# Teaching Large Language Models to Express Knowledge Boundary from Their Own Signals

Anonymous ACL submission

## Abstract

Large language models (LLMs) have achieved great success, but their occasional content fabrication, or hallucination, limits their practical application. Hallucination arises because LLMs struggle to admit ignorance due to inadequate training on knowledge boundaries. We call it a limitation of LLMs that they can not accurately express their knowledge boundary, answering questions they know while admitting ignorance to questions they do not know. In this paper, we aim to teach LLMs to recognize and express their knowledge boundary, so they can reduce hallucinations caused by fabricating when they do not know. We propose CoKE, which first probes LLMs' knowledge boundary via internal confidence given a set of questions, and then leverages the probing results to elicit the expression of the knowledge boundary. Extensive experiments show CoKE helps LLMs express knowledge boundaries, answering known questions while declining unknown ones, significantly improving in-domain and out-of-domain performance.

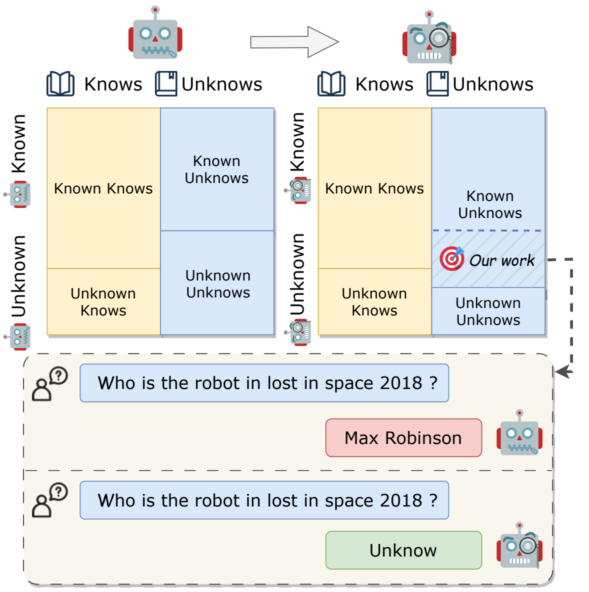


Figure 1: The evolution of the Known-Unknown Quadrant. The yellow portion represents the model's parametric knowledge. Our method increases the "Known Unknows", helping the model recognize and articulate its knowledge limitations.

## 1 Introduction

Large language models (LLMs) have emerged as an increasingly pivotal cornerstone for the development of artificial general intelligence. They exhibit powerful intellectual capabilities and vast storage of knowledge (Brown et al., 2020; Ouyang et al., 2022; Achiam et al., 2023), which enables them to generate valuable content. Recent research demonstrates that LLMs excel in passing various professional examinations requiring expert knowledge in domains like medical (Jin et al., 2021) and legal (Cui et al., 2023). Nevertheless, human users are hardly willing to seek professional suggestions from LLMs, due greatly to **hallucinations** in LLMs. Hallucinations in LLMs refer to the phenomenon that existing LLMs frequently generate untruthful information (Zhang et al., 2023b; Ji et al., 2023), which greatly undermines people's trust and acceptance of LLM-generated content.

An important cause of hallucinations is the model's insufficiency in knowledge boundary expression, which originates from the learning paradigm of LLMs. Pre-training and instruction fine-tuning serve as the two indispensable learning stages for current LLMs. The learning mechanism of these stages is to encourage LLMs to generate the provided text, which also makes LLMs prone to fabricating content when LLMs do not possess relevant knowledge (joh, 2023; Gekhman et al., 2024). Hence, LLMs are hardly instructed to express their ignorance, which is a lack of accurate knowledge boundary expression. Given a specific LLM and a question set, the corresponding question-answer pairs can be categorized based on two factors: (1) whether the model has corresponding parametric knowledge (knows v.s. unknows), and (2) whether the model is aware of the first factor (known v.s. unknown), as is depicted in Figure 1. Hallucinations frequently occur in the "Unknown Unknows" scenarios, where the model is unaware that it should explain its ignorance like humans, instead of struggling to give a hallucinated response.

Fine-tuning models to express knowledge boundaries faces two significant challenges. The first challenge is how to efficiently obtain data that reflects the internal knowledge of a specific model. Even if evaluation questions are easy to construct, obtaining expert-level answers in certain fields is costly. Additionally, since the model might produce correct answers in different forms from the reference answers, evaluating their correctness is also challenging (Kadavath et al., 2022; Zou et al., 2023). The second challenge is enabling the model to express its knowledge boundary robustly (Ren et al., 2023). We expect consistent knowledge boundary expression across prompts and generalization across domains.

