From Sampling to Cognition: Modeling Internal Cognitive Confidence in Language Models for Robust Uncertainty Calibration

Hao Li[1], Tao He[1], Jiafeng Liang[1], Zheng Chu[1], Ming Liu[1,2*]

[1]Harbin Institute of Technology, Harbin, China
[2]PengCheng Laboratory, Shenzhen, China
{haoli, the, jfliang, zchu, mliu}@ir.hit.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of tasks, yet they generally lack self-awareness, often displaying overconfidence when confronted with questions beyond their knowledge boundaries. This limitation severely hinders their trustworthiness in high-stakes scenarios. Existing calibration methods typically rely on sampling accuracy, derived from multiple outputs, as a proxy for model confidence. However, this coarse-grained metric fails to capture the model's internal cognitive states, such as confusion, hallucination, or persistent belief in false knowledge. To address this, we propose `CogConf` (Cognitive Confidence), a cognitively grounded uncertainty signal that extends sampling accuracy by incorporating the semantic diversity of incorrect answers and the model's abstention behaviors. By shifting the focus from sampling-based to cognition-oriented uncertainty modeling, `CogConf` offers a more faithful reflection of the model's internal beliefs. Building on this signal, we introduce CogAlign, a simple yet effective alignment framework that explicitly aligns the model's verbalized confidence with `CogConf`, thereby producing uncertainty estimates that better reflect the model's internal cognition. Experimental results on six knowledge-intensive in-domain and out-of-domain QA datasets demonstrate that `CogConf` robustly characterizes the model's internal uncertainty. Building on this foundation, CogAlign guides the model's expression to significantly enhance the trustworthiness and utility of its uncertainty calibration without compromising its underlying QA capabilities, while also demonstrating strong cross-task generalization and output stability. Offering a new pathway toward building more trustworthy LLMs.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable advancements in their capabilities, rapidly expanding their footprint across diverse domains (Jiang et al. 2023; Dubey et al. 2024). However, this growth in capability has not been accompanied by a commensurate increase in trustworthiness (Xiong et al. 2024). This growing "Capability-Trustworthiness Gap" severely impedes their practical application and deployment in high-stakes, sensitive domains such as healthcare, law, and finance. As a result, to build more reliable and secure LLMs, improving the model's ability to assess and communicate its own uncertainty has become an urgent research focus.

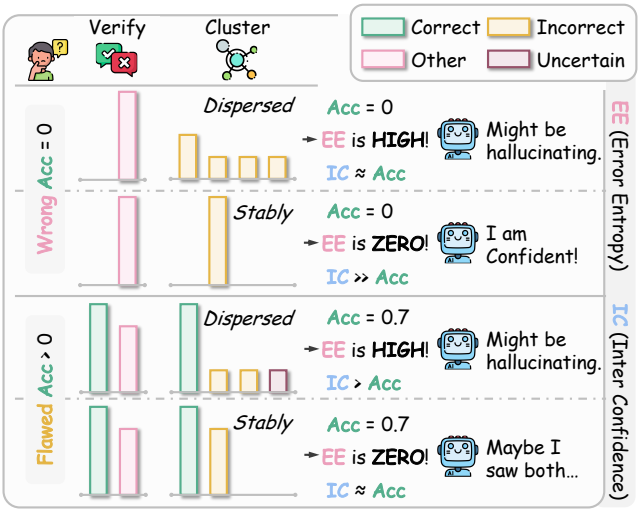

Figure 1: An illustration of four cognitive failure patterns across two main scenarios: *Wrong* and *Flawed*. The patterns reveal how Sampling Accuracy (`SA`) alone is insufficient to capture the model's distinct internal confidence states.

A trustworthy LLM should possess fundamental self-awareness, enabling it to recognize its own knowledge boundaries (Amayuelas et al. 2024). However, the prevailing next-token prediction paradigm, along with the lack of uncertainty expression in training corpora, leads models to exhibit a systematic tendency toward overconfidence (Huang et al. 2025b). When faced with unfamiliar or ambiguous contexts, they often generate responses that appear plausible but are factually incorrect, a phenomenon known as hallucination. To mitigate such hallucinations and improve model reliability, a key research direction is to calibrate the model's internal uncertainty. Existing approaches often encourage models to abstain when uncertain (Zhang et al. 2024a) or to estimate answer credibility using output probabilities (Tian et al. 2023), prompt engineering (Tian et al. 2023), or external evaluators. Among these methods, Sampling Accuracy

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

(SA), computed from multiple generations, is widely used as a simple and intuitive proxy for model confidence (Xu et al. 2024; Cheng et al. 2024). Another line of research focuses on detecting hallucinations in generated content. These methods typically involve multiple answer samples, analyze their semantic distributions, and compute Semantic Entropy (SE) (Farquhar et al. 2024) to assess the stability or factual deviation of model outputs. SE is regarded as a key signal for gauging the model's degree of confusion.

