Teaching LLMs to Abstain via Fine-Grained Semantic Confidence Reward

Anonymous ACL submission

Abstract

Mitigating hallucinations in Large Language Models (LLMs) is critical for their reliable deployment. Existing methods typically fine-tune LLMs to abstain from answering questions beyond their knowledge scope. However, these methods often rely on coarse-grained signals to guide LLMs to abstain, such as overall confidence or uncertainty scores on multiple sampled answers, which may result in an imprecise awareness of the model's own knowledge boundaries. To this end, we propose a novel reinforcement learning framework built on **Fine-grained Semantic Confidence Reward (FiSCoRe)**, which guides LLMs to abstain via sample-specific confidence. Specifically, our method operates by sampling multiple candidate answers and conducting semantic clustering, then training the LLM to retain answers within high-confidence clusters and discard those within low-confidence ones, thereby promoting accurate post-hoc abstention. Additionally, we propose a new metric for evaluating the reliability of abstention fine-tuning tasks more comprehensively. Our method significantly enhances reliability in both in-domain and out-of-distribution benchmarks.

1 Introduction

Large language models (LLMs) have achieved remarkable success, demonstrating powerful capabilities in content generation, complex reasoning, and software development (Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2024). Their widespread adoption has established them as essential tools in numerous applications. However, a critical vulnerability threatens their reliability: their propensity to hallucinate, fabricating plausible but false information when faced with questions beyond their knowledge (Zhang et al., 2023). Such hallucinations erode user trust and can propagate misinformation. Reliance on fabricated content in high-stakes domains such as medicine or law can have severe consequences. Therefore, mitigating hallucinations is essential for the safe and responsible deployment of LLMs.

A promising direction for mitigating hallucination is to fine-tune LLMs to abstain from answering questions that lie beyond their knowledge scope and answer those within it (Wen et al., 2025; Cheng et al., 2024). Supervised methods often categorize a training set into "known" and "unknown" questions based on the correctness or accuracy of the model's answers, then train the model to answer the former and abstain from the latter (Zhang et al., 2024; Cheng et al., 2024). To avoid reliance on ground-truth labels, other approaches leverage uncertainty metrics such as semantic entropy to partition the training set into known and unknown questions (Zhang et al., 2024; Tjandra et al., 2024; Xue et al., 2025). However, these methods rely on coarse, global signals, such as a single correctness label or an aggregated uncertainty score, to guide abstention decisions. These approaches do not account for the nuanced confidence levels of individual samples, resulting in an ambiguous decision boundary. Consequently, models may either excessively abstain from answerable questions or fail to abstain when facing low-confidence inputs, which prevents the model from balancing truthfulness and helpfulness.

To overcome these limitations, we propose **Fine-grained Semantic Confidence Reward (FiSCoRe, /f́ɪs.kɔːr/)**, a novel reinforcement learning framework that replaces the conventional global uncertainty reward with a fine-grained, **per-sample** confidence reward. The core insight is that an answer's intrinsic confidence is reflected in its semantic consensus among all samples, and such confidence can be fully utilized for factual alignment. Although existing methods also sample multiple answers, they typically distill them into a single aggregated uncertainty score. In contrast, as shown in Figure 1, our approach rewards the model

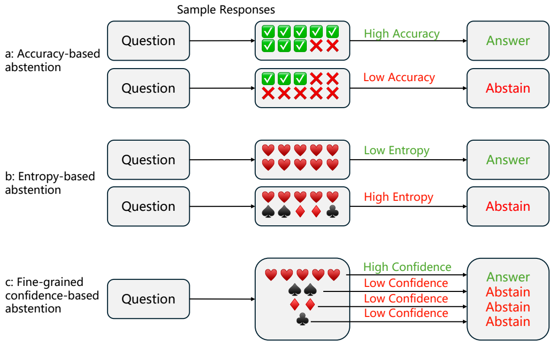

Figure 1: (a) Accuracy-based abstention. (b) Entropy-based abstention. (c) Ours: fine-grained confidence-based abstention.

for accurately assessing the confidence of each sample. Specifically, we first generate multiple candidate responses and group them into clusters based on their semantic equivalence. We use the cluster size as a proxy for the intrinsic confidence of its member samples: a large cluster implies high confidence, whereas a small cluster implies low confidence. A positive reward is granted only when the model's expressed confidence (e.g., "sure" or "unsure") aligns correctly with this intrinsic confidence level. This fine-grained signal incentivizes the model to align its verbalized confidence with semantic consensus, thus enabling it to learn a more nuanced knowledge boundary and make more accurate abstention decisions.

Existing studies indicate that helpfulness and truthfulness are like two ends of a seesaw — enhancing one often leads to a decline in the other (Zhu et al., 2025). We observe that current metrics fail to accurately reflect both aspects, making them inadequate for assessing abstention methods. To address this, we propose a reliability metric grounded in the model's self-awareness of its knowledge boundaries. This metric synthesizes helpfulness and truthfulness, aiming to prevent models from over-rejecting queries or generating hallucinations.

Our main contributions are as follows:

- We design a fine-grained semantic confidence reward (see Section 3.2), which significantly improves the reliability of LLMs.
- We propose a new metric (see Section 2.2) that integrates the helpfulness and truthfulness of LLMs into a single reliability metric, offering a comprehensive assessment of its awareness of its own knowledge boundaries.
- We experiment with various LLMs and test them on both in-distribution and out-of-distribution data. The results demonstrate that our approach outperforms baselines in most cases, particularly on out-of-distribution data, verifying the effectiveness and generalization capability of the proposed fine-grained semantic confidence reward.

R GT	Correctly answered	Wrongly answered	Abstained
Known	N_1	N_2	N_3
Unknown	–	N_4	N_5

Table 1: Abstention confusion matrix. R denotes the answer of the refined (fine-tuned) model. GT denotes the initial model. "Known" denotes that the initial model answered correctly, while "unknown" indicates that the initial model answered wrongly.

2 Problem Formulation

2.1 Abstention Fine-Tuning

To mitigate the risk of hallucination inherent in LLMs, a primary alignment strategy is to fine-tune models to selectively abstain from answering questions that fall beyond their knowledge boundaries (Wen et al., 2025; Li et al., 2025a). The objective is to produce a model that is both helpful, by correctly answering questions within its knowledge boundary, and truthful, by abstaining from those outside it. To formally measure progress towards this goal, we establish an evaluation framework that assesses the performance of a refined LLM relative to the initial LLM before fine-tuning. As illustrated in the abstention confusion matrix in Table 1, we partition the evaluation benchmark into two distinct categories based on the performance of the initial LLM:

- **Known Questions**: The set of questions that the initial LLM answers correctly.
- **Unknown Questions**: The set of questions that the initial LLM answers incorrectly.

The refined LLM is then evaluated on its ability to preserve accuracy on known questions while learning to abstain from unknown questions.

2.2 Reliability Evaluation

2.2.1 Helpfulness and Truthfulness

Helpfulness evaluates a model's capacity to preserve accuracy on questions within its known scope, preventing unnecessary abstention and forgetting. Truthfulness measures its ability to abstain from questions outside that scope, preventing fabricated responses. Following Kim et al. (2024); Feng et al. (2024), we adopt two fine-grained F1 metrics $\mathrm{F1}_{ans}$ for helpfulness evaluation and $\mathrm{F1}_{abs}$ for truthfulness evaluation:

$$\begin{aligned}\mathrm{F1}_{ans} &= \frac{2 \cdot \mathrm{Precision}_{ans} \cdot \mathrm{Recall}_{ans}}{\mathrm{Precision}_{ans} + \mathrm{Recall}_{ans}} \\ &= 2N_1/(2N_1 + 2N_2 + N_3 + N_4), \quad (1)\end{aligned}$$

$$\begin{aligned}\mathrm{F1}_{abs} &= \frac{2 \cdot \mathrm{Precision}_{abs} \cdot \mathrm{Recall}_{abs}}{\mathrm{Precision}_{abs} + \mathrm{Recall}_{abs}} \\ &= 2N_5/(N_3 + N_4 + 2N_5), \quad (2)\end{aligned}$$

2.2.2 Reliability Metric

Since increasing truthfulness often leads to a decrease in helpfulness, using either metric alone introduces bias. Therefore, we propose $\mathrm{F1}_{rel}$, defined as the harmonic mean of $\mathrm{F1}_{ans}$ and $\mathrm{F1}_{abs}$, which provides a holistic measure of reliability:

$$\mathrm{F1}_{rel} = \frac{2 \cdot \mathrm{F1}_{ans} \cdot \mathrm{F1}_{abs}}{\mathrm{F1}_{ans} + \mathrm{F1}_{abs}} = 4N_1N_5/(4N_1N_5 + 2N_2N_5 + N_1N_3 + N_1N_4 + N_3N_5 + N_4N_5). \quad (3)$$