To address the above two challenges, we propose COKE, an **Co**nfidence-derived **K**nowledge boundary **E**xpression method which teaches LLMs to express knowledge boundaries and decline unanswerable questions, leveraging their internal signals. Our method consists of two stages: a probing stage and a training stage. In the probing stage, we use the model's internal signals reflecting confidence to distinguish between answerable and unanswerable questions, avoiding reliance on external annotations. This allows for easy collection of large data and avoids conflicts between the model's internal knowledge and annotations. In the training stage, we construct prompts for each question using three representative types: prior awareness, direct awareness, and posterior awareness. Then, we apply regularization by incorporating the squared differences in confidence across different prompts for the same question into the loss function to enhance consistency. This training setup helps the model semantically learn to express knowledge boundary better, thereby enhancing its generalization ability.

To evaluate the model's knowledge boundary expression capability, we design an evaluation framework that comprehensively assesses the model's performance in both "knows" and "unknows" scenarios. We conduct extensive experiments on both in-domain and out-of-domain datasets. Results show that the model learns to use internal signals to help express knowledge boundary. Compared to directly using model signals for determination, the models trained with our method demonstrate better performance and generalization.

In summary, our contributions are:

- We explore which signals within the model itself can indicate the model's confidence, and find that using the minimum token probability signal from the model's response yields the best results.
- We propose a novel unsupervised method that leverages internal model signals and multi-prompt consistency regularization to enable the model to express its knowledge boundary clearly.
- We develop a framework for evaluating a model's ability to express its knowledge boundary, and experimental results demonstrate that the model can learn signals about the confidence of its knowledge and articulate its knowledge boundary.

## 2 Related Work

### 2.1 Knowledge Boundary Perception

While models are equipped with extensive parametric knowledge, some studies indicate their inability to discern the knowledge they possess from what they lack, thus failing to articulate their knowledge boundary (Yin et al., 2023; Ren et al., 2023). In terms of enhancing a model's awareness of its knowledge boundary, efforts can be categorized into two parts: one focuses on enabling the model to fully utilize its inherent knowledge, thereby shrinking the ratio of the model's "Unknown Knows" (Wei et al., 2022; Li et al., 2023; Tian et al., 2024). The other part focuses on enabling the model to acknowledge the knowledge it lacks, thereby reducing the ratio of the model's "Unknown Unknows". R-tuning (Zhang et al., 2023a) uses labeled data to judge the correctness of model responses and trains the model using the SFT method. Yang et al. (2023) and Kang et al. (2024) explore training methods based on RL. Focused on this aspect, our work investigates how to enable models to express knowledge boundaries without annotated data, while also considering consistent knowledge boundary expression across prompts and generalization across domains.

### 2.2 Uncertainty-based Hallucination Detection

Some work on hallucination detection focuses on obtaining calibrated confidence from LLMs. One segment of work involves utilizing the information from these models to compute a score that signifies

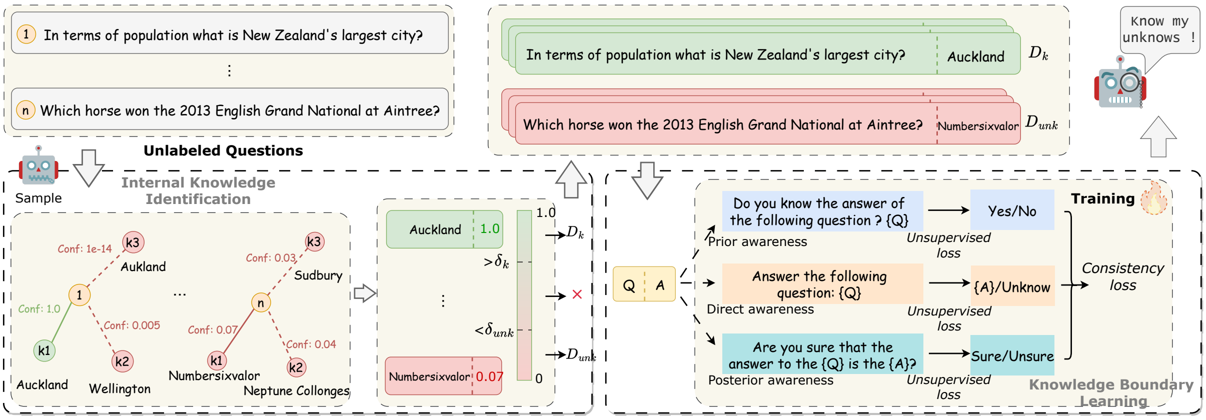


Figure 2: The procedure of COKE, which consists of two stages. In the first stage, the model makes predictions for unlabeled questions. We obtain two parts, $D_k$ and $D_{unk}$, based on the model confidence. In the second stage, we train with different prompts for the same question and use unsupervised loss and consistency loss to teach the model to express the knowledge boundary.

