Purging the Gray Zone: Latent-Geometric Denoising for Precise Knowledge Boundary Awareness

Anonymous ACL submission

Abstract

Large language models (LLMs) often exhibit hallucinations due to their inability to accurately perceive their own knowledge boundaries. Existing abstention fine-tuning methods typically partition datasets directly based on response accuracy, causing models to suffer from severe label noise near the decision boundaries and consequently exhibit high rates of abstentions or hallucinations. This paper adopts a latent space representation perspective, revealing a "gray zone" near the decision hyperplane where internal belief ambiguity constitutes the core performance bottleneck. Based on this insight, we propose GEODE (**Geo**metric **De**noising) framework for abstention fine-tuning. This method constructs a truth hyperplane using linear probes and performs "geometric denoising" by employing geometric distance as a continuous abstention decision confidence metric. This approach filters out ambiguous boundary samples while retaining high-fidelity signals for fine-tuning. Experiments across multiple models (Llama3, Qwen3) and benchmark datasets (TriviaQA, NQ, SciQ, SimpleQA) demonstrate that GEODE significantly enhances model truthfulness and exhibits outstanding generalization capabilities in out-of-distribution (OOD) scenarios.

1 Introduction

Large Language Models (LLMs) have demonstrated outstanding performance across various natural language processing tasks (Grattafiori et al., 2024; Yang et al., 2025). However, it is universally acknowledged that LLMs exhibit hallucination—that is, generating responses that are factually inaccurate or fabricating answers (Zhang et al., 2025a). This issue underscores the urgent need to develop effective hallucination detection and mitigation methods.

A practicable solution to mitigate hallucination phenomena involves fine-tuning LLMs to remain active when answering known questions and refrain from responding to queries beyond their knowledge scope (Wen et al., 2025; Li et al., 2025a). Specifically, these methods typically classify training data into "known" and "unknown" questions based on the correctness or accuracy of model responses, training models to answer the known set while replying "I don't know" to the unknown set (Zhang et al., 2024a; Cheng et al., 2024). To reduce reliance on ground truth labels, alternative approaches utilize uncertainty metrics like semantic entropy to partition training data into known and unknown questions (Tjandra et al., 2024; Xue et al., 2025).

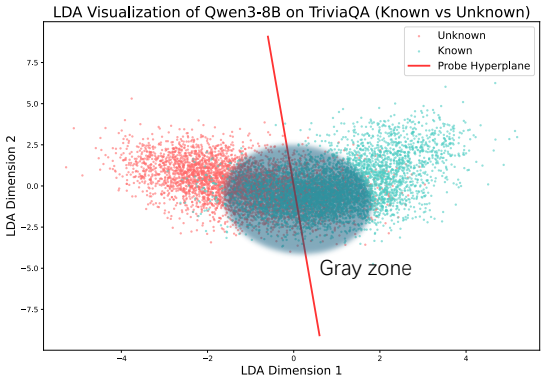

Figure 1: Visualization of the hidden states of questions that are known and unknwon to the model. "Gray zone" refers to the overlapping area.

This type of method has demonstrated significant effectiveness across various models. However, current abstention fine-tuning often fails when internal confidence misaligns with external correctness and vague internal beliefs. Relying on stochastic accuracy to partition "known" and "unknown" sets introduces significant label noise—such as "lucky guesses" or formatting-driven failures. Training on these noisy heuristics forces the model to learn a jagged, contradictory decision boundary, ultimately leading to over-refusal or persistent hallucinations.

We employ *probing* method to analyze such cases. As shown in Figure 1, we visualize the known and unknown sets, with the red line representing the boundary obtained through training. It can be observed that the two sets exhibit significant overlap. Notably, while the central region exhibits overlap and boundary crossing, this issue does not occur in data points farther from the hyperplane. This indicates that a significant portion of the current abstention fine-tuning data contains noise. Therefore, a potential improvement is to discard noisy data and retain clean data in the training set so that the LLM can learn to distinguish the known from unknown cases more effectively.

To this end, we propose **Geo**metric **De**noising (GEODE) method for abstention fine-tuning, inspired by the linear representation hypothesis. This approach selects effective samples for abstention fine-tuning through linear probes, thereby enhancing LLMs' self-awareness of their knowledge boundaries. The basic principle is, if the hidden state representation of a question is highly distant from the probe's hyperplane, then the model can readily decide to reject or answer it. However, if it is very close to the hyperplane, then it indicates a difficult case to decide whether to reject or answer. Guided by this principle, we select samples with high probe confidence (distant from the probe hyperplane) and discard those with low probe confidence (close to the probe hyperplane).

Our main contributions are as follows:

1. **Internal Representation Perspective**: We offer a novel diagnostic perspective on abstention fine-tuning by analyzing the latent space of LLMs. Our analysis reveals that suboptimal performance frequently stems from a "grey zone" near the latent decision boundary, where ambiguous representations introduce significant label noise.