The $\mathrm{F1}_{rel}$ metric holistically considers all categories in the abstention confusion matrix in Table 1 and exhibits the following desiderata: it monotonically increases with the correctly answered known questions (N_1) and correct abstention from unknown questions (N_5); it decreases monotonically with both types of error: incorrect answer (N_2, N_3) and over-abstention (N_4), and therefore is intuitively reasonable. In the ideal case ($N_2 = N_3 = N_4 = 0$, $N_1 + N_5 = N$), $\mathrm{F1}_{rel} = 1$; in the worst case ($N_1 = N_5 = 0$, $N_2 + N_3 + N_4 = N$), $\mathrm{F1}_{rel} = 0$, where N is the number of all questions. This metric effectively balances helpfulness and truthfulness, mitigating the risks of overly conservative or aggressive model behavior. A recent metric, reliability score (RS) (Xu et al., 2024a), has been proposed to achieve similar balancing effect, but we show that it is flawed in encouraging hallucination (see Appendix E.1).

3 Method

Inspired by the generalization capabilities of reinforcement learning (Chu et al., 2025) and its success in reasoning (Guo et al., 2025; Zuo et al., 2025), we incorporate abstention decisions into the rollout process and train LLMs via fine-grained confidence reward (FiSCoRe).

3.1 Group Relative Policy Optimization

We reward the model based on the sample-specific semantic confidence among a group of its own sampled responses, which requires a reinforcement learning algorithm capable of handling multiple samples per prompt. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) naturally fits this setting: it samples a group of G outputs (or *rollouts*) for each question x, which directly supports our semantic clustering–based confidence reward computation. Formally, for each question x, GRPO samples a group of G outputs $Y = \{y_1, y_2, \ldots, y_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and optimizes the new policy π_θ by maximizing the following objective:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\Big[\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \\ \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t} - \beta\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})\Big)\Big] \quad (4)\end{aligned}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta\left(y_{i,t}|q, y_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(y_{i,t}|q, y_{i,<t}\right)}$, $\hat{A}_{i,t}$ is the normalized advantage of the ith rollout in the current group, ϵ is the clipping ratio, and β is the KL loss coefficient. The GRPO pipeline is shown in Figure 2 (a).

3.2 Reward Function Design

Figure 2 (b) illustrates the proposed FiSCoRe. We first partition the responses using a proxy model that captures semantic equivalence, then introduce the confidence reward and auxiliary rewards.

3.2.1 Semantic Confidence Reward

Some studies employ reinforcement learning with confidence rewards to enhance the reasoning performance of LLMs (Zuo et al., 2025; Shafayat et al., 2025). We extend this idea to the abstention domain, which guides LLMs to answer when confident and abstain otherwise.

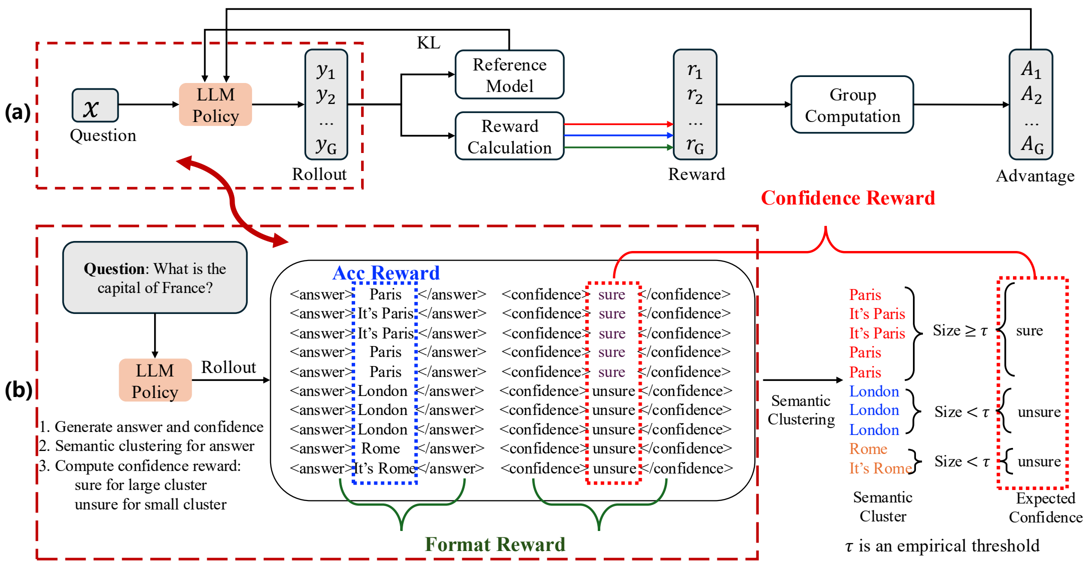
(a)
x
Question
LLM Policy
y_1 y_2 … y_G
Rollout
KL
Reference Model
Reward Calculation
r_1 r_2 … r_G
Reward
Group Computation
A_1 A_2 … A_G
Advantage
Confidence Reward
(b)
Question: What is the capital of France?
LLM Policy
Rollout
1. Generate answer and confidence
2. Semantic clustering for answer
3. Compute confidence reward:
sure for large cluster
unsure for small cluster
Acc Reward
<answer> Paris </answer> <confidence> sure </confidence>
<answer> It's Paris </answer> <confidence> sure </confidence>
<answer> It's Paris </answer> <confidence> sure </confidence>
<answer> Paris </answer> <confidence> sure </confidence>
<answer> Paris </answer> <confidence> sure </confidence>
<answer> London </answer> <confidence> unsure </confidence>
<answer> London </answer> <confidence> unsure </confidence>
<answer> London </answer> <confidence> unsure </confidence>
<answer> Rome </answer> <confidence> unsure </confidence>
<answer> It's Rome </answer> <confidence> unsure </confidence>
Format Reward
Semantic Clustering
Paris
It's Paris
It's Paris
Paris
Paris
London
London
London
Rome
It's Rome
Size ≥ τ
Size < τ
Size < τ
sure
unsure
unsure
Semantic Cluster
Expected Confidence
τ is an empirical threshold

Figure 2: The training overview of our method. (a) The GRPO pipeline. (b) A detailed example of how FISCORE works.

Semantic Clustering To compute the semantic confidence reward for individual rollouts, we group responses by semantic equivalence via semantic clustering (Kuhn et al., 2023). Equivalence is determined by bidirectional entailment using DeBERTa (He et al., 2021): two texts are semantically equivalent when they mutually entail each other. Following Kuhn et al. (2023), a greedy algorithm assigns each response to an existing cluster if it is semantically equivalent to any member of that cluster.

Semantic Confidence The formulation of FISCORE is based on the intuition that generative semantic consensus reflects model confidence. We assume that a larger semantic cluster implies strong consensus among the generated responses, reflecting higher confidence. Conversely, smaller or singleton clusters suggest divergent or less-supported meanings, thus implying lower confidence. Following this principle, we quantify the confidence of each rollout using a measure derived from the cardinality of the semantic cluster it belongs to. This per-sample approach distinguishes our method from work that uses semantic entropy, which aggregates all clusters into a single scalar and therefore can only produce one shared label for the entire group; using cluster size is precisely what makes per-sample labeling possible. Cluster size is nevertheless only a confidence proxy rather than a correctness oracle: a large cluster can still correspond to an incorrect answer. We therefore combine the confidence reward with an auxiliary accuracy reward, described below, to prevent the policy from collapsing toward high-consensus but wrong responses. For a given rollout y_i and its corresponding semantic cluster C_i, we define the target confidence label $\hat{U}_i$ according to semantic cluster size:

$$\hat{U}_i = \begin{cases} \text{sure}, & |C_i| \geq \tau, \\ \text{unsure}, & |C_i| < \tau. \end{cases} \tag{5}$$

The confidence reward is then defined as:

$$R_c = \mathbf{1}[U_i = \hat{U}_i], \tag{6}$$

Here, τ is an empirical threshold of abstention, which we set to $\lceil \frac{G}{2} \rceil$ in all experiments, G is the total number of rollouts and U is the model's expressed confidence, either "sure" or "unsure", extracted from the `<confidence> </confidence>` tag within the rollout. By incentivizing this alignment, the reward function effectively teaches the model a reliable abstention policy: to answer when semantic confidence is high and to abstain otherwise.