However, no single-dimensional signal can fully capture the complexity of a model's internal belief states. To better highlight the limitations of existing metrics, we categorize failure cases into two scenarios and examine two representative semantic distribution patterns within each (see Figure 1): In *Wrong* ($\texttt{SA} = 0$) scenarios, a *Stably Wrong* case features sampled answers concentrated on a single incorrect option, reflecting the model's strong belief in false knowledge, its internal confidence is thus significantly higher than the SA. In contrast, a *Dispersed Wrong* case shows samples fluctuating among several incorrect options, indicating substantial uncertainty; though SA remains zero, the model's confidence is much lower. In *Flawed* ($0 < \texttt{SA} < 1$) scenarios, a *Stably Flawed* case involves semantically similar incorrect options forming a stable distraction. Although the model does not firmly select the correct answer, it retains some belief in it, and its confidence should be close to SA. Conversely, in a *Dispersed Flawed* case, the incorrect answers are scattered and mutually competitive, weakening their interference and making it easier for the model to identify the correct choice, thus, the model's confidence is higher than SA. These observations suggest that SA ignores the structural concentration of internal beliefs, while SE ignores whether those beliefs are correct. Neither alone can adequately reflect the model's true confidence state.

To more accurately characterize the internal cognitive state of LLMs, we propose a novel internal confidence signal, CogConf (Cognitive Confidence), as a cognitively-grounded alternative to the commonly used SA. Unlike traditional metrics that focus on external behavior, CogConf incorporates the semantic entropy of incorrect answers (ESE) to quantify the distributional structure of the model's internal beliefs when it errs. This signal is further combined with the proportion of abstention responses to provide a holistic measure of cognitive uncertainty.

Building on this signal, we introduce CogAlign, a simple but effective framework for verbal confidence calibration. This framework guides the model to generate verbal confidence expressions that align with the CogConf signal, enabling more trustworthy and cognitively-grounded self-assessment. Specifically, the framework consists of three core stages: (i) *Sampling and Semantic Clustering*, where we perform multiple answer generations for each input question and cluster the responses based on semantic similarity to calculate the proportion of each answer type, including correct answers (to obtain SA), refusals, and different categories of incorrect answers; (ii) *CogConf Quantification*, where we first categorize all cases into two core scenarios based on the value of SA (*Wrong* and *Flawed*), then calculate its ESE in a corresponding manner, and based on this, further construct the comprehensive confidence metric CogConf; and (iii) *Confidence Alignment Training*, where we design a reinforcement learning objective containing both a *consistency reward* and a *correction bonus* to drive the model to generate verbal confidence scores that align with CogConf.

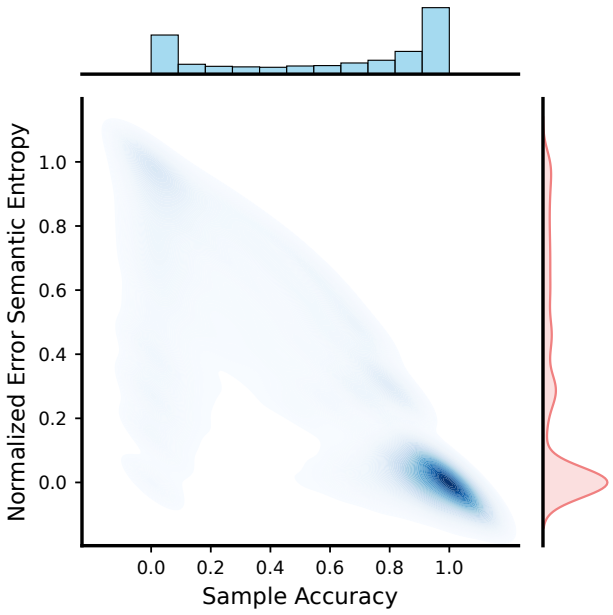

Figure 2: Joint distribution of Sampling Accuracy and Error Semantic Entropy on TriviaQA. The vertical spread of entropy demonstrates the unreliability of Sampling Accuracy as a proxy for the model's internal cognitive state.

We systematically evaluated our approach on six question-answering benchmarks spanning diverse knowledge domains and both in-domain and out-of-domain settings: TriviaQA (Joshi et al. 2017), CommonsenseQA (Talmor et al. 2019), ScienceQA (Lu et al. 2022), SportQA (Xia et al. 2024), StrategyQA (Geva et al. 2021), and ARC-Challenge (Bhakthavatsalam et al. 2021). Results show that CogAlign significantly enhances model calibration without compromising its original QA accuracy. In particular, the average Expected Calibration Error was reduced from 25.4% to 6.4%, while the Area Under the Receiver Operating Characteristic Curve improved from 56.5% to 67.3%, substantially boosting the reliability and discriminative power of the model's confidence estimates.

2 Preliminary Study: The Misalignment of Sampling Accuracy and Cognitive State

Existing research in confidence calibration is largely predicated on a core assumption: SA can serve as an effective proxy for a model's true confidence. This section presents a preliminary study that empirically challenges this assumption. Our central thesis is that the single metric SA cannot comprehensively characterize a model's complex internal cognitive state, as it overlooks the semantic distribution structure of the answers, a crucial factor in revealing the