2. **Latent-Guided Denoised Dataset Curation**: We propose GEODE, a framework that leverages internal probes to curate high-quality fine-tuning datasets. By using geometric distance from the truthfulness hyperplane as a confidence metric, GEODE purges ambiguous boundary samples. This geometric denoising ensures the model trains on linearly separable signals, leading to sharper knowledge boundaries.

3. **Empirical Superiority**: Extensive experiments across multiple architectures and benchmarks demonstrate that GEODE significantly outperforms baselines. Our method also shows superior generalization in out-of-distribution (OOD) and abstention tasks.

2 Related Work

2.1 Hidden States of LLMs

Recent work suggests there is a "truthfulness" direction in latent space (Marks and Tegmark, 2024; Azaria and Mitchell, 2023). Liu et al. (2024) suggest there is a universal truthfulness hyperplane that exists within LLMs generalizing on cross-task, cross-domain, and in-domain. Some work probes the last token of a question to predict whether the model can answer it correctly without generating any tokens (Slobodkin et al., 2023; Snyder et al., 2024; Gottesman and Geva, 2024). To more effectively distinguish facts from errors, some work designs more complex features to train truthfulness probes and utilize information from model-generated answers (Orgad et al., 2025; Li et al., 2025b; Zhang et al., 2025b). Li et al. (2025b); Orgad et al. (2025). Truthfulness vectors can also be employed for hallucination mitigation by steering (Ji et al., 2025; Zhang et al., 2024b). Recent works suggest that models' own internal judgments often lead to better overall factuality (Newman et al., 2025; Liang et al., 2024). In this work, we employ the truthfulness hyperplane as an internal confidence classifier to guide abstention fine-tuning for the model.

2.2 Abstention Fine-tuning

Abstention fine-tuning is a technique that teaches the model to abstain from answering questions that it does not know, while maintaining accuracy on known questions (Wen et al., 2025). Zhang et al. (2024a); Tjandra et al. (2024); Cheng et al. (2024) construct abstention-aware dataset based on whether the model can answer correctly, defining this as the model's knowledge boundary. Then they fine-tune the model to refuse answering questions beyond its knowledge boundary while responding to those within it. Xu et al. (2024); Cheng et al. (2024); Brahman et al. (2024) use Direct Preference Optimization (DPO) (Rafailov et al., 2023) to train models to admit uncertainty when encountering unknown questions rather than outputting incorrect answers. Li et al. (2025c) employ adaptive contrastive learning to optimize LLMs' abstention preferences. Huang et al. (2025); Cohen et al.

(2024) incorporate a dedicated "rejection" token into the model's vocabulary and formulate an objective function that redistributes probability mass toward this token when the model is uncertain. Zheng et al. (2025); An and Xu (2025) train models to output binary confidence labels ("sure" vs. "unsure") after generating an answer as a proxy for abstention, enabling them to reject low-confidence answers. Abstention fine-tuning may result in models being overly conservative or overly aggressive (Cheng et al., 2024; Zhu et al., 2025). In this work, we construct fine-tuning datasets based on the model's internal beliefs to mitigate issues of over-rejection and over-hallucination.

3 Method

Our work aims to develop a hidden state approach to enhance the model's awareness of its own knowledge boundaries. The main steps in our approach are to sample and arrange these diagnostic data into known/unknown subsets, train a hidden-state probe to quantify the model's confidence in its responses, and finally perform targeted abstention fine-tuning on the curated samples to teach the model to abstain from answering questions outside its knowledge scope.

3.1 Identify the Knowledge Boundary

First, the base LLM is prompted to answer every question within a source training dataset ($D0$). This acts as a diagnostic phase to examine what knowledge the model has actually internalized during its initial training. The model's responses are then split into two distinct subsets based on the accuracy in answering each question.

Known Knowledge ($D0_{\text{ik}}$): These are the samples that the LLM correctly answered. The "ik" stands for "I know". These question-answer pairs are retained in their original form to reinforce the retention of correct information during fine-tuning.

Unknown Knowledge ($D0_{\text{idk}}$): These are the samples that the LLM provided an incorrect response. For these cases, the original (incorrect) answer is discarded and replaced with a refusal message–specifically "I don't know" in abstention fine-tuning (hence, "idk" for short).