3.2.2 Auxiliary Rewards

Accuracy Reward. To prevent the model from engaging in reward hacking by learning to maximize the confidence reward on a uniform but incorrect answer, we introduce an auxiliary accuracy reward. This reward is granted only when the answer is semantically equivalent to the ground truth solution.

$$R_a = \begin{cases} 1, & \text{if the answer is correct} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where the answer is extracted between `<answer> </answer>` tags.

Format Reward. We encourage the model to enclose the answer within `<answer> </answer>` tags and to present the confidence enclosed within `<confidence> </confidence>` tags. The format reward R_f is composed of two components. For each correctly used tag (`<answer>`, `</answer>`, `<confidence>`, `</confidence>`) that appears in the rollout, we assign a reward of 0.125. To allow the model to determine its confidence based on its answer, it is necessary to place the answer before the confidence expression. We assign an additional reward of 0.5 if the tags appear in the correct order.

The total reward R_{total} is defined as a weighted sum of three distinct components:

$$R_{total} = w_c R_c + w_a R_a + w_f R_f, \tag{8}$$

where w_c, w_a, w_f are the respective weights for each reward component. Notably, only when the format is correct ($R_f = 1$) will the other two rewards be added.

4 Experiments

4.1 Experiment Setting

Datasets. We use four datasets of question-answering (QA) task for evaluation, including **Pararel** (Elazar et al., 2021) which contains cloze-style questions of relation prediction, **TriviaQA** (Joshi et al., 2017) which contains general knowledge QA pairs, **Natural Questions (NQ)** (Kwiatkowski et al., 2019) which contains questions from aggregated queries to Google Search, and **SciQ** (Welbl et al., 2017) which contains science exam questions. We use the **Pararel** dataset for training.

Baselines. We compare FISCORE with the following baselines.

1. **Prompting method**: In-Context Learning (ICL), In-Context Learning with refusal examples (ICL-IDK), and In-Context Learning with confidence expression (ICL-Unsure). We take the performance of ICL as the performance of the initial model.
2. **Supervised fine-tuning methods**: R-Tuning (Zhang et al., 2024), R-Tuning-U (Zhang et al., 2024), and SE-Tuning (Tjandra et al., 2024).
3. **RL-based method**: GRPO with semantic entropy (GRPO-SE), which uses the same training data as SE-Tuning and is built on GRPO.

Evaluation. We use four metrics: F1_{ans}, F1_{abs}, F1_{rel}, and accuracy (Acc) for evaluation, which are detailed in Section 2.2. We evaluate the correctness of the model's response by prompting with 6-shot examples. Following Xue et al. (2025), we adopt bidirectional string matching to evaluate answer correctness. Here, we treat a response as an abstention if the confidence part is unsure. More details about the evaluation are listed in Appendix B.

Implementation Details. We choose Llama3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024) in our experiments. The reinforcement learning framework is built on Open-R1 (Face, 2025). We set the reward weights $w_c = 1$, $w_a = 4$ and $w_f = 2$. We use Deberta-v2-xlarge-mnli (He et al., 2021) as the NLI model for semantic clustering. The sampling number is set to 10, and the threshold τ is set to 5. All experiments are implemented on Nvidia L40-48GB GPUs. The training epoch is set to 1. We set the sampling temperature to 1.0 during the GRPO training. During inference, we utilize the vLLM framework to accelerate the process and employ a greedy search strategy to generate responses. Hyperparameters and additional configurations are detailed in Appendix A and Appendix D .

4.2 Main Results

We present the main experimental results of FISCORE and all baseline methods in Table 2, grouped by backbone LLM. The results reveal several key findings:

Generalizable Reliability. We focus on the metric F1_{rel} that assesses a model's overall reliability by balancing helpfulness (F1_{ans}) and truthfulness (F1_{abs}). A key finding is that the high reliability demonstrated by some baselines on in-domain (ID) data is brittle and does not generalize. For example, R-Tuning using the Llama model achieves a high F1_{rel} score of 79.0 on the Pararel (ID) dataset by directly learning to abstain from unknown questions. However, this specialized performance collapses on out-of-distribution (OOD) tasks, with

Method	Pararel (ID)				TriviaQA (OOD)				NQ (OOD)				SciQ (OOD)			
	F1_{ans}	F1_{abs}	F1_{rel}	Acc	F1_{ans}	F1_{abs}	F1_{rel}	Acc	F1_{ans}	F1_{abs}	F1_{rel}	Acc	F1_{ans}	F1_{abs}	F1_{rel}	Acc
Llama3-8B-Instruct																
ICL	-	-	-	48.8	-	-	-	**68.4**	-	-	-	**37.1**	-	-	-	**65.3**
ICL-IDK	62.1	64.3	63.2	31.7	69.8	32.6	44.5	55.3	43.5	67.9	53.0	18.1	71.8	43.9	54.5	56.4
ICL-Unsure	70.4	68.2	69.3	41.9	67.7	46.4	55.0	51.7	45.9	18.0	25.9	30.4	64.7	17.9	28.0	53.1
R-Tuning	78.3	**79.6**	**79.0**	40.5	47.7	51.5	49.5	26.3	31.2	72.0	43.5	10.8	32.1	51.4	39.5	14.6
R-Tuning-U	73.8	76.6	75.2	40.5	42.6	53.2	47.3	20.4	31.8	**74.0**	44.5	10.3	49.9	49.7	49.8	28.0
SE-Tuning	78.2	75.0	76.6	45.9	44.9	50.7	47.7	24.0	27.7	72.2	40.0	9.4	22.4	50.4	31.0	9.2
GRPO-SE	70.9	77.3	73.9	35.1	70.5	60.2	65.0	47.3	37.2	73.9	49.5	13.0	64.4	51.2	57.0	46.4
FISCORE	**83.3**	67.1	74.3	**59.4**	**82.6**	**62.0**	**70.8**	61.7	**53.1**	72.5	**61.3**	21.7	**74.9**	**53.3**	**62.3**	52.4
Qwen2.5-7B-Instruct																
ICL	-	-	-	44.1	-	-	-	**51.7**	-	-	-	**27.6**	-	-	-	**68.8**
ICL-IDK	60.1	68.6	64.1	27.4	66.6	49.3	56.6	44.9	38.1	50.1	43.3	19.4	71.6	14.8	24.5	60.5
ICL-Unsure	77.7	82.9	80.2	35.8	71.2	51.7	59.9	49.8	40.2	39.7	39.9	21.9	71.4	37.6	49.2	57.3
R-Tuning	79.3	**86.4**	**82.7**	31.8	44.0	71.0	54.4	15.1	25.1	84.5	38.7	4.5	43.8	53.9	48.3	19.9
R-Tuning-U	62.1	80.5	70.1	21.3	40.1	69.8	51.0	13.6	32.3	**84.8**	46.7	6.3	43.9	52.5	47.8	20.4
SE-Tuning	71.4	82.3	76.5	28.0	33.8	69.3	45.4	10.7	32.8	84.2	47.2	6.7	55.4	52.7	54.0	34.5
GRPO-SE	75.2	83.3	79.0	31.7	71.6	72.4	72.0	41.6	42.8	78.0	55.3	14.2	76.7	50.7	61.0	57.3
FISCORE	**80.8**	72.9	76.6	**51.9**	**83.3**	**79.1**	**81.1**	51.3	**57.3**	80.1	**66.8**	20.4	**79.2**	**54.5**	**64.6**	58.0

Table 2: Performance on in-domain and out-of-domain knowledge-intensive QA datasets. All results are multiplied by 100. Notably, ICL is the initial model, so we do not calculate its F1 scores.

its score dropping to just 49.5 of F1_{rel} in TriviaQA. In contrast, the proposed FISCORE demonstrates a more robust, reliable performance that excels at generalization. It achieves this by attaining a superior balance between answering known questions (F1_{ans} of 83.3) and correctly abstaining (F1_{abs} of 67.1). This balanced approach leads to a high F1_{rel} score of 74.3 on the ID dataset, which shows a much smaller drop to 70.8 on the OOD TriviaQA dataset. This consistent outperformance on all three OOD datasets proves that the knowledge boundary learned by our approach is more generalizable. Compared to SE-Tuning, GRPO-SE shows stronger generalization: SE-Tuning performs well on ID data but degrades significantly on OOD sets, while GRPO-SE maintains robust performance across both, highlighting the advantage of reinforcement learning over supervised fine-tuning.