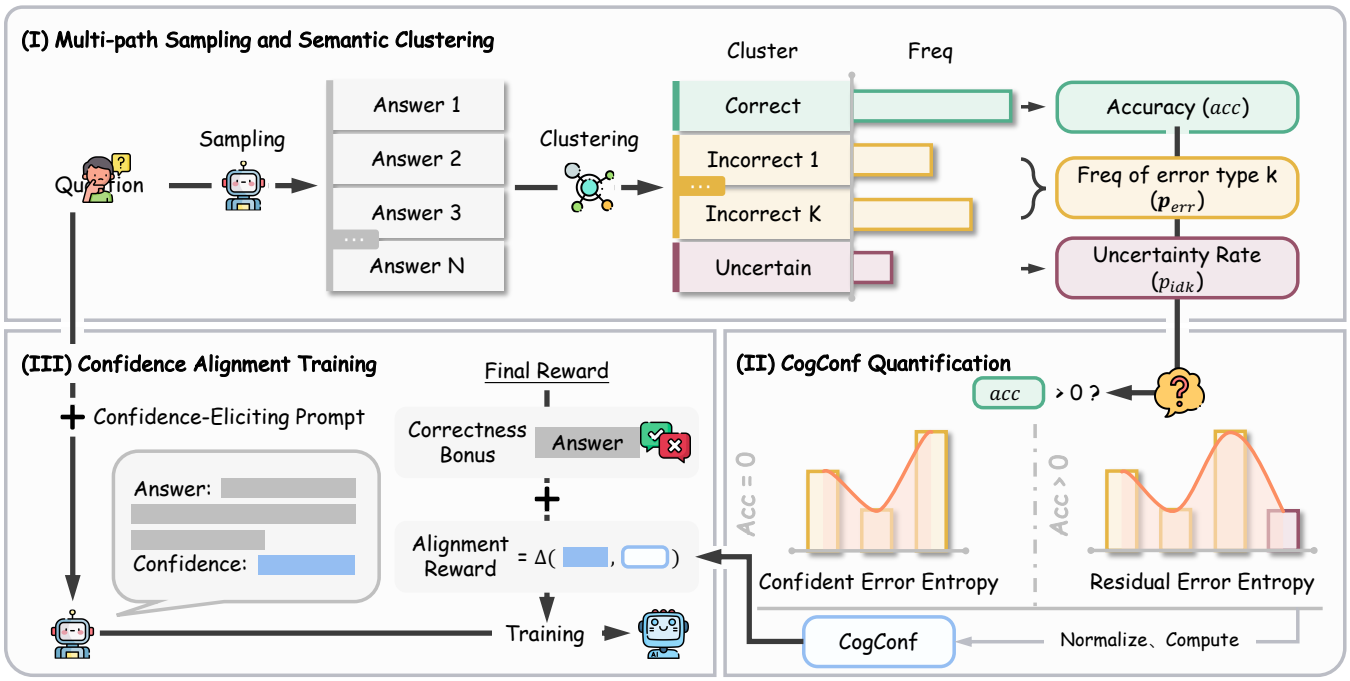

Figure 3: Overview of the CogAlign framework. (i) Multiple answers are sampled and semantically clustered to categorize responses. (ii) Our cognitive signal, CogConf, is then computed by adaptively integrating features like SA and ESE based on whether the model is completely *Wrong* (SA $= 0$) or partially correct (*Flawed*, SA > 0). (iii) Finally, reinforcement learning aligns the model's verbal confidence with the CogConf target.

model's true confidence. We use the ESE as a proxy for this structure. To validate this thesis, we conduct an in-depth analysis of test results from the Llama-3-8B model on the TriviaQA dataset, with $N = 10$ responses per question.

As visualized in Figure 2, the joint distribution of SA and ESE intuitively reveals this limitation. The densest region of the plot is at the bottom-right vertex (SA $= 1$, ESE $= 0$), indicating that the model answers correctly with high confidence in the majority of cases. However, in scenarios where the model errs, we clearly identify four failure modes defined by the combination of SA and ESE, each with a distinctly different cognitive meaning: In the *Wrong* (SA $= 0$) category, we observe two states: the *Stably Wrong* state, located in the bottom-left corner (ESE ≈ 0), where the low-entropy distribution indicates the model holds a strong but erroneous belief in specific false knowledge; and the *Dispersed Wrong* state, which spreads upwards, where the high-entropy distribution reflects a state of high cognitive uncertainty. Similarly, in the *Flawed* ($0 <$ SA < 1) category, two states also exist: the *Stably Flawed* state, which forms in the lower-middle region (ESE ≈ 0) and represents the correct answer competing with a strong, incorrect belief; and the *Dispersed Flawed* state, which spreads into high-entropy regions, indicating that the competitors to the correct answer are a series of semantically diverse, weak interferences.

The concurrent existence of these four cognitive states powerfully demonstrates the inadequacy of the single SA metric. SA conflates states of high-confidence error with states of high uncertainty (when SA $= 0$) and also conflates scenarios with strong competitors versus those with weak interferences (when SA > 0), yet the underlying mechanisms and appropriate confidence levels for these states are fundamentally different. This finding provides a strong empirical motivation for us to develop a new confidence metric that can perceive both accuracy and semantic structure, aiming to more faithfully reflect the model's cognitive state.

3 Methodology

This section details our proposed framework, CogAlign, which calibrates verbal confidence by first quantifying the model's internal cognitive state and then aligning its linguistic expression to this state. Throughout this process, it intricately models the semantic distribution of errors, thereby overcoming the limitations of coarse-grained metrics like Sampling Accuracy. Furthermore, CogAlign leverages a cognitively-grounded signal, CogConf, to supervise confidence expression, consequently mitigating the model's tendency for overconfidence and hallucination.

Firstly, we perform multi-path sampling and semantic clustering to deconstruct the model's response distribution. Then, we quantify this distribution into our novel cognitive signal, CogConf. Finally, we employ confidence alignment training to guide the model in generating verbal confidence that faithfully reflects this internal signal. An overview of the methodology is illustrated in Figure 3. We will detail each of these stages in the following sections: Multi-path Sampling and Semantic Clustering in §3.1, CogConf Quantification in §3.2, and Confidence Alignment Training in §3.3.