the model's uncertainty about knowledge (Manakul et al., 2023; Duan et al., 2023; Kuhn et al., 2023; Varshney et al., 2023). Another segment of work seeks to enable the model to express verbalized uncertainty (Lin et al., 2022; Xiong et al., 2023; Tian et al., 2023). Our work concentrates on enabling the model to explicitly express whether it is capable of answering, rather than generating a probability score. By allowing the model to express its knowledge boundary autonomously, users no longer need to concern themselves with detecting hallucinations, such as by setting uncertainty thresholds.

## 3 Knowledge Boundary Expression

### 3.1 Problem Formulation

We focus on exploring LLMs' capacity to perceive their internal knowledge. For a series of questions $Q = \{q_1, q_2, \ldots, q_n\}$, we categorize the questions based on whether the model has the knowledge required to answer them into two parts: questions that can be answered $Q_k$ and questions that cannot be answered $Q_{unk}$. To minimize the interference from the model's reasoning ability, the questions used for testing the model are all single-hop questions that inquire about factual knowledge. For a given question $q$, the model $M$ generates a prediction based on its parameter knowledge $K_\theta$, represented as $y = M(K_\theta, q)$. We measure the model's awareness of its knowledge from two aspects: the awareness of the knowledge it possesses and the knowledge it does not possess. The former is represented as the ratio of the model's "Know Knows" to "Knows", denoted as $R_k$, while the latter is represented as the ratio of the model's "Know Unknows" to "Unknows", denoted as $R_{unk}$. Given a question $q \in Q_k$, $R_K$ is set to 1 if the model's response $y$ aligns with the knowledge k, and to 0 if the model either expresses uncertainty or provides an incorrect answer. For a question where $q \in Q_{unk}$, $R_{unk}$ is assigned 1 if the model expresses uncertainty, and 0 if it fabricates an incorrect answer. We evaluate the model's awareness of its knowledge by testing on two types of $q$ and calculating $S_{aware} = \frac{1}{2}(R_k + R_{unk})$. The model's awareness of its knowledge is more accurate as $S_{aware}$ approaches 1, and less accurate as it approaches 0.

### 3.2 Method

Our insight is that the learning mechanism of LLM enables the model to search for the nearest knowledge $k$ in its parameters as the answer to the query $q$. Although training allows the model to measure distances accurately, it does not teach it to refuse to answer based on the distance. Therefore, we hope the model can learn to use its signals to recognize when a large distance indicates a lack of knowledge to answer $q$. Our method involves two steps as shown in Figure 2: First, we use the model's own signals to detect knows and unknows; Second, we guide the model to learn these signals through instruction tuning, enabling it to express its knowledge boundary clearly.

#### 3.2.1 Internal Knowledge Identification

To identify whether the model possesses the knowledge required to answer question $q$, we calculate the model's confidence about its prediction. The confidence of the model's prediction serves as a measure of the distance between query $q$ and knowledge $k$. On the unlabeled question set Q, we let model M generate phrase-form predictions for each question. We only consider the distance between query $q$ and the closest prediction; therefore, we use greedy decoding to obtain the prediction.

We use three model signals to represent the model's confidence: Min-Prob, Fst-Prob, and Prod-Prob. Min-Prob denotes the minimum probability among the $m$ tokens that make up the model's prediction, $c = min(p_1, p_2, ..., p_m)$. Fst-Prob and Prod-Prob respectively represent the probability of the first token in the prediction and the product of all probabilities. Two conservative thresholds, $\delta_k$ and $\delta_{unk}$, are established to decide whether the model has enough knowledge to answer a question. For questions with $c$ below the threshold $\delta_{unk}$, indicating the model is fabricating an answer due to insufficient knowledge, we define this subset as $D_{unk} = \{(q_i, y_i, c_i) \mid c_i < \delta_{unk}\}$ and use it to train the model to express its lack of knowledge. For questions with $c$ above the threshold $\delta_k$, indicating the model possesses the necessary knowledge, we define this subset as $D_k = \{(q_i, y_i, c_i) \mid c_i > \delta_k\}$ and use it to train the model to express that it knows the answer with increased confidence.