Effectiveness of the Fine-grained Semantic Confidence Reward. We compare our approach directly with GRPO-SE, a baseline employing the same GRPO algorithm but utilizing a coarse semantic uncertainty reward. The results demonstrate a clear advantage: our method FISCORE significantly outperforms GRPO-SE across all OOD datasets and different LLMs in terms of F1_{rel}. Given that the primary distinction between the two methods is the reward function, this comparison provides empirical evidence that our fine-grained confidence reward is substantially more effective than a coarse, global entropy-based signal.

Comparison between Different LLMs. We fine-tuned both the Llama3-8B-Instruct and Qwen2.5-7B-Instruct models on the Pararel dataset using our FISCORE method alongside established baseline abstention techniques. The results on TriviaQA and NQ show that although Llama achieved higher raw accuracy than Qwen both before and after fine-tuning, its overall reliability (measured by F1_{rel}) remained lower. This disparity indicates that a model's extensive knowledge capacity does not ensure precise self-awareness of its knowledge boundaries. It also highlights that accuracy alone provides an inadequate proxy for gauging improvements in abstention alignment.

4.3 Ablation Study

Figure 3 summarizes the ablation results on the sampling number G and accuracy reward weight w_a. **Impact of Sampling Number.** We investigate the impact of the sampling number G on performance in Figure 3 (a). The results demonstrate that for FISCORE, both reliability (F1_{rel}) and accuracy improve as G increases, consistently outperforming the GRPO-SE baseline. Both metrics reach their optimal or near-optimal levels at $G = 10$. Since

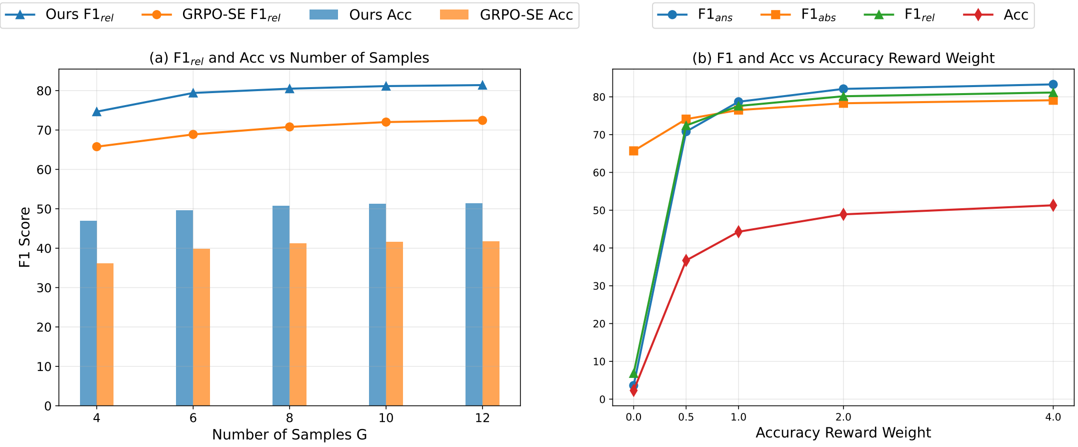

Figure 3: (a) Experiments of F1$_{rel}$ and accuracy of various sampling number G on TriviaQA on Qwen2.5-7B-Instruct. (b) Experiments of F1$_{ans}$, F1$_{abs}$ and F1$_{rel}$ of various accuracy reward weight w_a on TriviaQA on Qwen2.5-7B-Instruct.

performance gains become negligible beyond this point, we select $G = 10$ as the default setting for all experiments to maximize performance while maintaining computational efficiency.

Impact of Accuracy Reward Weight. We evaluate the impact of the accuracy reward weight w_a while fixing the confidence reward weight w_c at 1, as shown in Figure 3 (b). When w_a is set to 0, the model achieves a high F1$_{abs}$ but a near-zero F1$_{ans}$, and its Acc drops to only 2.3. This indicates that the model abstains from almost all questions. This occurs because the model receives rewards for generating low-confidence incorrect answers, a phenomenon known as reward hacking. Shafayat et al. (2025) demonstrate similar results: relying on self-confidence as the only reward leads to model collapse in the later stages of training. Therefore, it is necessary to combine accuracy reward and confidence reward for training. Introducing the accuracy reward improves F1$_{rel}$. As w_a increases from 0 to 4, Acc rises monotonically from 2.3 to 36.7, 44.3, 48.9, and 51.3, while F1$_{rel}$ also increases from 6.8 to 72.4, 77.6, 80.1, and 81.1. Notably, F1$_{abs}$ improves over the same range, showing that the gain is not merely caused by answering more questions, but by learning a better abstention boundary. When w_a is further increased to 8, Acc slightly increases to 51.7, but F1$_{abs}$ drops to 65.9 and F1$_{rel}$ decreases to 71.1, indicating an over-aggressive answering policy. These results show that the accuracy reward is essential for preventing collapse, but overly large accuracy weights can sacrifice truthfulness.

τ	F1$_{ans}$	F1$_{abs}$	F1$_{rel}$	Acc
3	79.3	65.6	71.8	54.7
4	82.4	75.4	78.8	53.0
5	83.3	79.1	81.1	51.3
6	82.1	78.3	80.1	49.9
7	74.7	81.2	77.8	36.5

Table 3: Ablation on threshold.

Setting	F1$_{ans}$	F1$_{abs}$	F1$_{rel}$	Acc
FiScoRe	83.3	79.1	81.1	51.3
w/o Format	71.5	43.6	54.2	51.5
w/o Confidence	69.5	2.1	4.1	56.1
w/ Continuous	83.0	82.9	82.9	49.1

Table 4: Ablation on reward functions.

Impact of Abstention Threshold. We conduct ablation experiments on the TriviaQA dataset using the Qwen2.5-7B-Instruct model to investigate the impact of the abstention threshold τ. As shown in Table 3, F1$_{abs}$ increases with rising τ, while F1$_{ans}$ and Acc exhibit a downward trend. At $\tau = 5$, truthfulness and helpfulness achieve a balanced level, yielding the maximum F1$_{rel}$. Therefore, different thresholds can be employed for training across scenarios with varying tolerance levels for hallucinations.

Ablation on Reward Functions. Table 4 presents the results of reward function ablation experiments conducted on the TriviaQA dataset using the Qwen2.5-7B-Instruct model. We first validate the necessity of the format reward. We remove the format reward and observe a decrease in $F1_{ans}$, a significant drop in $F1_{abs}$, while accuracy remains almost unaffected. The reason lies in semantic clustering: we extract answers within the answer tags using rule-based methods. The lack of format reward leads to inaccurate tags, which degrade semantic clustering performance and ultimately increase confidence bias. To validate the effectiveness of the proposed fine-grained semantic confidence reward, we removed this reward and observed that while accuracy remained stable, $F1_{abs}$ dropped to near zero. This demonstrates that the model possesses a certain degree of confidence awareness capability, and our confidence reward can significantly enhance this ability. Finally, we train Qwen2.5-7B-Instruct with continuous confidence rewards, where the model samples 10 candidate answers and predicts confidence scores scaled from 10% to 100%. We find that continuous confidence rewards achieve better reliability than binary discrete rewards. Continuous rewards do not require retraining when adapting the abstention threshold, resulting in lower applied costs.

5 Related Work

Uncertainty Estimation and Confidence Elicitation. Uncertainty estimation aims to quantify this risk before LLMs respond to a request, which is an effective tool for detecting and mitigating hallucinations of LLMs (Xia et al., 2025). Likelihood-based methods estimate uncertainty by computing probabilities for salient tokens, capturing the model's confidence in specific outputs (Duan et al., 2024; Lin et al., 2024). Sampling-based methods generate multiple candidate responses, cluster them, and quantify uncertainty based on the degree of inter-response consistency (Kuhn et al., 2023; Qiu and Miikkulainen, 2024a; Nikitin et al., 2024; Aichberger et al., 2025). Probing-based methods leverage internal model representations to train classifiers that predict uncertainty or detect hallucinations, often without requiring additional generation (Han et al., 2024; Marks and Tegmark, 2024; Gottesman and Geva, 2024; Azaria and Mitchell, 2023). Confidence elicitation aims to elicit the confidence of LLMs in their answers (Geng et al., 2024). Prompting methods directly prompt LLMs to express confidence (Lin et al., 2022; Kadavath et al., 2022). Recent studies (Xu et al., 2024b; Damani et al., 2025) employ chain-of-thought (CoT) training to teach LLMs to reason about their confidence. Zheng et al. (2025) adopt binary confidence elicitation ("sure" vs. "unsure") as a proxy for abstention, enabling models to discard low-confidence answers post hoc. In this work, we introduce a fine-grained semantic confidence reward to cultivate a more nuanced self-awareness of the model's confidence.