3.1 Multi-path Sampling and Semantic Clustering

For a given input question q, we first employ nucleus sampling to generate a set of N independent answer samples, denoted as $\mathcal{S} = \{s_1, \ldots, s_N\}$, from the language model $\mathcal{M}$. Each sampled answer s_i is then classified into one of three disjoint sets based on its content:

- **Correct Answers ($\mathcal{C}$):** The set of samples that match the ground-truth reference answer.
- **Incorrect Answers ($\mathcal{I}$):** The set of samples that are factually incorrect.
- **Abstention Answers ($\mathcal{U}$):** The set of samples expressing epistemic uncertainty (e.g., "I don't know"), which we identify by checking for semantic similarity to a predefined set of uncertainty expressions (Yin et al. 2023).

To analyze the structure of the model's errors, we further apply semantic clustering to the set of incorrect answers $\mathcal{I}$, yielding K distinct incorrect answer clusters: $\mathcal{I}_1, \ldots, \mathcal{I}_K$. Based on this categorization, we extract three key statistical features for constructing our `CogConf` signal:

Sampling Accuracy (`SA`). This metric provides a baseline measure of correctness.

$$\texttt{SA} = \frac{|\mathcal{C}|}{N} \tag{1}$$

Error Semantic Distribution ($\mathbf{p}_{\text{err}}$). To understand the consistency of errors, we compute the probability distribution over the K incorrect answer clusters. The proportion p_k for each cluster $\mathcal{I}_k$ is:

$$p_k = \frac{|\mathcal{I}_k|}{|\mathcal{I}|}, \quad \text{for } k = 1, \ldots, K \tag{2}$$

This distribution is the basis for calculating the **Semantic Entropy of Errors (`ESE`)**, which allows us to differentiate between concentrated, high-confidence errors and scattered, low-confidence hallucinations.

Abstention Rate (p_{abstain}). This metric quantifies the model's explicit expression of uncertainty.

$$p_{\text{abstain}} = \frac{|\mathcal{U}|}{N} \tag{3}$$

3.2 CogConf Quantification

Having extracted the statistical features, we now detail the construction of our cognitive confidence signal, `CogConf`. A key innovation of `CogConf` is its adaptive formulation, which distinguishes between the two fundamental failure scenarios identified in Section 2: *Wrong* ($\texttt{SA} = 0$) and **Flawed** ($\texttt{SA} > 0$). The rationale is that the role of an abstention answer changes: when the model is completely wrong, abstaining is a form of correct self-assessment; when it possesses some correct knowledge, abstaining is a form of error.

This motivates our scenario-specific calculation of the `ESE`.

Confident Error Entropy (ESE_{con}). This metric is used in the *Wrong* scenario to measure the consistency of confident mistakes. It is calculated exclusively over the K clusters of factually incorrect answers $\mathcal{I}$ using their internal distribution $\mathbf{p}_{\text{err}}$. A low ESE_{con} corresponds to the *Stably Wrong* state, while a high value corresponds to the *Dispersed Wrong* state.

$$\text{ESE}_{\text{con}} = -\sum_{k=1}^{K} p_k \log p_k \tag{4}$$

Residual Error Entropy (ESE_{res}). This metric is used in the *Flawed* scenario. It is calculated over the combined set of "residual errors," $\mathcal{E}_{\text{res}} = \mathcal{I} \cup \mathcal{U}$. We define a new probability distribution $\mathbf{q}_{\text{res}}$ over this combined set of $K^+ = K + 1$ categories. A low ESE_{res} corresponds to the *Stably Flawed* state (a strong competitor), while a high value corresponds to the *Dispersed Flawed* state (weak interferences).

$$\text{ESE}_{\text{res}} = -\sum_{j=1}^{K^+} q_j \log q_j \tag{5}$$

Final `CogConf` Formulation. The final `CogConf` score adaptively combines these components. The entropy values are first normalized by $\log N$ to ensure a consistent scale. The formulation includes a hyperparameter $\lambda \in [0, 1]$ to control the contribution of the entropy-based cognitive signal, which defaults to 0.25 in our experiments. Notably, when $\lambda = 0$, our `CogConf` formulation degenerates to the standard `SA`. This highlights λ's role as a switch that controls the impact of our cognitive adjustments. The final score is defined as:

$$\texttt{CogConf} = \begin{cases} \lambda \cdot (1 - p_{\text{abstain}}) \cdot \left(1 - \frac{\text{ESE}_{\text{con}}}{\log N}\right) & \text{if } \texttt{SA} = 0 \\ \texttt{SA} + \lambda \cdot (1 - \texttt{SA}) \cdot \frac{\text{ESE}_{\text{res}}}{\log N} & \text{if } \texttt{SA} > 0 \end{cases} \tag{6}$$

This formulation implements our core hypothesis: it rewards cognitive consistency (low entropy) when the model is completely wrong but rewards cognitive diversity (high entropy, indicating no strong incorrect competitor) when the model possesses partial knowledge.

3.3 Confidence Alignment Training

The final stage of our framework, CogAlign *Training*, supervises the model $\mathcal{M}$ to generate a verbal confidence score, c_{verbal}, that aligns with our cognitive signal, `CogConf`. This process uses the more reliable `CogConf` as the supervisory target, replacing the unstable sampling accuracy used in traditional methods.