#### 3.2.2 Knowledge Boundary Expression Learning

We guide the model in learning to express its knowledge boundaries clearly based on its own signals through instruction tuning. We believe that the model's expression of knowledge boundary awareness should possess two properties: honesty and consistency. Honesty requires the model to express whether it knows the answer to a question based on its certainty about the knowledge. For instance, it should not answer "I don't know" to questions it is certain about. For honesty, we fine-tune the model on the dataset obtained in the first step, enabling the model to admit its ignorance on $D_{unk}$ and maintain its answers on $D_k$. Consistency requires the model to have the same semantic expression about whether it knows the same knowledge under different prompt formulations.

For consistency, we consider three different prompts for knowledge boundary awareness inquiries, which we refer to as prior awareness, direct awareness, and posterior awareness. **Prior awareness** involves the model assessing its ability to answer a question before actually providing an answer, with prompts like "`Do you know the answer to the question 'panda is a national animal of which country' honestly?`". **Direct awareness** involves the model responding directly to a query, supplying the answer if it possesses the knowledge, and admitting ignorance if it doesn't, with prompts like "`Answer the question 'panda is a national animal of which country' `". **Posterior awareness** involves the model's capacity to evaluate the certainty of its answers, with prompts like "`Are you sure that the answer to the 'panda is a national animal of which country' is 'China' `".

We hope that the model can express the same knowledge boundary under different prompts for the same question. It means that if the model determines that it possesses the knowledge under the prompt of prior awareness, it should be able to provide the answer when queried, and express confidence in its response when reflecting upon its answer. We teach the model to recognize its knowledge boundary by constructing three types of prompts for the same question. We incorporate the difference in probabilities of identical semantic responses under various prompts into the loss function, thereby ensuring the model's consistency across different prompts. Specifically, the loss function is defined as:

$$L = L_{unsup} + L_{con} \tag{1}$$

$$L_{con} = \sum_{1 \le i,j \le 3} \| P(y_i|x_i) - P(y_j|x_j) \|^2 \tag{2}$$

Previous research emphasizes that the MLP layer is a key component for storing knowledge in the transformer architecture LLM (De Cao et al., 2021; Meng et al., 2022). Guided by these insights, we only fine-tune the weight matrix of the attention layer using LoRA (Hu et al., 2022). This strategy allows us not to change the internal knowledge of the model, but just let the model learn to express the of knowledge boundary based on the confidence of the knowledge.

## 4 Experimental Setup

**Datasets** We consider three open-domain QA datasets: TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023). These datasets are