Abstention Fine-Tuning. Many studies explore fine-tuning LLMs to abstain from answering questions beyond their knowledge boundaries (Wen et al., 2025). Brahman et al. (2024); Cheng et al. (2024); Xu et al. (2024a) propose learning from preferences via direct preference optimization (Rafailov et al., 2023) to train LLMs to abstain from answering unknown questions. Li et al. (2025b) introduce adaptive contrastive preference learning to calibrate abstention behavior. Zhang et al. (2024); Tjandra et al. (2024); Xue et al. (2025) leverage uncertainty to construct abstention-aware datasets, replacing the ground-truth labels of uncertain questions with "I don't know". Yang et al. (2025); Deng et al. (2024) train LLMs to reason about their own uncertainty and provide post-hoc explanations for why a given question is unanswerable. Huang et al. (2025); Cohen et al. (2024) introduce a special "rejection" token into the model's vocabulary and design an objective function that reallocates probability mass toward the "rejection" token when the model encounters uncertain predictions. Anonymous (2026) perform reliability alignment in the truthful subspace. In this work, we go beyond aggregated uncertainty by utilizing fine-grained, per-sample confidence to sharpen the model's awareness of its knowledge limits.

6 Conclusion

In this paper, we enhance the reliability of LLMs by enabling them to answer questions within their knowledge boundaries and abstain from those beyond. We propose $F1_{rel}$, a holistic reliability metric that integrates helpfulness and truthfulness into a single score monotonically sensitive to all error categories. We then introduce FISCORE, which derives fine-grained per-sample confidence from semantic clustering and trains the model to abstain when generative consensus is low. Experiments on in-domain and out-of-distribution datasets show that FISCORE achieves a superior balance between helpfulness and truthfulness, improving overall reliability.

Limitations

Our method has some limitations. Eliciting the answer and confidence simultaneously can introduce a possible cognitive bias (Xu et al., 2025); future research should decouple the accuracy reward from the confidence reward (Liu et al., 2026). First, we must manually set the threshold τ during GRPO training. Second, the $\mathrm{F1}_{ans}$, $\mathrm{F1}_{abs}$, and $\mathrm{F1}_{rel}$ metrics are only suitable for evaluating alignment methods, which means that we must know the knowledge that the model possesses before alignment. $\mathrm{F1}_{rel}$ can complicate cross-model comparability. Future evaluation metrics for LLMs should encourage models to acknowledge their knowledge gaps rather than guessing answers. Due to computational resource constraints, our experiments are limited to 7B/8B-scale models and a single training domain (Pararel). We leave the exploration of larger models and more diverse training corpora to future work. In addition, our main method adopts binary confidence ("sure" vs. "unsure") to ensure a fair comparison with all baselines, which also rely on binary abstention decisions. Although our ablation study shows that a continuous confidence reward can yield further improvements, using it as the primary method would make direct comparisons with existing baselines less controlled. We regard the continuous reward as a promising direction for future work.

References

Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. 2025. Sindex: Semantic inconsistency index for hallucination detection in llms. *Preprint*, arXiv:2503.05980.

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

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*.

Anonymous. 2026. Learning to abstain via hallucination-subspace alignment. Under review.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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? In *Forty-first International Conference on Machine Learning*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [idk] token. In *Advances in Neural Information Processing Systems*, volume 37, pages 10935–10958. Curran Associates, Inc.

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

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't just say "I don't know"! self-aligning large language models for responding to unknown questions with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673, Miami, Florida, USA. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024.

Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Daniela Gottesman and Mor Geva. 2024. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. 2025. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, Vienna, Austria. Association for Computational Linguistics.

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, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. 2025. When to speak, when to abstain: Contrastive decoding with abstention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, Vienna, Austria. Association for Computational Linguistics.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1989–2007, Miami, Florida, USA. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025a. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157, Vienna, Austria. Association for Computational Linguistics.

Yinghui Li, Haojing Huang, Jiayi Kuang, Yangning Li, Shu-Yu Guo, Chao Qu, Xiaoyu Tan, Hai-Tao Zheng, Ying Shen, and Philip S. Yu. 2025b. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations*.

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

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

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *Preprint*, arXiv:2601.05242.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xin Qiu and Risto Miikkulainen. 2024a. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xin Qiu and Risto Miikkulainen. 2024b. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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*, volume 36, pages 53728–53741. Curran Associates, Inc.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can large reasoning models self-train? *Preprint*, arXiv:2505.21444.

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

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. 2024. Fine-tuning large language models to appropriately abstain with semantic entropy. In *Neurips Safe Generative AI Workshop 2024*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

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*, pages 21381–21396, Vienna, Austria. Association for Computational Linguistics.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025. Do language models mirror human confidence? exploring psychological insights to address overconfidence in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25655–25672, Vienna, Austria. Association for Computational Linguistics.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

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

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025. UAlign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024, Vienna, Austria. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Junxiao Yang, Jinzhe Tu, Haoran Liu, Xiaoce Wang, Chujie Zheng, Zhexin Zhang, Shiyao Cui, Caishun Chen, Tiantian He, Hongning Wang, and 1 others. 2025. Barrel: Boundary-aware reasoning for factual and reliable lrms. *arXiv preprint arXiv:2505.13529*.

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 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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Hang Zheng, Hongshen Xu, Yuncong Liu, Shuai Fan, Lu Chen, Pascale Fung, and Kai Yu. 2025. Enhancing LLM reliability via explicit knowledge boundary modeling. In *Second Conference on Language Modeling*.

Runchuan Zhu, Xinke Jiang, Jiang Wu, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. 2025. GRAIT: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4006–4021, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. Ttrl: Test-time reinforcement learning. *Preprint*, arXiv:2504.16084.

A Experiment Details

A.1 Dataset

We test on the following publicly available datasets.

- **Pararel** (Elazar et al., 2021) is a dataset of factual knowledge and relations that are originally for evaluating masked language models. We follow Zhang et al. (2024), changing it into a question-answering format for generation evaluation. The validation set contains 5,584 QA pairs, which we use for evaluation.

- **TriviaQA** (Joshi et al., 2017) is a large-scale reading comprehension dataset containing over 95,000 trivia-style question–answer pairs. The validation set contains 11,313 QA pairs, which we use for evaluation.

- **Natural Questions (NQ)** (Kwiatkowski et al., 2019) are aggregated queries issued to the Google search engine. All questions of NQ can be answered using the contents of the English Wikipedia. The validation set contains 3,610 QA pairs, which we use for evaluation.

- **SciQ** (Welbl et al., 2017) is a crowdsourced dataset of 13,679 science questions. It covers topics from biology, chemistry, physics, and earth science from real educational materials. The validation set contains 1,000 QA pairs, which we use for evaluation.

A.2 Details of baselines

The core idea of existing abstention methods is to fine-tune LLMs to answer the questions within their knowledge scope and abstain from those outside it. It consists of two steps.

First, given a training dataset D, these methods partition it into two subsets, the answerable set D_0 and the unanswerable set D_1. For questions in D_0, the model is assumed to be confident about it, and the ground-truth answer is set to be the model’s standard response. For questions in D_1, the ground-truth text, it is changed to an abstention expression: “I don’t know”. R-Tuning (Zhang et al., 2024) prompts LLMs to answer questions from the training set, then classifies correctly answered questions into D_0 and incorrectly answered ones into D_1. R-Tuning-U (Zhang et al., 2024) and SE-Tuning (Tjandra et al., 2024) use uncertainty to partition the training set. R-Tuning adopts predictive entropy, which clusters lexical equivalent answers together. SE-Tuning adopts semantic entropy, which clusters semantically equivalent answers together. Let $C_1, ..., C_m$ be the classes that were clustered from sampled responses; the entropy is given by:

$$\begin{aligned} Entropy(\mathbf{x}) &\approx -\sum_{i=1}^{m} p_\theta(C_i \mid \mathbf{x}) \log p_\theta(C_i \mid \mathbf{x}) \\ &\approx -\sum_{i=1}^{m} \left(\frac{|C_i|}{M}\right) \log \left(\frac{|C_i|}{M}\right), \quad (9) \end{aligned}$$

where M is the total number of samples. Then, D_0 and D_1 are obtained by a threshold τ. Then, LLMs are fine-tuned on D_0 and D_1, learning to minimize the following objective:

$$\mathcal{L}(p_\theta) = -\sum_{x \in D_0, D_1} \sum_{t=1}^{|\mathbf{y}^{(x)}|} \log p_\theta(y_t^{(x)} \mid \mathrm{P}, \mathbf{x}, \mathbf{y}_{<t}^{(x)}), \quad (10)$$

where P is the prompt, $\mathbf{x}$ is the question, and $\mathbf{y}$ is the modified ground-truth text. For consistency, GRPO-SE is trained on the same processed dataset as SE-Tuning. Its objective function is Equation (4), enabling a controlled comparison of reward granularity and abstention behavior.