To achieve this, we adopt a Reinforcement Learning (RL) paradigm based on Proximal Policy Optimization (PPO) and design a simple reward function R that combines two intuitive components: (1) an **Alignment Reward**, which scores the model based on the proximity of its expressed confidence to our `CogConf` target, and (2) a **Correctness Bonus**, which gives an additional positive reward β when the generated answer is correct. The form of this reward function is

as follows:

$$R(a_{\text{gen}}, c_{\text{verbal}}) = \underbrace{\left(1 - 2 \cdot \left|\frac{c_{\text{verbal}}}{10} - \texttt{CogConf}(q)\right|\right)}_{\text{Alignment Reward}} + \underbrace{\beta \cdot \mathbb{I}(a_{\text{gen}} \text{ is correct})}_{\text{Correctness Bonus}} \quad (7)$$

where $\mathbb{I}(\cdot)$ is the indicator function and β is a hyperparameter (defaulting to 0.25 in our implementation). This function directly incentivizes the model to be both accurate and honest in its self-assessment.

4 Experimental Setup

Datasets. To evaluate the effectiveness and generalizability of our method, we use training data constructed from the TriviaQA (Joshi et al. 2017) and CommonsenseQA (Talmor et al. 2019) training sets. During evaluation, in addition to assessing performance on the in-domain test sets of TriviaQA and CommonsenseQA, we conduct zero-shot evaluations on four QA benchmarks from diverse domains: ScienceQA (science knowledge) (Lu et al. 2022), SportQA (sports) (Xia et al. 2024), StrategyQA (strategic reasoning) (Geva et al. 2021), and ARC-Challenge (multi-domain) (Bhakthavatsalam et al. 2021).

Baseline Methods. To evaluate the effectiveness of our approach, we compare it against four representative baselines:

- **Vanilla (Tian et al. 2023):** Instruct the model to generate a confidence score directly after providing the answer.
- **PPO (Schulman et al. 2017):** A standard reinforcement learning baseline that directly fine-tunes the language model by optimizing its policy against a scalar reward signal, using the Proximal Policy Optimization algorithm.
- **PPOM (Leng et al. 2025):** This method first trains a reward model to assess the alignment between correctness of the responses and the confidence expressed, rewarding responses with correctness and confidence that match well and penalizing misaligned ones. The calibrated reward model is then used to guide PPO fine-tuning of the language model.
- **RewardingDoubt (Stangel et al. 2025):** This approach uses a reward function based on the logarithmic scoring rule, which penalizes both overconfidence and underconfidence. It directly trains the model via PPO to express well-calibrated confidence alongside its answers.

Evaluation Metrics. We evaluate the models along two dimensions: answer correctness and confidence quality. Specifically, we adopt three core metrics: Accuracy (ACC) to measure task performance,Expected Calibration Error (ECE) (Guo et al. 2017) to quantify the gap between predicted confidence and actual accuracy, and Area Under the Receiver Operating Characteristic Curve (AUROC) (Hendrycks and Gimpel 2017) to assess the ability of confidence scores to distinguish between correct and incorrect answers.

Methods	ECE ↓	AUROC ↑	ACC ↑
Llama-3-8B			
Vanilla	26.5	59.4	69.8
PPO	30.9	50.6	68.9
PPOM	26.8	56.4	69.1
RewardingDoubt	9.8	62.1	68.8
CogAlign	**7.5**	**64.2**	**70.2**
align with SA	10.2	61.4	68.8
align with SE	11.5	62.0	68.4
Mistral-7B			
Vanilla	27.3	61.0	69.6
PPO	18.2	60.7	69.2
PPOM	18.7	59.6	**70.3**
RewardingDoubt	15.3	62.7	69.1
CogAlign	**7.8**	**72.7**	69.5
align with SA	14.8	70.2	68.1
align with SE	15.4	65.6	68.2

Table 1: In-domain average performance of Llama-3-8B and Mistral-7B on TriviaQA and CommonsenseQA. Arrows in the column headers denote the preferred direction for each metric (↓: lower is better; ↑: higher is better).

Implementation Details. Following prior work (Leng et al. 2025), we use instruction-tuned (non-RLHF) models to isolate RL's impact on confidence. For the backbone models, we perform all experiments using fine-tuned versions of Llama-3-8B (Dubey et al. 2024) and Mistral-7B (Jiang et al. 2023), two representative open source LLMs. During training, we perform 10 sampling passes per question to compute key indicators for reward modeling.

5 Results and Analysis

5.1 Main Experimental Results

We systematically evaluate all methods on six question-answering benchmarks, including both in-domain and out-of-domain settings. As shown in Table 1 and Table 2, CogAlign consistently improves model calibration while preserving task accuracy, especially in distribution-shifted scenarios. These findings underscore the strength of our cognitively grounded alignment approach.

In-Domain Calibration Performance. On in-domain benchmarks (Table 1), CogAlign delivers superior calibration across both models. For Llama-3-8B, it achieves the lowest ECE and highest AUROC, indicating that the model's predicted confidence better reflects its true correctness likelihood. Importantly, this gain in calibration does not degrade task accuracy, which remains competitive with or slightly exceeds the best baselines. Similar trends are observed for Mistral-7B, where CogAlign sharply improves AUROC (from 62.7 to 72.7) while maintaining high accuracy. These results suggest that CogAlign strengthens the model's internal uncertainty representation without compromising its ability to answer correctly.