3.2 Curating Denoised Dataset

We train the probe model using $D0_{\text{ik}}$ and $D0_{\text{idk}}$ from the previous step. Specifically, we use the hidden states representation for the question-related string as the feature $\mathbf{x} = f_{\text{LLM}}(q)$, and the correctness of the statement as a binary label $y = \mathbb{I}(q \Rightarrow a) \in \{0, 1\}$, to train a logistic regression probe, whose formula is:

$$f_{\text{probe}}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b), \tag{1}$$

where σ is the sigmoid function, and w denotes the linear weight and b is the bias term. Next, we define the confidence of LLM answering question q with a by measuring the distance from $\mathbf{x}$ to the learned hyperplane $(\mathbf{w}; b)$:

$$d(\mathbf{x}) = \frac{\left|\mathbf{w}^\top \mathbf{x} + b\right|}{\|\mathbf{w}\|_2} \tag{2}$$

$d(\mathbf{x})$ has a clear probabilistic meaning that reflects the model's confidence in answering the given question. $d(x) > 0$ means the model believes it can answer correctly, while $d(x) < 0$ means the opposite, and a larger $|d(x)|$ value indicates stronger confidence. It follows naturally that we can reformulate the goal of abstention fine-tuning as teaching the model to reject questions for which $d(x) < 0$ and to answer those for which $d(x) > 0$. Accordingly, we partition the data into subsets of varying difficulty based on the magnitude of $|d(x)|$. Instead of using all available data, we select only samples farthest from the decision boundary, i.e., those with large $|d(x)|$, retaining the top $X\%$ of samples for the fine-tuning task.

The detailed methods for extracting hidden state from LLMs, $\mathbf{x} = f_{\text{LLM}}(q)$, are as follows:

Hidden state of the question (TBG): We directly feed the question string q into the model and extract the hidden state of the last token of the question from the final layer, e.g., **token before generation** (TBG), which corresponds to the token immediately preceding the generation trigger for question encoding.

Hidden state of the answer (SLT): We first generate the model's answer a via few-shot learning and greedy decoding (with temperature set to 0), and then feed concatenated sequence $q \oplus a$ to the model, and retrieve the hidden state of the last token from the model's final layer, e.g., **second last token** (SLT); this token captures the contextual representation of the entire question-answer sequence right before the end-of-sequence token.

3.3 Abstention Fine-tuning on Subset

Given a dataset D^{selected} that consists of a known question set $D^{\text{selected}}_{\text{ik}}$ and an unknown question set

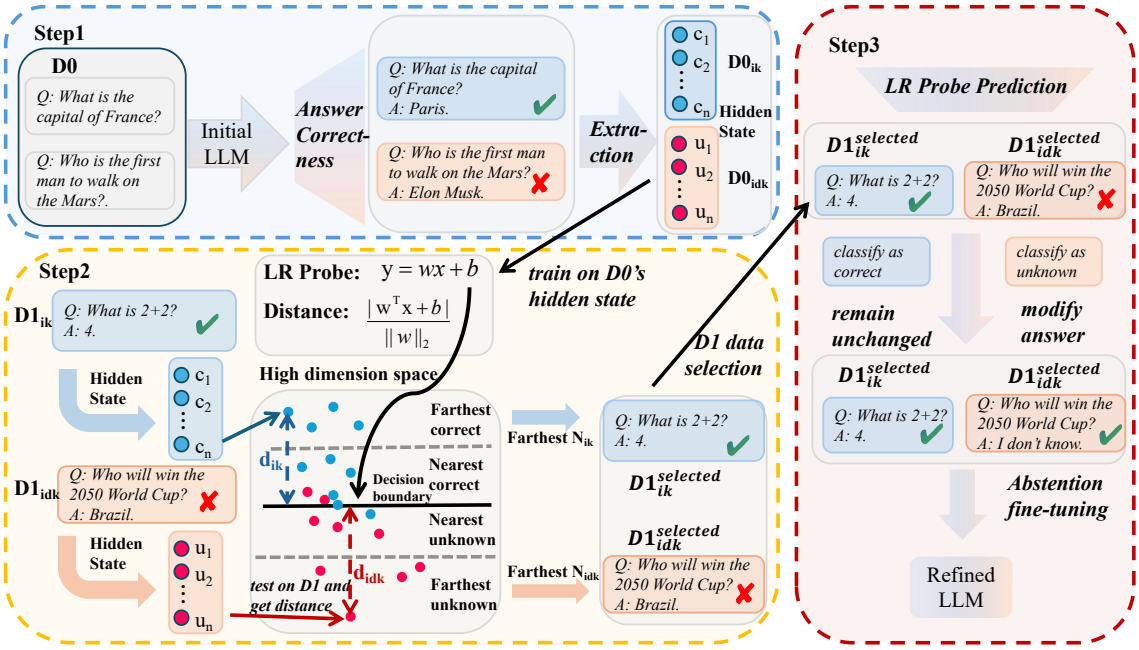

Figure 2: Overview of our method. GEODE contains three steps: (1) Identify the knowledge boundary by dividing the source data into two subsets ($D0_{ik}$ and $D0_{idk}$), and train a probe using these two subsets. (2) Calculate the distance from test data D1 to the hyperplane learned by the probe, and select the subset of samples that are farthest from the hyperplane $D1_{ik}^{selected}$ and $D1_{idk}^{selected}$. (3) Adjust the target answers based on the probe's prediction results (retain correct answers for $D1_{ik}^{selected}$ and replace incorrect answers with "I don't know" for $D1_{idk}^{selected}$), then conduct abstention fine-tuning on the selected subset.