| Model | Method | TriviaQA | | | NQ | | | PopQA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $K_{aware}$ | $U_{aware}$ | $S_{aware}$ | $K_{aware}$ | $U_{aware}$ | $S_{aware}$ | $K_{aware}$ | $U_{aware}$ | $S_{aware}$ |
| Llama2-Chat-7B | Orig. | 100 | 0 | 50.0 | 100 | 0 | 50.0 | 100 | 0 | 50.0 |
| | Fine-tune | 93.9 | 6.2 | 50.1 | 88.6 | 3.1 | 45.8 | 93.5 | 1.9 | 47.7 |
| | IDK-FT | 80.8 | 78.0 | 79.4 | 45.5 | 87.6 | 66.6 | 62.8 | 83.6 | 73.2 |
| | *Uncertainty-Based* | | | | | | | | | |
| | Min-Prob | 61.8 | 86.2 | 74.0 | 33.4 | 91.4 | 62.4 | 57.7 | 89.3 | 73.5 |
| | Fst-Prob | 74.6 | 69.8 | 72.2 | 51.5 | 79.1 | 65.3 | 65.1 | 82.6 | 73.9 |
| | Prod-Prob | 66.0 | 84.7 | **75.3** | 39.8 | 90.2 | 65.0 | 61.0 | 87.7 | 74.4 |
| | *Prompt-Based* | | | | | | | | | |
| | Prior | 96.3 | 7.5 | 51.9 | 97.0 | 10.3 | 53.6 | 65.4 | 31.8 | 48.6 |
| | Posterior | 70.5 | 57.9 | 64.2 | 62.7 | 55.6 | 59.1 | 31.6 | 82.8 | 57.2 |
| | IC-IDK | 86.4 | 25.8 | 56.1 | 53.6 | 65.1 | 59.3 | 42.3 | 85.3 | 63.8 |
| | Verb | 14.3 | 95.8 | 55.1 | 17.5 | 95.0 | 56.3 | 17.6 | 97.3 | 57.4 |
| | CoKE | 76.1 | 74.0 | 75.0 | 56.0 | 84.2 | **70.1** | 71.1 | 83.0 | **77.0** |
| Llama2-Chat-13B | Orig. | 100 | 0 | 50.0 | 100 | 0 | 50.0 | 100 | 0 | 50.0 |
| | Fine-tune | 96.7 | 7.1 | 51.9 | 95.0 | 2.8 | 48.9 | 95.7 | 2.9 | 49.1 |
| | IDK-FT | 82.5 | 81.6 | 82.0 | 53.9 | 84.6 | 69.3 | 65.4 | 82.0 | 73.6 |
| | *Uncertainty-Based* | | | | | | | | | |
| | Min-Prob | 91.6 | 44.5 | 68.1 | 88.1 | 43.4 | 65.8 | 84.6 | 57.2 | 70.9 |
| | Fst-Prob | 92.9 | 34.1 | 63.5 | 90.6 | 30.7 | 60.7 | 87.4 | 51.0 | 69.2 |
| | Prod-Prob | 90.6 | 50.9 | 70.7 | 85.8 | 50.2 | 68.0 | 84.9 | 59.3 | 72.1 |
| | *Prompt-Based* | | | | | | | | | |
| | Prior | 88.6 | 14.2 | 51.4 | 81.3 | 26.5 | 53.9 | 38.2 | 81.8 | 60.0 |
| | Posterior | 100 | 0.30 | 50.0 | 100 | 0.0 | 50.0 | 100 | 0.10 | 50.0 |
| | IC-IDK | 99.7 | 1.5 | 50.6 | 96.8 | 6.7 | 51.7 | 90.8 | 25.1 | 58.0 |
| | Verb | 60.0 | 68.9 | 64.4 | 44.7 | 89.8 | 67.3 | 50.8 | 81.8 | 66.3 |
| | CoKE | 71.6 | 74.9 | **73.3** | 68.3 | 70.2 | **69.2** | 70.1 | 82.6 | **76.4** |

Table 1: Comparison of the performance of our method and the baseline method across an in-domain dataset (TriviaQA) and out-of-domain datasets (NQ and PopQA). We present results on two model scales: Llama2-Chat-7B and Llama2-Chat-13B.

| Model | TriviaQA | NQ | PopQA |
|---|---|---|---|
| Llama2-Chat-7B | 45.2 | 16.6 | 21.7 |
| Llama2-Chat-13B | 52.0 | 21.9 | 23.5 |

Table 2: The accuracy of LLMs on our test data. It represents the portion of knowledge that the model knows and can answer (Known Knows).

broad-coverage, knowledge-intensive QA datasets, making them well-suited for evaluating LLMs' capacity to perceive their internal knowledge. We utilize the train set of TriviaQA as our training data, treating it as unsupervised data by not using the labels. Natural Questions and PopQA serve as the out-of-domain test sets since they were not involved during the training process. We use a closed-book and free-form setup evaluating our approach on 2000 samples from each test set of three datasets. We use exact match to determine whether the model answers correctly or expresses the unknown.

| Metric | Definition |
|---|---|
| $K_{aware}$ | Proportion of *correct answers* on $T_k$ |
| $U_{aware}$ | Proportion of *expressions of unknown* or *correct answers* on $T_{unk}$ |
| $S_{aware}$ | $\frac{1}{2}(K_{aware} + U_{aware})$ |

Table 3: Knowledge awareness metrics.

**Metrics** As mentioned in the 3, we evaluate the model's awareness of its knowledge from two aspects: the awareness of the knowledge it possesses and the awareness of the knowledge it does not possess. Since we cannot directly access the model's internal knowledge $K_\theta$, we divide the test sets into two parts based on whether the model's predictions match the groundtruth: $T_k$ represents the "Known Knows" of the model (as shown in Table 2); $T_{unk}$ contains both the "Unknown Unknows" and "Unknown Knows" cases. We define the evaluation