Methods	ScienceQA			SportQA			StrategyQA			ARC-Challenge		
	ECE ↓	AUROC ↑	ACC ↑	ECE ↓	AUROC ↑	ACC ↑	ECE ↓	AUROC ↑	ACC ↑	ECE ↓	AUROC ↑	ACC ↑
Llama-3-8B												
Vanilla	22.8	55.2	73.3	40.7	54.9	57.1	25.4	58.2	64.2	21.1	55.3	76.8
PPO	24.4	55.6	73.0	38.4	51.7	60.2	27.6	55.8	**69.7**	22.3	51.5	77.4
PPOM	27.7	54.2	69.6	48.0	52.3	50.0	25.6	56.4	66.9	22.2	53.0	76.5
RewardingDoubt	12.9	69.1	**77.1**	**7.5**	68.4	**65.0**	8.7	61.8	63.8	9.0	58.8	76.6
CogAlign	6.3	**70.4**	74.0	14.0	67.2	59.3	**3.3**	**64.1**	66.3	**2.6**	**59.9**	**78.1**
align with SA	**6.0**	69.2	73.2	21.3	67.0	51.1	5.4	62.0	63.7	2.9	58.4	77.3
align with SE	7.5	70.2	70.4	18.6	**69.4**	50.1	3.9	61.9	63.4	6.3	55.8	75.9
Mistral-7B												
Vanilla	19.3	59.6	77.2	31.2	57.6	66.1	22.5	55.0	72.6	20.6	56.0	77.0
PPO	7.2	63.1	78.8	18.5	61.9	68.2	12.4	58.0	74.2	8.1	61.4	**79.4**
PPOM	7.8	59.3	**79.9**	20.1	60.1	**69.0**	13.2	57.4	**74.4**	11.5	56.4	77.6
RewardingDoubt	6.6	69.2	78.2	19.4	63.3	64.8	8.9	60.0	71.7	**3.2**	65.5	77.1
CogAlign	**6.2**	**73.4**	78.5	**4.9**	73.4	65.1	**7.0**	**63.8**	72.2	6.2	**66.2**	76.9
align with SA	6.3	68.0	78.2	18.4	**74.8**	62.9	11.2	59.3	72.9	6.3	64.6	75.5
align with SE	15.6	65.3	79.8	8.1	69.9	64.7	7.7	59.1	69.4	13.3	61.6	76.5

Table 2: Performance comparison on four out-of-domain datasets: ScienceQA, SportQA, StrategyQA, and ARC-Challenge. Each metric is reported per dataset (without averaging) to capture dataset-specific generalization behavior. ECE (Expected Calibration Error; ↓) measures calibration accuracy, AUROC (↑) assesses confidence discrimination, and ACC (↑) reflects task accuracy. Best results are highlighted in **bold**. *align with* SA and *align with* SE represent ablation variants of our CogAlign method, designed to examine the effectiveness of the CogConf metric in capturing the model's internal confidence.

Out-of-Domain Generalization. The advantage of CogAlign becomes more pronounced under distribution shift. As shown in Table 2, when evaluated on out-of-domain datasets, our method yields substantial reductions in ECE and notable gains in AUROC across all tasks and models. For instance, Llama-3-8B with CogAlign achieves lower ECE and higher AUROC than all baselines, despite the inherent domain mismatch. This pattern holds across ScienceQA, SportQA, StrategyQA, and ARC-Challenge, highlighting that the confidence learned via CogConf generalizes better than reward signals that focus solely on task correctness or confidence penalization.

Comparative Analysis and Ablation Study. A closer examination of the baselines helps explain why CogAlign outperforms them. Vanilla models tend to be overconfident, while PPO-based methods may inadvertently encourage miscalibrated but correct responses. RewardingDoubt improves calibration by discouraging overconfidence, but its reward design lacks sensitivity to the semantic structure of errors. In contrast, CogAlign integrates both accuracy and the fine-grained semantics of incorrect responses to align internal beliefs with model uncertainty. This advantage is further supported by our ablation study. Variants that align confidence using only SA or only SE fall short of the full method across most metrics. While each provides partial benefits, only their combination captures the multifaceted nature of model confidence. These results validate the effectiveness of CogConf as a proxy for internal belief state, confirming its utility in guiding robust confidence modeling.

5.2 Validating the Effectiveness of CogConf

To evaluate the effectiveness of CogConf as a confidence proxy, we examine its ability to predict the correctness of the model's final output. An ideal proxy should exhibit strong alignment with answer accuracy, which we quantify using the AUROC metric. As shown in Figure 4, CogConf consistently achieves the highest AUROC scores across both Llama-3-8B and Mistral-7B, highlighting its superior predictive capability compared to other confidence signals.

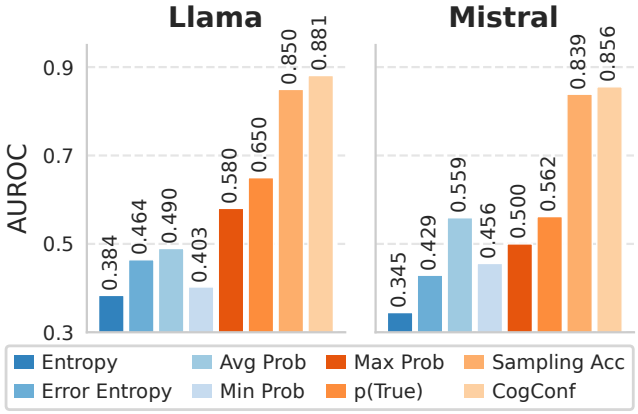

Figure 4: AUROC comparison of various confidence proxies on the two models. CogConf demonstrates the strongest predictive power, indicating the best alignment with the true correctness of the final answer.