$D_{idk}^{selected}$, we modify the ground truth of $D_{idk}^{selected}$ as "I don't know" and keep the ground truth of $D_{ik}^{selected}$ as the correct answer. Then we employ supervised fine-tuning with the cross-entropy loss:

$$\mathcal{L}_{(p_\theta)} = -\sum_{q \in D^{selected}} \sum_{t=1}^{|y^{(q)}|} \log p_\theta\left(y_t^{(q)} \mid \mathrm{I}, \mathbf{q}, \mathbf{y}_{t-1}^{(q)}\right), \tag{3}$$

in which $p_\theta(y_t^{(q)})$ is the model's predicted next-token probability distribution given the instruction (I), question (q), and the first $t-1$ tokens of the ground truth $\mathbf{y}_{t-1}^{(q)}$.

4 Experiments

4.1 Experimental Setting

Datasets We assess the effectiveness of GEODE on four open-ended question-answering tasks: **TriviaQA** (Joshi et al., 2017) which contains general knowledge QA pairs; **Natural Questions** (NQ) (Kwiatkowski et al., 2019) which contains questions from users' queries to search engines; **SciQ** (Welbl et al., 2017) which contains science exam questions across multiple disciplines; and **SimpleQA** (Wei et al., 2024) which is adversarially collected against GPT-4 responses, using short-fact

Algorithm 1 GEODE TBG Process

Require: Initial model M, QA dataset $D_{src} = D0 + D1$, percentage threshold X
Ensure: Fine-tuned model $M_{fine\text{-}tuned}$

Step 1: Identify Knowledge Boundary
  Test M on $D0$
  Split $D0$ into $D0_{ik}$ and $D0_{idk}$ by accuracy
Step 2: Curate Subsets Based on Distance
  Get representation $x = f_{LLM}(q)$, $q \in D0$
  Train linear probe on $D0$: $f_{probe}(x) = \sigma(w^\top x + b)$
  Compute distance on $D1$: $d(x) = \frac{|w^\top x + b|}{||w||_2}$
  Determine threshold θ as the $X\%$-th quantile of sorted $d(x)$
  Select top $X\%$ samples: $D1^{selected} \leftarrow \{x \mid |d(x)| > \theta\}$
Step 3: Abstention Fine-tuning
  Split $D1^{selected}$ into $D1_{ik}^{selected}$ $(d(x) > 0)$ and $D1_{idk}^{selected}$ $(d(x) < 0)$.
  Replace ground truth of $D1_{idk}^{selected}$ with "I don't know."
  Fine-tune M with cross-entropy loss on $D1^{selected}$
  Return $M_{fine\text{-}tuned}$

question-answering with single, indisputable answers to test whether the model truly "knows what it knows". We use 10K samples from the TriviaQA training set for probe training, with the rest used for SFT. The validation split of TriviaQA is used for the in-domain (ID) test. We use NQ, SciQ, and SimpleQA for out-of-distribution (OOD) tests. Detailed evaluation dataset information is provided in Appendix A.

Baselines We compare GEODE with the following existing methods for abstention fine-tuning.

1. **IDK (Cheng et al., 2024)** directly prompts the model to abstain from uncertain questions.
2. **Uncertainty (Xu et al., 2025)** first prompts the model to answer questions as accurately as possible, then prompts the model to output binary uncertainty (sure or unsure).
3. **R-Tuning (Zhang et al., 2024a)** randomly selects a set of questions, categorizes them as known or unknown based on the model's accuracy on each question, changes the ground truth for unknown questions to "I don't know," and then performs supervised fine-tuning.
4. **Probe-Tuning (Newman et al., 2025)**. The workflow of Probe-Tuning is identical to that of R-Tuning, with the key difference being that Probe-Pred utilizes the prediction results from truthful probes as the criteria for classifying questions into known and unknown categories.

GT \ R	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.

Evaluation For each test question, we classify the response as correct, wrong, or abstention. We use Llama3.1-8b-instruct (Grattafiori et al., 2024) as the judge to evaluate the correctness of answers generated by LLMs with 6-shot prompting. A response containing "I don't know" is counted as an abstention. To measure the performance of abstention fine-tuning methods, we adopt three metrics, each reflecting unique aspects of performance, based on the widely used abstention confusion matrix in Table 1. We use in-context learning (**ICL**) as the initial model.

Helpfulness ($\mathbf{F1_{ans}}$): For known questions, we calculate $F1_{ans}$ (Kim et al., 2024) as the harmonic mean of answerable recall ($\frac{N_1}{N_1+N_2+N_3}$) and answerable precision ($\frac{N_1}{N_1+N_2+N_4}$).