A.2.1 Details of GRPO-SE

GRPO-SE is an RL-based baseline that employs the GRPO framework with a coarse, global semantic entropy reward. It is trained on the same processed dataset as SE-Tuning, enabling a controlled comparison between RL and SFT approaches under identical data conditions.

Reward Function. Unlike FISCORE, which assigns a *per-sample* confidence label derived from the size of each rollout's semantic cluster, GRPO-SE computes a single semantic entropy score over all G rollouts and assigns one shared confidence label to the entire group. Concretely, given the G rollouts $\{y_1, \ldots, y_G\}$ for question x, we obtain semantic clusters $C_1, \ldots, C_m$ via the same NLI-based greedy algorithm described in Section 3.2. The semantic entropy is then computed as:

$$H(x) = -\sum_{k=1}^{m} \frac{|C_k|}{G} \log \frac{|C_k|}{G}. \tag{11}$$

A global binary confidence label $\hat{U}$ is assigned to all rollouts based on a threshold τ_H:

$$\hat{U} = \begin{cases} \text{sure}, & H(x) < \tau_H, \\ \text{unsure}, & H(x) \geq \tau_H. \end{cases} \tag{12}$$

The confidence reward for the i-th rollout is:

$$R_{\text{SE}} = \mathbf{1}[U_i = \hat{U}], \tag{13}$$

where U_i is the confidence expression ("sure" or "unsure") extracted from rollout y_i. The auxiliary accuracy reward R_a and format reward R_f are identical to those used in FISCORE (Equation 7 and Section 3.2). The total reward follows the same weighted sum as in Equation 8.

Key Distinction from FISCORE. The critical difference lies in the granularity of the confidence signal. In GRPO-SE, all G rollouts share the same label $\hat{U}$, determined by a single group-level entropy score. In contrast, FISCORE assigns a distinct label $\hat{U}_i$ to each rollout based on the cardinality of the semantic cluster it belongs to. This per-sample labeling enables FISCORE to provide richer, more discriminative training signals: within the same rollout group, responses falling in a large cluster are labeled "sure" while those in small or singleton clusters are labeled "unsure". This fine-grained signal helps the model learn more nuanced knowledge boundaries, as validated by the experimental results in Section 3.2.

A.3 Training Details

We use 15,000 samples from the Pararel dataset for training. Before training, we remove questions with semantic entropy below 1. These low-entropy samples lack sufficient semantic diversity, which can bias the binary confidence reward and lead the model to abstain too infrequently. Moreover, their simplicity diminishes the relative advantage signal in GRPO training, reducing the effectiveness of policy optimization. In the semantic clustering, we concatenate the question and the answer as the input of the NLI model. We run each experiment with three different random seeds and report the mean performance.

A.4 Training Template

You are a helpful and truthful AI assistant. Given a question, you should provide responses that include answer and confidence. You first answer the question as briefly as possible, enclosed by `<answer>` and `</answer>`, and then provide your confidence in sure or unsure about the answer, enclosed by `<confidence>` and `</confidence>`. Respond in the following format:
`<answer>` ... `</answer>`
`<confidence>` sure or unsure `</confidence>`
Question:

Table 5: Training template. The question is appended at the end during training and inference.

The training template is shown in Table 5. As we describe the confidence reward in Figure 2 (b), it is essential for the policy LLM to generate responses that conform to the specified format. To facilitate this, we design a prompt template to guide the policy model in producing an answer along with an associated confidence level. As shown in Table 5, the model is instructed to format its responses into two distinct parts. First, the model provides the answer wrapped in the `<answer> </answer>` tag. Second, the model expresses a binary confidence level (sure or unsure) wrapped in the `<confidence> </confidence>` tag. This `<answer> + <confidence>` structure forces the model to reflect on its confidence after providing an answer. In this work, we interpret an unsure confidence as an act of abstention.

B Evaluation Prompt

We use a 6-shot prompt across all methods. For the ICL baseline, the prompt consists of six examples of direct answering. In contrast, prompts for abstention-aware methods include a balanced mix of three answering examples and three abstention examples.

ICL Prompt

Answer the following questions as briefly as possible.

Question: `{demo question 1}`
Answer: `{demo answer 1}`

Question: `{demo question 2}`
Answer: `{demo answer 2}`
...
Question: `{input question}`
Answer:

ICL-IDK Prompt

Answer the following questions as briefly as possible. If you don't know the answer, please simply say "I don't know."

Question: `{demo question 1}`
Answer: `{demo answer 1}`

Question: `{demo question 2}`
Answer: `I don't know.`
...
Question: `{input question}`
Answer:

ICL-Unsure Prompt

You are a helpful and truthful AI Assistant. You should answer the question as briefly as possible, then present your confidence. If you are sure about your answer, please say "I am sure" after your answer; otherwise, say "I am unsure".

Question: `{demo question 1}`
Answer: `{demo answer 1}` I am sure.

Question: `{demo question 2}`
Answer: `{demo answer 2}` I am unsure.

...

Question: `{input_question}`
Answer:

R-Tuning, R-Tuning-U and SE-Tuning Prompt

Answer the following questions as briefly as possible. If you don't know the answer, please simply say "I don't know."

Question: `{demo question 1}`
Answer: `{demo answer 1}`

Question: `{demo question 2}`
Answer: `I don't know.`

...

Question: `{input question}`
Answer:

GRPO-SE and FiSCoRe Prompt

You are a helpful and truthful AI assistant that provides responses that include answers and confidence. You first answer the question as briefly as possible enclosed by `<answer>` and `</answer>` and then provide your confidence in sure or unsure about the answer enclosed by `<confidence>` and `</confidence>`. Respond in the following format:
`<answer>...</answer>`
`<confidence>sure/ unsure</confidence>`

Question: `{question 1}`
Answer: `<answer>` `{answer 1}` `</answer>`
`<confidence> sure </confidence>`

Question: `{question 2}`
Answer: `<answer>` `{answer 2}` `</answer>`
`<confidence> unsure </confidence>`

...
Question: `{input question}`
Answer:

C Details about semantic clustering

The Deberta-v2-xlarge-mnli (He et al., 2021) is fine-tuned on the MNLI dataset, which includes 392K premise-hypothesis pairs. For a fair comparison, SE-Tuning, GRPO-SE and our method all use the Deberta-v2-xlarge-mnli. We selected 40 questions from Pararel and TriviaQA respectively, and for each question we sampled ten examples for manual verification. We found that Deberta-v2-xlarge-mnli achieved accuracies of 98% and 96% on these datasets.

The greedy semantic clustering algorithm is parameter-free once the NLI equivalence predicate is fixed, so the main potential source of sensitivity is the choice of NLI model. To evaluate this factor, we replace Deberta-v2-xlarge-mnli with smaller DeBERTa variants and train FiSCoRe on Qwen2.5-7B-Instruct using the same setup. As shown in Table 6, performance is stable across NLI model sizes: even Deberta-base-mnli only reduces F1_{rel} by 0.7 points compared with Deberta-v2-xlarge-mnli. This suggests that FiSCoRe does not require a large or specialized NLI model for semantic clustering.

NLI	**F1**$_{ans}$	**F1**$_{abs}$	**F1**$_{rel}$	**Acc**
XLarge	83.3	79.1	81.1	51.3
Large	83.0	78.9	80.9	51.1
Base	82.5	78.4	80.4	50.6

Table 6: Sensitivity analysis of the DeBERTa NLI model used for semantic clustering on TriviaQA with Qwen2.5-7B-Instruct.