This performance improvement stems from CogConf's ability to distinguish deeper internal cognitive states that SA cannot. The traditional metric SA fails in certain scenarios. For instance, in the *Wrong* case, SA cannot differentiate between a *Stably Wrong* state, where the model holds a firm but erroneous belief, and a *Dispersed Wrong* state, reflecting high cognitive uncertainty. In contrast, by incorporating the ESE, CogConf can capture this difference in belief structure. CogConf also excels in the *Flawed* case. It can

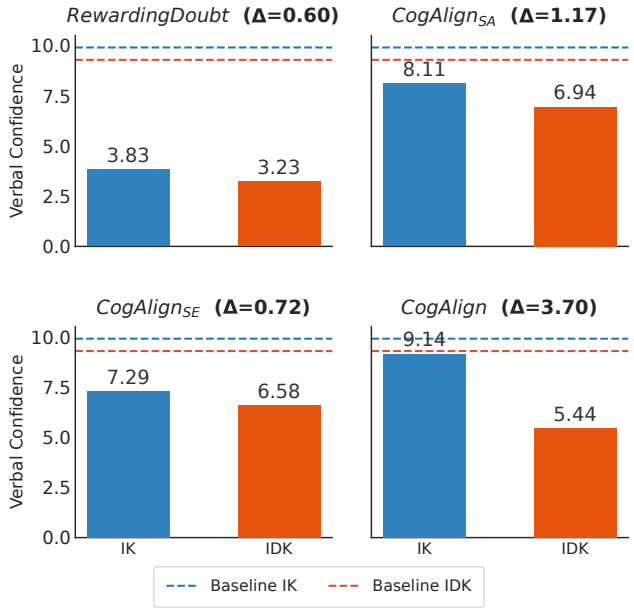

Figure 5: Average confidence scores of different methods on IK and IDK datasets. The blue and red dashed lines indicate the baseline confidence levels for IK (9.91) and IDK (9.29), respectively.

distinguish a *Stably Flawed* state, where the correct answer contends with a single, dominant incorrect alternative, from a *Dispersed Flawed* state, where the competitors are merely a series of semantically diverse and weak interferences. Although SA values may be similar in these two flawed states, the model's confidence in the final answer should be fundamentally different. CogConf successfully captures this distinction, assigning a higher confidence score to the latter case where the correct answer is more likely to prevail.

In summary, by integrating the semantic structure information, CogConf provides a more faithful representation of the model's true cognitive state, leading to superior performance in predicting the correctness of the final answer.

5.3 Analyzing the Differentiation of CogAlign

To analyze the confidence differentiation capability of different methods, we design an experiment to measure a model's self-awareness of its knowledge boundaries. To this end, we construct two distinct and representative subsets: **IK (I Know)**, comprising questions the model consistently answers correctly (SA $=1$); and **IDK (I Don't Know)**, representing the model's clear knowledge gaps, containing questions that the model consistently fails to answer correctly (SA $=0$) and for which its errors are highly dispersed (normalized ESE >0.8). This definition precisely corresponds to the "Dispersed Wrong" cognitive state.

The primary metric for this analysis is the average confidence gap (Δ) between the IK and IDK subsets, where a larger gap indicates superior differentiation capability. Figure 5 visualizes the results of this comparison. The analysis shows that the Vanilla baseline fails to distinguish between the subsets. The RewardingDoubt method exhibits a minimal differentiation capability ($\Delta = 0.6$), performing comparably to our ablation variant aligned only with SE (CogAlign_SE, $\Delta = 0.72$). In contrast, aligning with SA (CogAlign_SA) yields a more noticeable improvement ($\Delta = 1.17$). However, the CogAlign framework dramatically outperforms all other approaches, achieving the largest confidence gap by a wide margin ($\Delta = 3.70$).

The superior differentiation capability of CogAlign stems directly from the design of its supervisory signal, CogConf. Because the IDK set is explicitly constructed to represent the 'confused' cognitive state (high error entropy), CogConf—which is designed to recognize this state—provides a much clearer supervisory signal than simpler metrics. This confirms the superiority of CogConf for modeling internal cognitive states and fostering reliable self-assessment, thereby endowing the model with a stronger confidence differentiation capability.

6 Related Work

Confidence Estimation. Recent studies have explored various confidence estimation methods based on likelihood (Vazhentsev et al. 2023), prompt engineering (Lin, Hilton, and Evans 2022), and external evaluators (Han et al. 2024). SA, calculated from multiple samples, remains a mainstream proxy metric due to its intuitiveness and perceived reliability (Lyu et al. 2025). However, our research indicates that SA, as a coarse-grained external metric, struggles to capture the model's complex internal cognitive states—which is the core problem our work aims to solve.

Confidence Elicitation. Confidence elicitation aims to enable models to explicitly express their internal confidence rather than relying on post-hoc estimation. Prior research has explored supervised fine-tuning approaches that train models to reproduce externally estimated scores (Jang et al. 2025), reinforcement learning methods that induce intrinsic confidence through reward design (Stangel et al. 2025), and hybrid strategies that combine both paradigms (Xu et al. 2024). Our approach adopts a purely RL framework, aligning the model's expressed confidence directly with the proposed cognitive confidence signal CogConf.