Truthfulness ($\mathbf{F1_{abs}}$): For unknown questions, we calculate $F1_{abs}$ (Kim et al., 2024) as the harmonic mean of unanswerable recall ($\frac{N_5}{N_4+N_5}$) and unanswerable precision ($\frac{N_5}{N_3+N_5}$).

Reliability ($\mathbf{F1_{rel}}$): Existing studies indicate that enhancing helpfulness leads to a decline in factuality (Xu et al., 2024; An and Xu, 2025). Therefore, we calculate the harmonic mean of metrics $F1_{ans}$ and $F1_{abs}$ as a reliability metric for comprehensive evaluation (An and Xu, 2025).

Implementation Details In this work, we choose Llama3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) as the initial models. We conduct experiments using SFT. We set the X as 25%. We employ logistic regression probe with L2 regularization. We use the Swift[1] framework to conduct fine-tuning using the AdamW optimizer, setting epoch to 3, learning rate to 1e-5, and batch size to 16. We use grid search to select the optimal hyperparameters. All the experiments are implemented on 4 Nvidia L40-48GB GPUs. During inference, we utilize the vLLM framework[2] to accelerate the process and employ a greedy search strategy to generate responses. Results are averaged over three different random seeds.

4.2 Main Results

We show the main experimental results of GEODE and all baseline methods in Table 2. The key observations from our experiments are:

Effectiveness of GEODE The fundamental limitation of existing baselines lies in their susceptibility to the "gray zone"—the latent region where internal belief is ambiguous and misaligned with external correctness. As illustrated in Figure 1, partitioning data based solely on response accuracy results in significant overlap between known and unknown representations. By utilizing the geometric distance from the truthfulness hyperplane, GEODE effectively purges this noise. Our results demonstrate that this denoising process allows the

[1] https://github.com/modelscope/ms-swift
[2] https://github.com/vllm-project/vllm

Dataset	TriviaQA			NQ			SciQ			SimpleQA		
Method	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$
Llama3-8B-Instruct												
IDK	78.7	35.3	48.8	**61.4**	56.3	58.7	**83.3**	5.6	10.4	14.6	64.0	23.8
Uncertainty	67.7	46.4	55.0	45.9	18.0	25.9	64.7	17.9	28.0	10.0	52.2	16.8
R-Tuning	77.3	71.6	74.4	47.0	78.1	58.7	69.9	58.2	63.5	14.6	**96.1**	25.4
Probe-Tuning TBG	78.8	72.8	75.7	51.6	79.1	62.5	81.4	53.4	64.5	12.8	**96.1**	22.6
GEODE TBG	**80.9**	**73.7**	**77.1**	54.0	78.4	**64.0**	81.9	58.5	68.3	18.4	94.8	30.7
Probe-Tuning SLT	75.8	71.3	73.4	44.8	78.3	56.9	75.4	58.9	66.1	18.4	95.8	30.9
GEODE SLT	79.8	**73.7**	76.7	52.6	**79.4**	63.3	82.4	**59.8**	**69.3**	**18.6**	96.0	**31.2**
Qwen3-8B												
IDK	74.0	55.1	63.2	55.1	65.4	59.8	81.8	44.5	57.6	11.7	58.6	19.5
Uncertainty	75.9	38.8	51.3	**57.3**	52.1	54.6	70.3	38.6	49.8	12.2	66.6	20.6
R-Tuning	75.8	74.3	75.0	50.3	70.5	58.7	**86.5**	45.1	59.3	12.6	63.6	21.0
Probe-Tuning TBG	75.0	71.7	73.3	51.2	70.5	59.4	**86.5**	44.2	58.5	12.6	63.4	21.0
GEODE TBG	**77.7**	**77.4**	**77.6**	54.3	71.3	**61.6**	81.8	47.2	59.8	13.0	**67.3**	21.7
Probe-Tuning SLT	75.4	72.8	74.1	50.1	69.2	58.1	85.7	45.8	60.0	12.3	62.0	20.6
GEODE SLT	75.6	75.6	75.6	51.6	**71.7**	60.0	84.6	**48.5**	**61.6**	**13.8**	65.9	**22.8**

Table 2: Performance on in-domain and out-of-domain question answering benchmarks. All results are multiplied by 100. The best result is bolded. The second best is underlined.

model to fine-tune on pure representations of its knowledge boundary. Consequently, GEODE consistently outperforms all baselines in the reliability metric $F1_{rel}$) across both Llama3-8B-Instruct and Qwen3-8B. For instance, on the TriviaQA in-domain task, GEODE (TBG) achieves a top $F1_{rel}$ of 77.1 for Llama3 and 77.6 for Qwen, representing a significant margin over traditional R-Tuning and Probe-Tuning. GEODE also exhibits exceptional OOD generalization. On NQ, SciQ and SimpleQA, GEODE consistently maintains high $F1_{rel}$ scores.