Fine-tuning the DeBERTa model on manually annotated data may yield better results, but this process is labor-intensive. In addition, a small amount of one-way entailment answers may also impact performance. For example, for the question "What field does Robert Bunsen work in¿‘, DeBERTa classifies the answers "science" and "chemistry" into two different categories. However, they have a hierarchical relationship and are not suitable for grouping together into the same cluster, or separating completely. Hierarchical clustering (Abdaljalil et al., 2025) and finer-grained semantic classification (Qiu and Miikkulainen, 2024b) can mitigate this issue.

D Hyperparameters

For FiSCoRe on Llama3-8B-Instruct and Qwen2.5-7B-Instruct, we use the same hyperparameter setups, which are shown in Table 7.

Parameter	**Value**
General Settings	
bf16	true
use_vllm	true
vllm_device	auto
vllm_enforce_eager	true
vllm_gpu_memory_utilization	0.7
vllm_max_model_len	4608
do_eval	false
Training Configuration	
gradient_accumulation_steps	4
gradient_checkpointing	true
gradient_checkpointing_kwargs	use_reentrant: false
learning_rate	3.0e-06
lr_scheduler_type	cosine_with_min_lr
lr_scheduler_kwargs	min_lr_rate: 0.1
warmup_ratio	0.05
num_train_epochs	1
per_device_train_batch_size	10
per_device_eval_batch_size	10
threshold of abstention	5
Generation Settings	
max_prompt_length	256
max_completion_length	64
num_generations	10
temperature	1.0
seed	42, 32, 22
Logging and Saving	
log_completions	true
log_level	info
logging_first_step	true
logging_steps	1
logging_strategy	steps
save_strategy	steps
save_steps	50
report_to	wandb
Reward Weights	
confidence	1.0
format	2.0
accuracy	4.0

Table 7: Hyperparameter setups for the GRPO trainer.

E Reliability Metric

E.1 Analysis for Reliability Score

Xu et al. (2024a) propose the Reliability Score (RS), which encourages LLMs to offer maximal assistance while minimizing hallucinated errors. RS is defined as a weighted sum of accuracy and truthful rate, which is used by current studies (Xu et al., 2024a; Kim et al., 2025; Yang et al., 2025): $\text{RS}(\alpha) = \alpha \times \text{Truth.} + (1-\alpha) \times \text{Acc.}$, the weight α is the answering rate.

The components are:

$$N = N_1 + N_2 + N_3 + N_4 + N_5$$

$$\text{Acc.} = \frac{N_1}{N}$$

$$\text{Truth.} = \frac{N_1 + N_3 + N_5}{N}$$

$$\alpha = 1 - \frac{N_3 + N_5}{N} = \frac{N_1 + N_2 + N_4}{N}$$

Substituting these into the RS formula:

$$\begin{aligned} RS &= \left(\frac{N_1 + N_2 + N_4}{N}\right)\left(\frac{N_1 + N_3 + N_5}{N}\right) + \left(1 - \frac{N_1 + N_2 + N_4}{N}\right)\left(\frac{N_1}{N}\right) \\ &= \left(\frac{N_1 + N_2 + N_4}{N}\right)\left(\frac{N_1 + N_3 + N_5}{N}\right) + \left(\frac{N_3 + N_5}{N}\right)\left(\frac{N_1}{N}\right) \\ &= \frac{(N_1 + N_2 + N_4)(N_1 + N_3 + N_5) + N_1(N_3 + N_5)}{N^2} \end{aligned}$$

The partial derivative of RS with respect to N_4 is:

$$\frac{\partial \text{RS}}{\partial N_4} = \frac{N_3 + N_5 - (N_1 + N_2 + N_4)}{N^2} \tag{14}$$

The sign of this derivative is variable and contingent on the specific distribution of counts among the categories. It is not guaranteed to be negative. For instance, consider a perfect model evaluated on a dataset composed solely of unanswerable questions ($N_1 = N_2 = N_3 = 0$, with an initial state of $N_4 = 0$ and $N_5 = N$). The derivative at this point is:

$$\left.\frac{\partial \text{RS}}{\partial N_4}\right|_{N_4=0} = \frac{N_5 - N_4}{N^2} = \frac{N - 0}{N^2} = \frac{1}{N} > 0 \tag{15}$$

A positive derivative demonstrates that RS **perversely incentivizes** a perfect model to start making mistakes. An initial increase in N_4 from zero leads to an *increase* in the RS score. This behavior is antithetical to the fundamental goal of a reliability metric.

E.2 Analysis for the proposed reliability F1

The proposed reliability F1_{rel} is defined as:

$$\text{F1}_{rel} = \frac{2}{\frac{1}{\text{F1}_{ans}} + \frac{1}{\text{F1}_{abs}}}$$

First, we calculate the sum of the reciprocals:

$$\begin{aligned} \frac{1}{\text{F1}_{ans}} + \frac{1}{\text{F1}_{abs}} &= \frac{2N_1 + 2N_2 + N_3 + N_4}{2N_1} + \frac{N_3 + N_4 + 2N_5}{2N_5} \\ &= \frac{N_5(2N_1 + 2N_2 + N_3 + N_4) + N_1(N_3 + N_4 + 2N_5)}{2N_1N_5} \\ &= \frac{2N_1N_5 + 2N_2N_5 + N_3N_5 + N_4N_5 + N_1N_3 + N_1N_4 + 2N_1N_5}{2N_1N_5} \\ &= \frac{4N_1N_5 + 2N_2N_5 + N_1N_3 + N_1N_4 + N_3N_5 + N_4N_5}{2N_1N_5} \end{aligned}$$

Finally, substitute this result back into the formula for $\text{F1}_{rel.}$:

$$\begin{aligned}\text{F1}_{rel} &= \frac{2}{\frac{4N_1N_5+2N_2N_5+N_1N_3+N_1N_4+N_3N_5+N_4N_5}{2N_1N_5}} \\ &= \frac{4N_1N_5}{4N_1N_5+2N_2N_5+N_1N_3+N_1N_4+N_3N_5+N_4N_5}\end{aligned}$$

In stark contrast, the partial derivative of F1_{rel} with respect to N_4 is:

$$\frac{\partial \text{F1}_{rel}}{\partial N_4} = -\frac{\text{F1}_{rel}^2}{4}\left(\frac{1}{N_1} + \frac{N_3+N_4+N_5}{N_5^2}\right) \tag{16}$$

The sign of this derivative is **consistently non-positive**. We can analyze its components to confirm this:

- The term F1_{rel}^2 is always non-negative.
- The term within the parentheses, $\left(\frac{1}{N_1} + \frac{N_3+N_4+N_5}{N_5^2}\right)$, is a sum of non-negative quantities and is strictly positive for any non-trivial case where the metric is defined ($N_1 > 0$ and $N_5 > 0$).
- The leading factor of $-\frac{1}{4}$ ensures the entire expression is non-positive.

This non-positive derivative signifies the ideal behavior. It mathematically guarantees that any increase in the critical error count N_4 will **always lead to a decrease or no change** in the F1_{rel} score. The metric consistently and correctly penalizes unreliability.

Weighted F1_{rel}. F1_{rel} can be applied to a broader range of scenarios involving the trade-off between helpfulness and truthfulness. To make the reliability metric applicable across various scenarios, we add a weighting parameters into F1_{rel}:

$$\text{F}\beta_{rel} = (1+\beta^2)\cdot\frac{\text{F1}_{\text{ans}}\cdot\text{F1}_{\text{abs}}}{\beta^2\cdot\text{F1}_{\text{ans}}+\text{F1}_{\text{abs}}}, \tag{17}$$

when $\beta = 1$, the metric is equivalent to our original F1_{rel}. When $\beta = 0$, the metric is equivalent to F1_{ans}. When β approaches infinity, the metric is equivalent to F1_{abs}. In scenarios like search engines, a small β value is appropriate, while in scenarios like healthcare and law, a large β is more suitable.

Conclusion. The derivative analysis provides a rigorous mathematical basis for concluding that F1_{rel} is a superior metric for evaluating model reliability. While the RS metric can fail catastrophically by rewarding incorrect behavior, the F1_{rel} metric provides a robust, reliable, and interpretable gradient that always punishes critical errors. This makes F1_{rel} a far more suitable and trustworthy measure for developing reliable AI systems.

E.3 Special Case

To highlight the superiority of our proposed metric F1_{rel} (the harmonic mean of F1_{ans} and F1_{abs}) over the Reliability Score (RS), we analyze a specific, illustrative scenario: a dataset composed exclusively of unanswerable questions. A robust reliability metric must correctly identify and reward a model that abstains from such questions, while penalizing a model that makes incorrect guesses. Consider a dataset that contains only unanswerable questions. This implies:

$$N_1 = 0, \quad N_2 = 0, \quad N_3 = 0$$

the total number of samples is therefore $N = N_4 + N_5$.