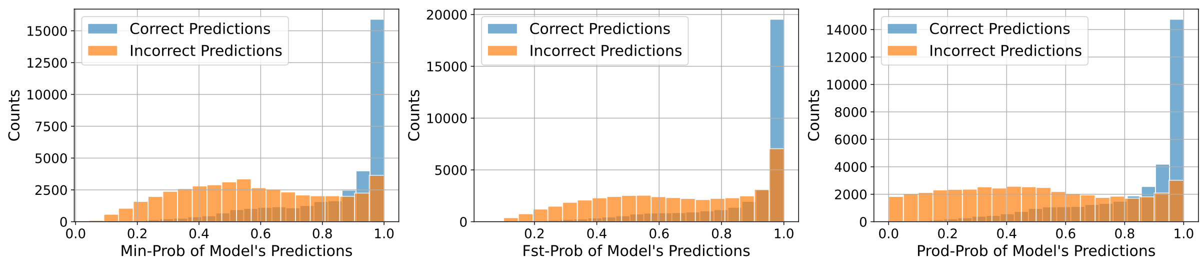


Figure 3: Distribution of model predictions regarding confidence for Llama2-Chat-7B on Trivia-QA. Confidence is calculated using Min-Prob, Fst-Prob, and Prod-Prob from left to right.

metrics as shown in Table 3.

**Baselines** We consider two different types of baselines: uncertainty-based methods and prompt-based methods. We also compared the original model (Orig.), the model fine-tuned with questions and their label (Fine-tune), and the model fine-tuned with question-label pairs, where responses to unknown questions are replaced by "Unknow" (IDK-FT).

The uncertainty-based methods obtain numerical confidence scores from the model's internal signals. Using labeled training data, we determine the optimal threshold for these scores that maximizes $S_{aware}$, and use this threshold to judge if the model knows the required knowledge for each question. The model's response consists of multiple tokens, and we experimented with three types of methods to calculate the final confidence score from the probabilities of these tokens:

- **Min token probability (Min-Prob)**: Use the smallest token probability in the model's prediction as the confidence score.
- **Product token probability (Prod-Prob)**: Use the product of the probabilities of all tokens in the model's prediction as the confidence score.
- **First token probability (Fst-Prob)**: Use the probability of the first token in the model's prediction as the confidence score.

The prompt-based methods use prompts to let models express their own knowledge boundary in natural language.

- **Prior prompt**: Similar to Ren et al. (2023) evaluating whether the model gives up on answering, we use the prompt "`Do you know the answer to the following question honestly? If you know, output Yes, otherwise output No, just say one word either Yes or No`" to directly ask the model if it knows the answer to the question.
- **Posterior prompt**: Kadavath et al. (2022) shows the model can evaluate the certainty of its answers. We use the prompt "`Are you sure that the answer to the following 'Q' is the following 'A'? If you are sure, output Sure, otherwise output Unsure, just say one word either Sure or Unsure`" to ask the model about the certainty of its answers.
- **In-context IDK (IC-IDK)**: Following Cohen et al. (2023), by integrating demonstrations into the prompt, we enable the model to express its knowledge boundary through in-context learning. These demonstrations include both the questions accurately answered by the model along with their responses, and the inaccurately answered questions, with their incorrect responses replaced by "Unknow".
- **Verbalize uncertainty (Verb)**: Resent work (Tian et al., 2023) suggest that LLMs' verbalized uncertainty exhibits a degree of calibration. We let the model output verbalized uncertainty, and search for the optimal threshold in the training set.

**Implementation Details** For our experiment, we choose to use the LLaMA2-Chat (Touvron et al., 2023) model. Based on the pre-trained LLaMA2 model, LLaMA2-Chat is a model that has undergone instruction tuning and RLHF, thereby acquiring the capability to follow instructions. We use the 7B and 13B versions of the LLaMA2-Chat model. In our approach, we sort the confidence scores calculated from the TriviaQA training set and designate the bottom 10% as $D_{unk}$ and the top 20% as $D_k$, collectively amounting to approximately 23,000 instances. We use LoRA for model fine-tuning, setting r=8, alpha=16, and dropout=0.05. During training, we set the initial learning rate to 1e-4, the final learning rate to 3e-4, the warmup phase to 300 steps, and we train for 700 steps. We