Latent Representations: TBG vs. SLT TBG and SLT exhibit comparable performance across benchmarks, with no single strategy consistently outperforming the other. While their absolute gains are similar, they reflect cognitive states at different stages: TBG captures the model's initial confidence prior to output, whereas SLT incorporates the semantic context of the generated response. This result underscores the robustness of the GEODE framework regarding representation positioning. It demonstrates that geometric denoising effectively identifies and reduces noise regardless of the specific extraction point, confirming the universal utility of latent-geometric distance.

4.3 Evaluation on RAG Setting

Experimental Setting We evaluate the performance of the abstention methods in the Retrieval Augmented Generation (RAG) scenario. We use the RAG-Bench (Fang et al., 2024) dataset, which includes two settings: (1) **Golden**: golden retrieval, which contains contexts that include correct answers. (2) **Golden & RRN**: golden retrieval with relevant retrieval noise, which includes golden retrieval and context relevant to the question statement but lacks the correct answers. We feed the context alongside the question into the model.

Experimental Results Table 3 shows the performance in the RAG setting. Similar to the main experimental results, IDK and R-Tuning performed worse than abstention fine-tuning based on probing. Overall, our method achieves the best performance relative to the baseline. Within the SLT-based setting, our method not only achieves superior accuracy but also yields a lower hallucination rate compared to Probe-Tuning. This dual improvement highlights the efficacy of geometric denoising in establishing more reliable knowledge boundaries. Our method maintains the highest reliability ($F1_{rel}$) even under noisy retrieval scenarios, demonstrating its robustness in practical applications.

	Golden					Golden & RRN				
	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	Acc.	Hallu.	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	Acc.	Hallu.
ICL	-	-	-	71.2	28.8	-	-	-	63.8	36.2
IDK	76.1	45.5	57.0	58.1	17.9	71.6	45.6	55.7	52.7	23.6
R-Tuning	80.8	52.5	63.7	61.0	11.8	74.4	53.3	62.1	50.4	15.3
Probe-tuning TBG	83.8	47.9	61.0	67.3	13.0	81.1	48.6	60.8	63.3	16.8
GEODE TBG	75.2	52.5	61.8	51.5	9.2	74.3	56.9	64.5	49.1	11.4
Probe-Tuning SLT	79.1	52.7	63.2	54.8	11.9	75.9	54.7	63.5	50.0	15.7
GEODE SLT	79.9	61.1	69.3	55.8	9.2	75.8	57.9	65.7	53.4	13.0

Table 3: RAG results. Acc. is accuracy. Hallu. is hallucination rate.

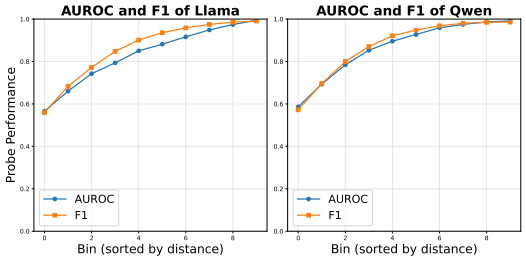

Figure 3: Probing performance vs. distance. Bin0 denotes the nearest subset, and bin9 the farthest subset. Accuracy and F1 score increase as distance increases.

Dataset	TriviaQA			NQ		
Metric	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$	$F1_{ans}$	$F1_{abs}$	$F1_{rel}$
Ours TBG	**77.7**	**77.4**	**77.6**	**54.3**	71.3	**61.6**
-middle	74.9	74.2	74.6	50.0	70.2	58.4
-nearest	75.2	71.2	73.2	49.7	69.6	58.0
Ours SLT	75.6	75.6	75.6	51.6	71.7	60.0
-middle	75.4	72.8	74.1	50.1	69.2	58.1
-nearest	73.8	67.5	70.5	48.9	67.2	56.6
R-Tuning-01	75.5	77.0	76.2	51.6	**71.8**	60.0

Table 4: Ablation study on distance-based data partitioning for abstention fine-tuning. Samples are partitioned into three tiers based on their geometric distance $|d(x)|$ to the probing hyperplane.

