 2023b. Defining a new NLP playground. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11932–11951, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. Transactions on Machine Learning Research.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. 2023. Chain of hindsight aligns language models with feedback. arXiv preprint arXiv:2302.02676.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. Preprint, arXiv:2406.13415.

Gerald E. Myers. 1986. Introspection and self-knowledge. American Philosophical Quarterly, 23(2):199–207.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations. Preprint, arXiv:2212.09251.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. Preprint, arXiv:2305.18290.

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

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. Preprint, arXiv:2305.14975.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-QA evaluation. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. Preprint, arXiv:2201.11903.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. Preprint, arXiv:2308.03958.

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.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. Preprint, arXiv:2403.18349.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi 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

Models	Dataset	0	1	2	3	4	5	6	7	8	9	10
Mistral-7B	TQA	38.1	4.7	3.0	2.9	3.3	8.1	6.5	6.3	6.7	9.2	15.2
	NQ	24.6	7.5	5.4	4.1	3.5	5.8	4.7	3.9	5.5	6.3	14.8
	GSM8K	26.1	16.5	11.5	9.9	8.6	8.2	5.4	4.2	4.5	2.9	2.4
LLaMA2-7B	TQA	24.6	6.5	5.1	4.9	5.5	6.3	5.6	6.5	8.8	11.5	14.8
	NQ	39.8	8.9	6.4	5.2	5.0	4.5	4.6	5.2	5.5	6.9	8.0
	GSM8K	40.9	22.3	12.2	8.4	5.6	3.7	3.3	1.7	1.0	0.6	0.2
LLaMA2-13B	TQA	21.5	5.6	4.1	3.5	3.8	5.1	3.7	4.4	7.2	14.0	27.0
	NQ	37.8	9.2	5.6	4.3	4.5	4.2	4.1	4.1	5.0	8.1	13.0
	GSM8K	61.3	20.4	10.0	4.1	1.8	1.2	0.6	0.5	0.1	0.0	0.0

Table 7: The proportion of uncertain questions with CRes ranging from 0 to 10 in the test sets of three datasets (test set): TriviaQA (TQA), Natural Questions (NQ), and GSM8K.

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Confidence distribution of Data

Table 7 shows the data distribution across various CRes values.

B Prefixes

Uncertain Prefixes :

- I'm not sure of the answer to that, but here's what I would suggest these.
- I can't give you a definitive answer, but I can offer some insights based on my knowledge.
- I'm not sure how to answer that, but here's what I think.

Unknown Prefixes:

- I don't know the answer to that, but I would suggest these.
- I don't have personal experience with that, but here are some general tips.
- I'm not knowledgeable about that topic, but I can try to find the answer for you.

C Prompts and details of baselines

C.1 Instructions of I-Prompt

The prompt we used for I-Prompt and testing other baselines is : "Please answer the following question, your answer should be as simple as possible. If you possess relevant knowledge, respond with confidence. If you are unfamiliar with the required knowledge, please honestly inform the user.Please answer this question:{question}"

For the GSM8K dataset, our prompt is: "Please solve the following math problem. If you possess relevant knowledge, respond with confidence. If you are unfamiliar with the required knowledge, please honestly inform the user your confidence.: {question}"

R-tuning adopts the original prompt from Zhang et al. (2024): "Question:{question} Answer: "

The prompt of Pf-SFT follows its input format mentioned in Appendix C.3.

C.2 Evaluation of I-Prompt

In addition to the prefixes designed in our experiment to indicate *unknown* or *uncertain*, if the following statements appear, they can also be considered as the model expressing of *unknown* or *uncertain*. These statements are summarized from the outputs of the vanilla model (Mistral-7B and LLaMA2), as shown in Table 8.

C.3 The details of PF-SFT

Pf-SFT needs data marking the correct and incorrect outputs with "GOOD" and "BAD" labels, respectively. Like this :

CRes	Statements
0	I am not an expert in the specific area.
1∼3	My Confidence: Low.
4∼6	I'm not entirely confident about this answer.
7∼9	I am 90% confident that.

Table 8: Statements representing the confidence level of the vanilla model.

- Question: What is the title of the French National Anthem?
- Label: *known*
- Expected answer: **GOOD**: The French National Anthem is called "La Marseillaise. **BAD**: I'm not sure how to answer that. My confidence in answer is moderate. The French National Anthem is called "La Marseillaise."

This is a data sample for Pf-SFT. Since the label is 'known,' we want the model to learn to give a direct answer rather than express uncertainty. We place the desired response after 'GOOD' and the incorrect response after 'BAD.' Since we have five class, there can be four different types of responses following 'BAD. The idea behind this method is to learn human preferences through SFT approach.