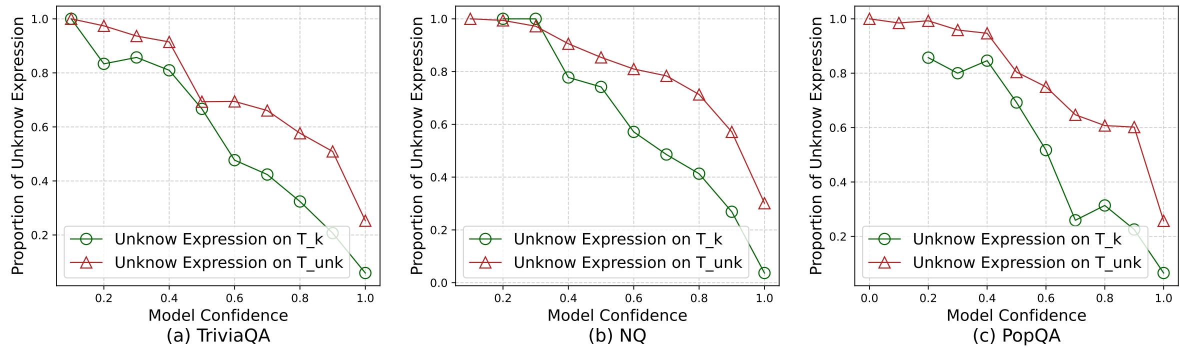


Figure 4: Model's "Unknow" expression ratio in question groups under different confidence scores (using minimum token probability). As the model's confidence score decreases, the ratio of "Unknow" expressions increases. The model exhibits a higher "Unknow" expression ratio on $T_{unk}$ compared to $T_k$.

conduct all our experiments on 4 NVIDIA A800 80GB GPUs.

## 5 Results and Analysis

### 5.1 Overall Performance

We present our main results on the in-domain and out-of-domain datasets in Table 1. Generally, we have the following findings:

Across all settings, we outperform prompt-based methods by a large gap. On Llama2-Chat-7B, our method obtains an $S_{aware}$ of 75.0 compared to $\leq$ 64.2 by prompt-based methods on TriviaQA, and obtains an $S_{aware}$ of 77.0 compared to $\leq 63.8$ by prompt-based methods on PopQA. Models struggle to accurately express knowledge boundaries when it comes to the prior prompt, in-context learning, and posterior prompts. Meanwhile, models can express verbalized uncertainty through prompts, and their accuracy improves with larger models, but remains limited for models with fewer than 13 billion parameters. Interestingly, as the model size increases, although the accuracy on the dataset improves, the model's ability for self-awareness does not show significant improvement in most cases. We believe that this capability might require even larger models to be evident.

Compared to uncertainty-based methods that leverage labeled data for threshold determination, our method can significantly outperform in most settings. This demonstrates that our method enables the model to effectively learn its confidence signals. Meanwhile, the model's performance surpasses the uncertainty-based methods that are used for training, indicating that the model can generalize and utilize information beyond the training signals. On out-of-domain datasets, our method significantly outperforms uncertainty-based methods, indicating that thresholds derived from a dataset have poor transferability, while our method exhibits better generalization.

| **Training Signal** | **TriviaQA** | **NQ** | **PopQA** |
|---|---|---|---|
| Fst-Prob | 74.9 | 69.3 | 76.2 |
| Prod-Prob | 73.9 | 69.8 | 76.3 |
| Min-Prob | **75.0** | **70.1** | **77.0** |

Table 4: Different signals serve as the model's confidence score in training the expression of knowledge boundary. The metric is represented by the $S_{aware}$.

Compared to IDK-FT, which uses labels to identify answerable and unanswerable questions, our method of using the model's own signals demonstrates better generalization. Although our method performs worse than IDK-FT on in-domain test sets, it significantly outperforms this supervised fine-tuning approach on out-of-domain datasets. This indicates that by leveraging the model's internal signals to teach LLMs to express knowledge boundaries, COKE not only avoids reliance on labeled data but also achieves better generalization.

### 5.2 Analysis

After demonstrating the effectiveness of our method, we conduct detailed analyses to further understand our method and find out why it works.

**Do signals effectively reflect model confidence?** We illustrate the effectiveness of the confidence calculation method through an empirical study. We obtain the model confidence for Llama2-chat-7B on the Trivia-QA training set using three different methods. We divide the model's responses into two parts based on whether the answers are correct and calculate the sample distribution for each part. As shown in Figure 3, there is a significant difference in the confidence distribution between the Correct Predictions and Incorrect Predictions. Predictions with confidence less than $0.4$ are mostly incorrect, while the confidence of correct predictions is generally $1.0$. This indicates that the model signals can reflect the model's confidence, implying whether the model possesses the corresponding knowledge.