5 Analysis

5.1 Ablation Studies

To validate our hypothesis that the geometric distance to the probing hyperplane serves as a reliable proxy for sample quality, we partitioned the training data into three tiers: Ours-Farthest (75%-100%), Ours-Middle (37.5%-62.5%), and Ours-Nearest (0-25%). As shown in Table 4, a clear performance gradient is observed across all metrics. Specifically, on Qwen3-8B (TriviaQA), Ours-Farthest achieves an $F1_{rel}$ of 77.6, markedly surpassing the Middle (74.6) and Nearest (73.2) tiers. This performance decay is inherently rooted in the noise density of the decision boundary: samples located near the hyperplane (the "grey zone") represent instances where the model's internal representations for known and unknown knowledge are highly entangled. Fine-tuning on these ambiguous samples introduces contradictory gradients that blur the model's knowledge boundaries. By selectively training on the "farthest" samples, GEODE effectively performs geometric denoising, ensuring that the model learns from high-confidence, linearly separable signals. To further illustrate this distance-quality correlation, we bin the training data into ten equal subsets sorted by distance in ascending order. As visualized in Figure 3, the probe's predictive accuracy scales monotonically with distance. Notably, the AUROC for samples nearest to the hyperplane drops below 0.6, approaching random chance, confirming that proximity to the boundary is a primary source of aleatoric noise. In contrast, distal samples provide a clean signal for knowledge boundaries.

We further compare this geometric denoising approach with R-Tuning-01, which conceptually shares a similar objective by training only on samples with 100% response consistency (either all correct or all incorrect) across 10 times of independent samples. While R-Tuning-01 attempts to filter ambiguity via external output stability, it remains inferior to our method.

5.2 Effects of Positive Proportions in Training

We fix the training set size at 20,000 and conduct experiments under different positive sample pro-

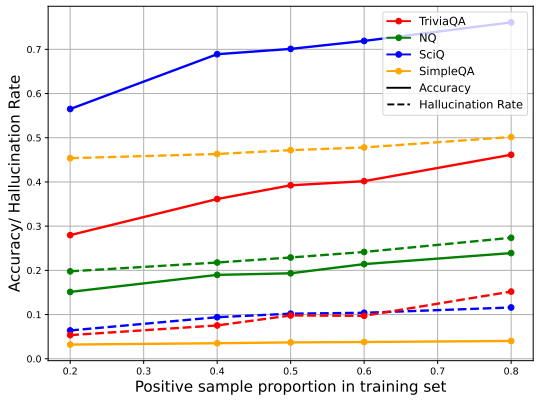

Figure 4: Accuracy and hallucination rate according to different positive proportions in the training set.

Method	Alcuna	FalseQA	SA
Llama3-8b-Instruction			
IDK	78.4	49.8	50.9
Uncertainty	19.2	8.5	15.4
R-Tuning	98.3	91.8	98.1
Probe-Tuning TBG	99.2	91.8	**99.7**
GEODE TBG	98.2	92.4	99.2
Probe-Tuning SLT	**99.7**	**95.3**	**99.7**
GEODE SLT	99.0	90.6	98.7
Qwen3-8B			
IDK	85.9	62.4	74.4
Uncertainty	38.5	24.5	39.6
R-Tuning	78.2	67.5	82.6
Probe-Tuning TBG	84.0	**72.2**	84.1
GEODE TBG	**94.5**	67.6	**88.3**
Probe-Tuning SLT	69.7	65.2	80.6
GEODE SLT	94.3	70.5	86.9

Table 5: Abstention rates (%) across specialized benchmarks. High scores indicate effective identification of unanswerable or deceptive queries.

portions. As shown in Figure 4, both accuracy and hallucination rates increase as the proportion of positive samples in the training set go up. High-accuracy datasets like TriviaQA and SciQ exhibit significant sensitivity to changes in the positive-to-negative ratio within the training set, whereas NQ and SimpleQA show less pronounced variation. This suggests that excessively high negative sample ratios may cause models to over-abstain from known questions. The results also imply that abstention fine-tuning cannot eliminate over-abstention and hallucination simultaneously.

5.3 Evaluation on Unanswerable Datasets

Experimental Setup To evaluate the model's ability for identifying unanswerable queries, we test performance on three specialized benchmarks: (1) **Alcuna** (Yin et al., 2023a), containing synthetic entity-based questions; (2) **FalseQA** (Hu et al., 2023), containing questions contradicting common sense; and (3) **Self-Aware (SA)** (Yin et al., 2023b), containing inherently unanswerable questions.

Performance Analysis As presented in Table 5, GEODE demonstrates highly competitive abstention rates, particularly with the TBG variant. While performance varies across benchmarks, our method shows notable strengths in specific scenarios. For instance, on the Alcuna dataset using Qwen3-8B, GEODE (TBG) achieves an abstention rate of 94.5%, a substantial improvement over the 78.2% reached by R-Tuning. We found that the rejection rate on Alcuna was higher than that on FalseQA and SA. Notably, overall rejection rates are higher on Alcuna than on FalseQA and SA. We attribute this to the nature of entities: while FalseQA and SA involve real-world concepts, Alcuna uses synthetic ones. This suggests that familiar domains trigger an illusion of knowledge, where the presence of known entities induces generative overconfidence. Consequently, models struggle more to recognize their epistemic limits in familiar contexts than in entirely novel, synthetic ones.