We use such data for SFT training. Compared to CIT(Ours), PF-SFT requires more complex data, but it does not require further training. During testing, we extract the content between 'GOOD' and 'BAD' as the model's answer, which it considers "GOOD" responses for the question.

It is possible that the relatively longer output with human preference made learning more difficult for the model, which is why the performance of Pf-SFT is not as good as CIT.

C.4 The details of S-Token

S-Token involves only classification, not generation. For each question, we first calculate the confidence type, then classify the generated response from I-Prompt (zero-shot, greedy decoding), and compute LeM-K and LeM-U.

C.5 The details of CEF and TOPK-S

Both methods require counting the frequency of the same answers. However, since LLMs rarely generate perfectly same outputs, we consider two sampled responses to express the same meaning if their cosine similarity exceeds 80%, where the cosine similarity is computed using Sentence-Transformers. Following (Cole et al., 2023), we set K to 1 and sampling times to 10.

C.6 The details of R-Tuning

During data construction, answers for *known* questions are generated by the model itself. For *uncertain* questions with 3 classes, we use our model's responses when correct, and golden answers when incorrect. For *unknown* question, we fully use the golden answers.

C.7 The details of IDK-DPO

We train the LLM to output the certain prefix (in Appendix B) and confidence level for *uncertain* and *unknown* questions while reserving direct answers for *known* ones.

D Statistics of CIs dataset

The data distribution in our CIS dataset is shown in Table 9. We use balanced data to construct the train, dev, and test sets. It is important to clarify that in the GSM8K dataset, after our sampling, the *known* class contains only 31 and 2 questions for Mistral-7B and LLaMA2-7B, respectively. In the *uncertain* class of LLaMA2-7B, there are also classes with fewer than 50 samples. Therefore, the test data for GSM8K is imbalanced. All balanced data selection is based on question classification, without any manual intervention

E Training Details

The β value for DPO is set 0.1, with 1 training epochs after SFT. During testing, we use the vllm[4] framework and employ greedy decoding for generation. All methods are evaluated on test set in our KIs dataset. LLaMA2

In actual training, we use sub-prefixes such as "My confidence is 0.2", "My confidence is 0.5" and "My confidence is 0.8" to represent the three confidence levels of the uncertain class (using the median of each CRes category divided by 10). Training in this way enables most methods to achieve better classification performance. After the model generates responses, we replace the sub-prefixes with those in Table 1 to help user better understand the model's confidence level. Sub-prefixes

[4]https://github.com/vllm-project/vllm

Dataset	TQA			NQ	GSM8K*
	Train	**Dev**	**Test**	**Test**	**Test**
			Mistral-7B-Instruct		
Total	50000	1500	5000	2500	631
N/class	10000	300	1000	500	150
			LLaMA2-7B-Chat		
Total	50000	1500	5000	1250	122
N/class	10000	300	1000	250	30

Table 9: Details of the train, dev and test set of CIs dataset. *Since the *known* category in the GSM8K dataset contains only 31 samples for Mistral and 2 samples for LLaMA2-7B. Moreover, in the *uncertain* class of Llama2-7B, there are also classes with fewer than 50 samples. the test data for GSM8K is imbalanced.

are simply symbolic of confidence levels. The final replacement does not affect model performance and makes the output more user-friendly.

F Helpfulness Evaluation

We present fifty question and responses from two methods side by side, asking three researchers specialized in natural language processing as evaluators to determine which response is more helpful for the given question or if there is no significant difference between the two. Evaluators need to consider differences of responses in terms of QA accuracy, relevance to the question and the richness of the information provided. The order of the responses is randomized to avoid position preference. The evaluation instructions are presented as follows.

Prompt Please evaluate the quality of these model responses based on the following criteria and clearly identify which model's response is better in terms of helpfulness.Helpfulness: When the model responds correctly, it should provide detailed information. When the responses are incorrect or when expressing uncertainty, the model should offer useful information, guesses or suggestions to help users solve question. For example, in the case of complex questions, the model may provide partial but relevant knowledge necessary to solve the problem. Question: {question}.Response A:{res_A}.esponse B:{res_B}. After evaluating responses A and B, select one of the following options: (1) the output of the first method is more helpful than the second; (2) the output of the second method is more helpful than the first; (3) there is no significant difference between the two outputs.

G Human Evaluation of Correct steps

We extracted sixty *uncertain* questions from the original GSM8K test set, along with the CIT method's responses to them, ensuring that these questions were correctly classified. Each confidence level includes twenty questions. We assigned the question-response pairs to three volunteers. Each volunteer was required to examine the proportion of correct reasoning steps to the total steps in the reasoning path. Afterward, they exchanged data and repeated the process. Here, one "step" refers to one calculation process (Wei et al., 2023). Finally, we calculate the average percentage of correct reasoning steps of total steps for each confidence level. The evaluation instructions are presented as follows.