7 Conclusion

To address the gap between external sampling accuracy and internal cognitive states, we propose CogConf—a confidence signal that captures internal uncertainty by combining answer correctness with semantic entropy. Building on this, we introduce CogAlign, a training framework that aligns the model's expressed confidence with its internal beliefs, shifting supervision from outcome-driven to introspective calibration. Experiments on six question-answering benchmarks, spanning both in-domain and out-of-domain settings, show that CogAlign achieves robust calibration and strong generalization, without sacrificing task accuracy. Overall, our work offers a new perspective on modeling cognitive uncertainty, advancing the development of more trustworthy and self-aware AI systems.

Acknowledgements

The research in this article is supported by the National Key Research and Development Project (2022YFF0903301), the National Science Foundation of China (U22B2059, 62276083), Key Research and Development Program of Heilongjiang Providence (2022ZX01A28) and the 5G Application Innovation Joint Research Institute's Project (A003).

References

Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; and Wang, W. Y. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 6416–6432. Association for Computational Linguistics.

Bhakthavatsalam, S.; Khashabi, D.; Khot, T.; Mishra, B. D.; Richardson, K.; Sabharwal, A.; Schoenick, C.; Tafjord, O.; and Clark, P. 2021. Think you have Solved Direct-Answer Question Answering? Try ARC-DA, the Direct-Answer AI2 Reasoning Challenge. *CoRR*, abs/2102.03315.

Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; He, Z.; Chen, K.; and Qiu, X. 2024. Can AI Assistants Know What They Don't Know? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Rozière, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I. M.; Misra, I.; Evtimov, I.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017): 625–630.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Trans. Assoc. Comput. Linguistics*, 9: 346–361.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR.

Han, H.; Li, T.; Chen, S.; Shi, J.; Du, C.; Xiao, Y.; Liang, J.; and Lin, X. 2024. Enhancing Confidence Expression in Large Language Models Through Learning from Past Experience. *CoRR*, abs/2404.10315.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Huang, C.; Huang, L.; Leng, J.; Liu, J.; and Huang, J. 2025a. Efficient Test-Time Scaling via Self-Calibration. *CoRR*, abs/2503.00031.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025b. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2): 42:1–42:55.

Jang, C.; Choi, M.; Kim, Y.; Lee, H.; and Lee, J. 2025. Verbalized Confidence Triggers Self-Verification: Emergent Behavior Without Explicit Reasoning Supervision. *CoRR*, abs/2506.03723.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Kang, Z.; Zhao, X.; and Song, D. 2025. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. *CoRR*, abs/2502.18581.

Leng, J.; Huang, C.; Zhu, B.; and Huang, J. 2025. Taming Overconfidence in LLMs: Reward Calibration in RLHF. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Li, M.; Zhao, Y.; Zhang, W.; Li, S.; Xie, W.; Ng, S.; Chua, T.; and Deng, Y. 2025. Knowledge Boundary of Large Language Models: A Survey. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 5131–5157. Association for Computational Linguistics.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching Models to Express Their Uncertainty in Words. *Trans. Mach. Learn. Res.*, 2022.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.; Zhu, S.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Lyu, Q.; Shridhar, K.; Malaviya, C.; Zhang, L.; Elazar, Y.; Tandon, N.; Apidianaki, M.; Sachan, M.; and Callison-Burch, C. 2025. Calibrating Large Language Models with Sample Consistency. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 19260–19268. AAAI Press.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Stangel, P.; Bani-Harouni, D.; Pellegrini, C.; Özsoy, E.; Zaripova, K.; Keicher, M.; and Navab, N. 2025. Rewarding Doubt: A Reinforcement Learning Approach to Confidence Calibration of Large Language Models. *CoRR*, abs/2503.02623.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4149–4158. Association for Computational Linguistics.

Tao, S.; Yao, L.; Ding, H.; Xie, Y.; Cao, Q.; Sun, F.; Gao, J.; Shen, H.; and Ding, B. 2024. When to Trust LLMs: Aligning Confidence with Response Quality. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 5984–5996. Association for Computational Linguistics.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 5433–5442. Association for Computational Linguistics.

Vazhentsev, A.; Tsvigun, A.; Vashurin, R.; Petrakov, S.; Vasilev, D.; Panov, M.; Panchenko, A.; and Shelmanov, A. 2023. Efficient Out-of-Domain Detection for Sequence to Sequence Models. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 1430–1454. Association for Computational Linguistics.

Xia, H.; Yang, Z.; Wang, Y.; Tracy, R.; Zhao, Y.; Huang, D.; Chen, Z.; Zhu, Y.; Wang, Y.; and Shen, W. 2024. SportQA: A Benchmark for Sports Understanding in Large Language Models. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 5061–5081. Association for Computational Linguistics.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xu, T.; Wu, S.; Diao, S.; Liu, X.; Wang, X.; Chen, Y.; and Gao, J. 2024. SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 5985–5998. Association for Computational Linguistics.

Xue, B.; Mi, F.; Zhu, Q.; Wang, H.; Wang, R.; Wang, S.; Yu, E.; Hu, X.; and Wong, K. 2025. UAlign: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 6002–6024. Association for Computational Linguistics.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 8653–8665. Association for Computational Linguistics.

Zhang, H.; Diao, S.; Lin, Y.; Fung, Y. R.; Lian, Q.; Wang, X.; Chen, Y.; Ji, H.; and Zhang, T. 2024a. R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 7113–7139. Association for Computational Linguistics.

Zhang, M.; Huang, M.; Shi, R.; Guo, L.; Peng, C.; Yan, P.; Zhou, Y.; and Qiu, X. 2024b. Calibrating the Confidence of Large Language Models by Eliciting Fidelity. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 2959–2979. Association for Computational Linguistics.