**Have LLMs learned to use their signals?** To determine if our model uses confidence scores to express its knowledge boundary, we examined its responses under various confidence levels. Figure 4 shows the proportion of questions where the model responds with "Unknown" based on different confidence scores. We found that the model rarely responds with "Unknown" when confidence is high and frequently does so when confidence is low. For instance, with a confidence score below 0.4, the model almost always responds "Unknown", while near a score of 1.0, it confidently provides answers. This indicates the model effectively uses confidence scores to delineate its knowledge boundaries and generalizes well to out-of-domain data. Notably, the model responds "Unknown" more often at the same confidence level for out-of-domain questions compared to in-domain ones. This suggests the model has learned to use additional implicit information beyond just the confidence score. Training with this signal helps reduce noise from using minimum token probability alone and enhances performance compared to methods solely based on uncertainty.

**Which signal more accurately represents the confidence of LLMs?** We explore different signals in terms of their accuracy in reflecting the model's knowledge boundary and their impact on our method. As demonstrated in Table 1, in the uncertainty-based method, the performance variations using different signals are slight, with the multi-token probability production standing out as the best. As a training signal, the use of the minimum probability of multi-token outperforms other signals on both in-domain and out-of-domain datasets, as illustrated in Table 4. We consider that the minimum probability of multi-token is more easily mastered by the model. We leave the discovery of better signals reflecting the model's knowledge boundary and the utilization of multi-signal training for future work.

| Method | TriviaQA | | NQ | | PopQA | |
|---|---|---|---|---|---|---|
| | $S_{aware}$ | Con. | $S_{aware}$ | Con. | $S_{aware}$ | Con. |
| orig. | 50.0 | 53.4 | 50.0 | 45.6 | 50.0 | 17.7 |
| CoKE | 75.0 | **85.0** | **70.1** | **83.5** | **77.0** | **87.6** |
| w/o *Con-loss* | **75.6** | 42.0 | 69.2 | 45.0 | 74.8 | 60.6 |

Table 5: The consistency of the model's knowledge boundary expression under different prompts.

**What are the benefits of training the model with consistency loss?** We investigate the benefits of teaching a model to express knowledge boundary by using the strategy of constructing different prompts for the same question and applying a consistency regularization loss function. By adopting this strategy, we discover that it not only improves the model's ability to generalize, but also ensures a consistent expression of knowledge boundary under different prompts. Results from Table 5 indicate that the application of consistency loss, despite causing a slight decrease in $S_{aware}$ on the in-domain dataset, leads to substantial improvements on the out-of-domain dataset, thereby demonstrating enhanced generalization. We also reported the consistency of the model's expression of knowledge boundary under different prompts, as shown in Table 5. Here we focus on the model's expression consistency under prior prompts, posterior prompts, and direct inquiries. We notice that the model adopted with consistency loss is capable of expressing consistent knowledge boundaries for most questions under different prompts.

## 6 Conclusion

In this paper, we target the knowledge boundary awareness problem and propose CoKE, a novel unsupervised approach for this task. Our approach is built on detecting signals of the model expressing knowledge boundary, and teaching the model to use its own signals to express the idea of knowledge boundary. Through comprehensive experiments on in-domain and out-of-domain datasets, we show that our method can teach the model to use its own signals, significantly enhancing the model's ability to accurately express knowledge boundary. Our work can be extended by seeking more internal signals that better reflect the model's confidence and exploring how to combine these signals to train the model, inspiring further research into models autonomously improving their ability to express knowledge boundaries without human annotations.

## Limitations

We note three limitations of our current work. First is the accuracy of the evaluation methods. Because of the lack of a method to discover the internal knowledge of the model, we divided $T_k$ and $T_{unk}$ based on whether the model's answer matches the groundtruth, ignoring the impact of the model's erroneous beliefs. Another limitation is that to prevent exposure bias and the influence of multiple pieces of knowledge, we focused on the expression of knowledge boundary under short-form answers, without investigating the issue of long-form generation. Last, we focused on the model's ability to express the boundary of its internal knowledge, not extending to scenarios like self-awareness with external knowledge (e.g., RAG scenarios) or reasoning abilities (e.g., mathematics or logical reasoning).

## Ethical Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

**Risks** We propose COKE, which teaches models to express their knowledge boundaries using internal signals, thereby reducing hallucinations caused by fabricating answers when they do not know. Our experiments demonstrate that our method significantly reduces the instances of models fabricating answers to unknown questions. However, models may still occasionally produce fabricated answers in certain scenarios. Therefore, in practical applications, it is important to note that our method does not completely eliminate hallucinations, and there remains a risk of models generating fabricated content. Caution is advised in fields with stringent requirements.

## References

2023. John schulman - reinforcement learning from human feedback: Progress and challenges.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore. Association for Computational Linguistics.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

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, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

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*, 7:453–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023a. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.