Instruction The following is a math problem and its response. Carefully review each calculation step in the response for logical coherence and accuracy, determining whether each step is reasonable and correct. If an error in a previous step causes an error in the current calculation, the current step should also be marked incorrect. Finally, provide the percentage of correct calculation steps to the total number of steps in this response.

H Case studies of CIT and baselines

As shown in Table 10, the LLM trained with our Confidence Introspection Training (CIT) method is capable of providing information relevant to the question. Even with a confidence level of only 0.5 (for Question1), the responses can still include the correct answer. Compared to I-Prompt, our method maintains the fundamental function of the LLM to provide useful information to users and presents the model's confidence level.

Compared to IDK-DPO, it is evident that the CIT method provides substantially more information when facing Questions 2 and 3. Although the LLM does not know the golden answer for Question 3, the LLM trained with CIT still offers substantial helpful information, whereas IDK-DPO merely responds with "I don't know." This demonstrates that the CIT method can effectively prevent the model from becoming unhelpful.

Question1	Whose backing band is known as The Miami Sound Machine?	
Methods	**Responses**	**LeM**
I-Prompt	The Miami Sound Machine is the backing band for the singer Gloria Estefan. They gained international fame in the late 1980s and early 1990s with their fusion of Latin rhythms and pop music. I am confident in my answer.	✓
IDK-DPO	I'm not sure how to answer that.	✗
CIT	I'm not sure of the answer to that, but here's what I would suggest these.My confidence is 0.5.The Miami Sound Machine is a band that was formed in Miami, Florida, in the late 1960s. They gained international fame in the 1980s with their fusion of Latin rhythms, jazz, and pop music. The band's most famous member is probably lead singer Gloria Estefan. Other notable members include her husband, Emilio Estefan, who was the band's founder and producer, and drummer Enrique Rodriguez. The Miami Sound Machine disbanded in the late 1990s but reunited in 2006 for a world tour.	✓
Question2	Michelangelo Merisi (or Amerighi) is the birth name of which artist?	
Methods	**Responses**	**LeM**
I-Prompt	Caravaggio, Confidence: Low. The name Michelangelo Merisi (or Amerighi) is not associated with Caravaggio. Michelangelo Merisi da Caravaggio was an Italian painter active in the late 16th and early 17th centuries, best known for his dramatic use of chiaroscuro, a technique that employs strong contrasts between light and dark. Caravaggio's full name was actually Michelangelo Merisi da Caravaggio. However, I cannot confirm that the given name is the birth name of Caravaggio with absolute certainty. It is commonly accepted that this is his full name, but there may be some variation in the sources. I would recommend double-checking with a reliable art history reference for the most accurate information.	✓
IDK-DPO	Michelangelo Merisi is the birth name of the artist Caravaggio.	✓
CIT	Michelangelo Merisi, also known as Il Caravaggio, was an Italian painter. He is famous for his dramatic use of chiaroscuro, a technique that emphasizes the contrast between light and dark in a painting. Caravaggio's works often depict scenes of religious and mythological subjects, as well as everyday life in 16th-century Italy. Some of his most famous paintings include "The Calling of St. Matthew", "The Supper at Emmaus" and "The Entombment of Christ".	✓
Question3	What is the significance of 16th June 1904?	
Methods	**Responses**	**LeM**
I-Prompt	I am not confident in answering this question without further context. The significance of June 16, 1904, depends on the specific historical or cultural context in which it is being considered. It could be a significant date in the history of a particular country, organization, or individual. For example, it could be the date of a battle, a treaty signing, a birth or death, or an important event in the history of a particular field or industry. Without more information, it is not possible to determine the significance of this date with certainty.	✗
IDK-DPO	I'm not knowledgeable about that topic.	✗
CIT	I don't know the answer to that, but I would suggest these. The significance of June 16th, 1904, is that it marks the beginning of the Mexican Revolution. On this day, the Mexican Liberal Party, led by Francisco Madero, nominated him as their presidential candidate. Madero's election in November 1912 marked the end of the long dictatorship of Porfirio Diaz. However, Madero's presidency was short-lived as he was assassinated in February 1913. The Mexican Revolution, which lasted from 1910 to 1920, brought about significant political and social changes in Mexico.	✗

Table 10: The responses of Mistral-7B after our Confidence Introspection Training (CIT), IDK-DPO(Cheng et al., 2024) and I-Prompt. ✓ and ✗ denote whether the response contains the golden answer.