We evaluate two distinct models on this dataset:

- **Model A (The Perfect Abstainer):** This model performs perfectly by correctly identifying all questions as unanswerable and abstaining every time, defined by $N_5 = N, N_4 = 0$.
- **Model B (The Reckless Guesser):** This model is flawed. It abstains half the time but provides an incorrect answer for the other half, defined by $N_5 = 0.5N, N_4 = 0.5N$.

Our expectation: A valid metric should award Model A a perfect (or maximum) score and Model B a significantly lower score.

Analysis of the Reliability Score (RS). The simplified formula for RS in this case is:

$$\mathrm{RS} = (\alpha \cdot \mathrm{Cov}) + ((1-\alpha) \cdot \mathrm{Acc}) = (\frac{N_4}{N} \cdot \frac{N_5}{N}) + ((1 - \frac{N_4}{N}) \cdot 0) = \frac{N_4 N_5}{N^2}.$$

For Model A:

$$\mathrm{RS}_A = \frac{(0)(N)}{N^2} = 0$$

For Model B:

$$\mathrm{RS}_B = \frac{(0.5N)(0.5N)}{N^2} = \frac{0.25N^2}{N^2} = 0.25$$

RS erroneously assigns Model A a score of 0 and Model B a higher score of 0.25.

Evaluation of the proposed F1$_{rel}$ metric. In this scenario, F1$_{rel}$ depends solely on F1$_{abs}$, reflecting abstention performance:

$$\mathrm{F1}_{rel} = \mathrm{F1}_{abs} = \frac{2N_5}{N_3 + N_4 + 2N_5}$$

For Model A:

$$\mathrm{F1}_{absA} = \frac{2(N)}{0 + 0 + 2(N)} = \frac{2N}{2N} = \mathbf{1}$$

This assigns a perfect score of 1, reflecting optimal performance. For Model B:

$$\mathrm{F1}_{absB} = \frac{2(0.5N)}{0 + 0.5N + 2(0.5N)} = \frac{N}{0.5N + N} = \frac{N}{1.5N} = \frac{2}{3} \approx 0.67$$

This appropriately penalizes Model B for incorrect answers.

Conclusion The Reliability Score (RS) exhibits a critical flaw by rewarding incorrect answers on unanswerable questions, contradicting the principles of reliability. In contrast, F1$_{rel}$, via its F1$_{abs}$ component, accurately assigns Model A a perfect score and penalizes Model B, demonstrating its robustness and suitability as a reliability metric.

F Case Study

We present detailed examples using Llama3-8B-Instruct in Table 8. In question 1, both ICL-IDK and GRPO-SE exhibit over-abstention. While ICL achieves an accuracy of 70%, ICL-IDK generates two correct answers, one hallucinated response, and seven abstentions. GRPO-SE produces six correct answers but discards five of them. In contrast, FISCORE successfully retains all correct answers while discarding all incorrect ones. In question 2, GRPO-SE demonstrates overaggressive behavior by retaining all incorrect answers, whereas FISCORE appropriately discards three incorrect responses, achieving calibrated post-hoc abstention.

Note on inference cost. The 10 sampled responses shown in Table 8 are generated solely for the purpose of illustrating the rollout distribution that *would be* observed during training. During actual inference, no multiple sampling is required: the model performs a single greedy decoding pass and outputs one answer together with one confidence expression. The multi-sample overhead therefore applies only to training, which is already analyzed in Appendix H.

G Abstention Rate

Table 9 shows the abstention rate for each method on each dataset.

Question1: The world's largest marketer of fruit juices, what is the juice arm of the Coca Cola company?
Answer: maid

ICL	ICL-IDK	GRPO-SE	FiSCoRe
minute maid	maid	<answer> maid </answer> <confidence> sure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	maid	<answer> maid </answer> <confidence> unsure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	simply	<answer> maid </answer> <confidence> unsure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	I don't know.	<answer> maid </answer> <confidence> unsure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	I don't know.	<answer> maid </answer> <confidence> unsure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	I don't know.	<answer> maid </answer> <confidence> unsure </confidence>	<answer> maid </answer> <confidence> sure </confidence>
minute maid	I don't know.	<answer> simply </answer> <confidence> unsure </confidence>	<answer> simply </answer> <confidence> unsure </confidence>
simply	I don't know.	<answer> simply </answer> <confidence> unsure </confidence>	<answer> simply </answer> <confidence> unsure </confidence>
fanta	I don't know.	<answer> fanta </answer> <confidence> unsure </confidence>	<answer> fanta </answer> <confidence> unsure </confidence>
fanta	I don't know.	<answer> fanta </answer> <confidence> unsure </confidence>	<answer> fanta </answer> <confidence> unsure </confidence>

Question2: Which mythical beings were said to inhabit the slopes of Mount Etna?
Answer: Cyclops

ICL	ICL-IDK	GRPO-SE	FiSCoRe
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	Cyclops	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Cyclops	I don't know.	<answer> Cyclops </answer> <confidence> sure </confidence>	<answer> Cyclops </answer> <confidence> sure </confidence>
Centaurs	I don't know.	<answer> Centaurs </answer> <confidence> sure </confidence>	<answer> Centaurs </answer> <confidence> unsure </confidence>
Centaurs	I don't know.	<answer> Centaurs </answer> <confidence> sure </confidence>	<answer> Centaurs </answer> <confidence> unsure </confidence>
Centaurs	Centaurs	<answer> Centaurs </answer> <confidence> sure </confidence>	<answer> Centaurs </answer> <confidence> unsure </confidence>

Table 8: Case studies of different methods on Llama3-8B-Instruct. We sample 10 answers for each question with the temperature of 0.2.

	Llama3-8B-Instruct				Qwen-2.5-7B-Instruct			
	Pararel	**TriviaQA**	**NQ**	**SciQ**	**Pararel**	**TriviaQA**	**NQ**	**SciQ**
ICL-IDK	48.30	12.35	55.79	15.40	54.91	21.97	28.67	4.10
ICL-Unsure	37.23	17.86	7.26	4.60	56.45	18.99	21.66	11.60
R-Tuning	54.03	59.69	70.69	75.30	66.47	84.10	92.85	78.00
R-Tuning-U	56.86	73.62	75.32	55.50	77.69	85.38	90.01	76.70
SE-Tuning	43.81	63.18	71.16	83.90	69.03	88.78	88.34	58.51
GRPO-SE	55.93	35.76	68.59	24.20	63.53	40.08	62.91	22.30
FISCORE	21.25	23.03	58.25	28.50	28.28	35.79	59.97	24.90

Table 9: Abstention rate.

H Computation Cost

R-Tuning (Zhang et al., 2024) trains LLMs to respond with "I don't know" explicitly when fed an uncertain question. In contrast, FISCORE enables self-reflective abstention or answering, at the cost of increased computation. We compare ICL, R-Tuning, and FISCORE in terms of the average generation length of all questions in the test sets and the generation time on 100 randomly sampled questions using an NVIDIA L40 GPU. As shown in Table 10, R-Tuning exhibits slightly slower generation speed than ICL, though in certain cases it performs comparably or even faster. In contrast, FISCORE incurs a 3-4x increase in generation time relative to ICL. This overhead arises from the structured output format: for answered questions, our method introduces additional tokens for tags and confidence expression; for abstained questions, it further generates a full-length answer to justify the refusal, thereby increasing computational cost.

Model	**Avg. tokens**				**Avg. inference time (s)**			
	Pararel	**TriviaQA**	**NQ**	**SciQ**	**Pararel**	**TriviaQA**	**NQ**	**SciQ**
Llama3-8b-Instruct								
ICL	2.86	3.93	5.14	3.47	7.75	11.93	12.36	10.39
R-Tuning	5.33	5.74	6.47	3.97	13.84	16.14	15.31	20.77
FISCORE	15.83	16.82	17.70	16.06	31.79	35.29	35.13	32.82
Qwen2.5-7B-Instruct								
ICL	3.36	5.54	12.44	3.64	12.23	12.49	19.49	11.04
R-Tuning	4.35	4.92	5.84	3.41	13.44	15.54	16.62	20.73
FISCORE	12.17	16.84	19.36	15.29	42.69	44.54	54.20	42.49

Table 10: The results of the computation cost. Left: average number of tokens of the model's response. Right: average generation time of 100 random questions.