6 Conclusion

In this work, we introduce GEODE, a novel framework for abstention fine-tuning that leverages the latent geometry of LLMs. Moving beyond traditional methods that rely solely on external response accuracy, we propose a diagnostic perspective by analyzing the model's internal representation space. We found a critical "grey zone" near the latent decision hyperplane, where ambiguous internal beliefs introduce significant noise that hinders a model's ability to perceive its own knowledge boundaries. By employing geometric denoising, GEODE systematically purges these ambiguous boundary samples, ensuring that the model is fine-tuned on high-fidelity, linearly separable signals. Extensive experiments across multiple models (Llama3, Qwen3) and benchmarks demonstrate that our approach significantly enhances model reliability and exhibits superior generalization in OOD, RAG, and deceptive scenarios.

Limitations

Our method has been validated on models with up to 8B parameters; however, it has not been evaluated on models with larger parameter sizes. We employ linear representations to model truthfulness, but linear methods may not fully capture the complexity of truth-related behaviors, and a multi-dimensional framework is necessary for more accurate modeling in larger models as suggested by recent evidence (Yu et al., 2025). Additionally, our experiments have yet to include reasoning tasks or long-form generation, which are critical for further evaluating the robustness and scalability of our approach in addressing more intricate linguistic challenges. Although this work employs highly discriminative data for fine-tuning, abstention fine-tuning carries an unavoidable risk: over-abstention, which may cause the model to refuse to output answers it was highly confident about prior to fine-tuning. Future research may need to continuously teach models to express uncertainty during the pre-training stage.

References

Hao An and Yang Xu. 2025. Teaching llms to abstain via fine-grained semantic confidence reward. *Preprint*, arXiv:2510.24020.

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*.

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.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. 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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

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.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3769–3793, Suzhou, China. 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.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sanggoo 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.

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.

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, and Fakhri Karray. 2025b. HD-NDEs: Neural differential equations for hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6186, 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. 2025c. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *Preprint*, arXiv:2401.15449.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024. On the universal truthfulness hyperplane inside LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224, Miami, Florida, USA. Association for Computational Linguistics.

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*.

Benjamin Newman, Abhilasha Ravichander, Jaehun Jung, Rui Xin, Hamish Ivison, Yegor Kuznetsov, Pang Wei Koh, and Yejin Choi. 2025. The curious case of factuality finetuning: Models' internal beliefs can improve factuality. *Preprint*, arXiv:2507.08371.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

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.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 2721–2732, New York, NY, USA. Association for Computing Machinery.

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*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

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.

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. 2024. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

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, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xunjian Yin, Baizhou Huang, and Xiaojun Wan. 2023a. ALCUNA: Large language models meet new knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1397–1414, Singapore. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023b. 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.

Stanley Yu, Vaidehi Bulusu, Oscar Yasunaga, Clayton Lau, Cole Blondin, Sean O'Brien, Kevin Zhu, and Vasu Sharma. 2025. From directions to cones: Exploring multidimensional representations of propositional facts in llms. *Preprint*, arXiv:2505.21800.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. 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.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024b. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025a. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025b. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17986–18002, Vienna, Austria. Association for Computational Linguistics.

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.

A Dataset Details

The table displays statistical information about the datasets, where SimpleQA and NQ are more challenging datasets, while TriviaQA and NQ are relatively simpler. Alcuna, FalseQA and Self-Aware contain unanswerable questions.

	Size	Llama Acc.	Qwen Acc.
TriviaQA Train	87622	66.17	54.93
TriviaQA Val	11313	65.55	54.65
NQ	3610	40.86	34.04
SciQ	1000	72.50	81.60
SimpleQA	4326	6.80	5.59
Alcuna	2001	-	-
FalseQA	1374	-	-
Self-Aware	1032	-	-

Table 6: Dataset details. Llama (Llama3-8B-Instruct) and Qwen (Qwen3-8B) are the initial models used in experiments. Acc. (Accuracy) is for reference.

B Prompts

During training, we use the following instruction:

You are a helpful and truthful AI assistant. You should answer the question as briefly as possible, if you don't know, please just say 'I don't know.".

We use a 6-shot prompt for evaluation 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:

Abstention-aware 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:

Uncertainty Prompt

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:

LLM judge Prompt

We are assessing the quality of answers to the following question: *input question*
The following are expected answers to this question: *input ground truth*
The proposed answer is *proposed answer*
Within the context of the question, does the proposed answer mean the same as the expected answer?
Respond only with yes or no.
Here are some examples:
Question: *demo question 1*
Expected answer: *demo ground truth 1*
Proposed answer: *demo correct answer 1*
Response: *yes*

Question: *demo question 2*
Expected answer: *demo ground truth 2*
Proposed answer: *demo wrong answer 2*
Response: *no*

...

Now evaluate the following:
Question: *input question*
Expected answer: *input ground truth*
Proposed answer: *input proposed answer*
Response: