Can RAG Models Know What They Don't Know? Analyzing and Improving Knowledge Boundary Perception

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) provides models with external knowledge to help mitigate hallucinations, but this external knowledge may contain irrelevant, distracting, or conflicting contents. This paper investigates the impact of external knowledge on model's internal perception of knowledge boundaries. We first conduct experiments to compare different detection methods with and without external documents, which reveal that external knowledge impairs models' ability to distinguish between known and unknown information, causing them to treat the unknown as known. Building on this finding, we refine training strategies to enhance the perception of knowledge boundary and propose a knowledge-boundary-based controlled generation framework. This enables models to dynamically determine knowledge reliance and reject unknown questions. Experiments demonstrate that our framework substantially improves generation quality with negligible additional training overhead. Code is submitted with the paper and will be publicly available.

1 Introduction

Large Language Models (LLMs) internalize extensive knowledge during training and show strong performance across diverse tasks (Yang et al., 2024a; Dubey et al., 2024; Yang et al., 2025). Retrieval-augmented Generation (RAG) supplies external knowledge to help models integrate new information and reduce hallucinations (Gao et al., 2023; Asai et al., 2023). However, external knowledge may contain contents that interfere with or conflict with internal knowledge, adversely affecting output quality (Shi et al., 2023; Xu et al., 2024).

Enabling models to balance between different knowledge sources, answer based on correct knowledge, and decline questions beyond their scope is a requirement for model honesty (Li et al., 2024; Wen et al., 2025). Existing works often require additional training to help models filter irrelevant documents (Pan et al., 2024; Yoran et al., 2024), refuse unknown questions (Zhang et al., 2024a; Yang et al., 2024b), or express knowledge boundaries (Chen et al., 2024; Xue et al., 2025). In contrast, this paper explores a solution that leverages a model's inherent perception of knowledge boundaries, eliminating the need for further training.

A model's perception of the knowledge boundaries refers to its ability to distinguish between the known and the unknown, often assessed through uncertainty estimation methods (Li et al., 2025). These techniques, which vary by source and computation, typically assign an uncertainty or confidence score to the model's output (Huang et al., 2024; Xia et al., 2025; Vashurin et al., 2024). A low score suggests low confidence, leading the model to treat the question as unknown. However, existing approaches primarily evaluate uncertainty based on the question alone, without accounting for the influence of external documents.

This observation motivates our first research question: *How do external documents influence a model's perception of knowledge boundaries?* To address this, we empirically investigate the performance of multiple uncertainty estimation methods with and without external documents. The experimental results indicate that external documents interfere with a model's ability to distinguish unknown from known information, making the model more likely to treat unknown as known. Consequently, learning to reject unknown questions becomes more challenging in RAG settings.

This limitation motivates our second question: *How can we enhance the perception and utilization of knowledge boundaries in RAG?* To address this, we first introduce refined training strategies to improve estimators based on internal hidden states, thereby enhancing the perception of knowledge boundaries. Building on this, we propose a knowledge-boundary-based controlled generation framework that integrates prior and posterior con-

trols, enabling the model to dynamically adjust knowledge reliance, reject unknown queries, and better utilize documents to answer known questions. The estimator within our framework is lightweight, requiring only minimal training data. The framework avoids retraining the large generator and adds only a modest inference overhead of a few tokens, yet yields substantial gains in generation quality. Experiments on knowledge boundary datasets show that our approach outperforms baselines, significantly improving output quality while enabling reliable refusals without compromising the validity of responses.

2 Related Works

2.1 Abstain in RAG

While Retrieval-Augmented Generation (RAG) supplies models with external knowledge, this knowledge can be irrelevant, distracting, or contradictory to a model's internal knowledge (Shi et al., 2023; Yoran et al., 2024; Wu et al., 2024). This interference complicates a model's ability to abstain from answering questions beyond its knowledge scope (Wen et al., 2025). Previous research on models' abstain has proposed constructing specialized datasets (Peng et al., 2025; Amayuelas et al., 2024; Kirichenko et al., 2025) or employing methods such as supervised fine-tuning (Yang et al., 2024b; Zhang et al., 2024a) and reinforcement learning (Cheng et al., 2024; Kang et al., 2025) to teach models to reject uncertain queries, thereby reducing hallucinations and improving reliability. In contrast, this paper focuses on the inference stage and does not require additional model training to achieve reasonable abstention.

2.2 Knowledge Boundary

Models must balance and integrate internal and external knowledge (Bi et al., 2025a; Xu et al., 2024). Existing approaches address this by aligning the model to factuality (Tian et al., 2024; Li et al., 2023; Zhang et al., 2024b) or context faithfulness (Huang et al., 2025; Zhou et al., 2023), or adjusting the weighting towards external knowledge during the decoding phase (Shi et al., 2024b; Bi et al., 2025b; Jin et al., 2024). We achieve this by leveraging the model's perception of knowledge boundaries (Zeng et al., 2024; Chen et al., 2024; Ashok and May, 2025a; Zhou et al., 2025). The perception of the knowledge boundaries in an LLM lies in distinguishing between the known and the unknown (Li et al., 2025). Since knowledge itself is opaque within LLMs, research often shifts focus from "knowledge" to "questions," determining whether a model should abstain based on its capacity to answer correctly (Yang et al., 2024b; Sun et al., 2025; Wen et al., 2025).

2.3 Uncertainty Estimation

Assessing whether a model can answer a given question involves uncertainty estimation (Xia et al., 2025; Huang et al., 2024; Vashurin et al., 2024). The objective of uncertainty estimation is to quantify a model's own uncertainty during generation. Related methodologies can be categorized as follows: verbalizing methods (Tian et al., 2023; Ren et al., 2023), latent information methods, consistency-based methods (Lyu et al., 2025; Lin et al., 2024) and semantic clustering methods (Kuhn et al., 2023; Nikitin et al., 2024). The uncertainty score is sometimes also referred to as confidence score. While distinctions exist between the two in certain scenarios (Huang et al., 2024; Lin et al., 2024), this paper does not involve multi-sample generation results, so both terms convey the same meaning and are used interchangeably. We focus on latent information methods, which avoid repeated sampling and incur lower computational costs, providing uncertainty scores based on either the model's internal state or its generated outputs (Ni et al., 2025; Sriramanan et al., 2024).

3 Investigation

In this section, we address the research question: *How do external documents influence a model's perception of knowledge boundaries?*

3.1 Preliminaries

Building on the NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017) and WebQ (Berant et al., 2013) datasets, Sun et al. (2025) proposes dividing data into four quadrants based on the boundaries of internal and external knowledge within the model. The internal boundary is defined by the model's performance without retrieved documents, while the external boundary is annotated according to the provided documents. Consequently, differences in a model's internal knowledge result in varying quadrant assignments. Due to resource constraints, we primarily consider the following models: Qwen2.5-7B, Qwen2.5-7B-Instruct (Yang et al., 2024a), Llama-3.1-8B and Llama-3.1-8B-Instruct (Dubey et al., 2024). Table 1 presents the

Dataset	✘✘	✘✔	✔✘	✔✔
Qwen2.5-7B				
NQ	978	1818	140	674
TriviaQA	2748	3493	1565	3507
WebQ	566	767	135	564
Qwen2.5-7B-Instruct				
NQ	960	1729	158	763
TriviaQA	2440	2960	1873	4040
WebQ	548	703	153	628
Llama-3.1-8B				
NQ	934	1654	184	838
TriviaQA	2187	2508	2126	4492
WebQ	533	660	168	671
Llama-3.1-8B-Instruct				
NQ	796	1318	322	1174
TriviaQA	1836	2069	2477	4931
WebQ	451	536	250	795

Table 1: Statistics of test sets across the four knowledge quadrants. ✔and ✘correspond to internal knowledge, while ✔and ✘correspond to external knowledge.

statistics of test data with 3 documents, categorized into four knowledge quadrants for each models.

For data across the four knowledge boundary quadrants, the model should demonstrate correspondingly distinct behaviors. Ideally, for Type ✘✘data, the model should abstain from answering. For Type ✘✔and Type ✔✔data, it should leverage the external documents to generate answers. For Type ✔✘data, where the external knowledge contains interference, the model should rely on its internal knowledge to answer. Achieving such selective generation requires a reliable perception of knowledge boundaries. The subsequent experiments are conducted based on these datasets.

3.2 Uncertainty Estimation Methods

The perception of knowledge boundaries is closely related to model uncertainty estimation. In this paper, we mainly consider latent information methods due to their efficiency and reliance on a model's internal information for scoring. These methods can be categorized by their information source: predicted probability-based, probability distribution-based and hidden states-based approaches (Xia et al., 2025). Additionally, we distinguish methods by their requirements during the LLM generation process: "Prior" methods can compute scores before an answer is fully generated, while "Posterior" methods require the complete generated output for computation (Chen et al., 2024). Table 2 summarizes the uncertainty estimation methods considered and their applicability. The brief descriptions for each method are provided in Appendix A.

Method	Prior	Posterior
Probability-Level		
Perplexity (Fomicheva et al., 2020)	✔	✔
Fst Prob (Allen-Zhu and Li, 2024)	✔	✔
Min Prob (Varshney et al., 2023)		✔
Prod Prob (Varshney et al., 2023)		✔
Logprob (Abbasi-Yadkori et al., 2024)		✔
P(True) (Kadavath et al., 2022)	✔	
Internal Confidence (Chen and Varoquaux, 2025)	✔	
Distribution-Level		
Predictive Entropy (Malinin and Gales, 2021)	✔	✔
Mink Entropy (Shi et al., 2024a)	✔	✔
Attentional Entropy (Duan et al., 2024)	✔	
LogTokU (Ma et al., 2025)		✔
Hidden states-Level		
Hidden Score (Sriramanan et al., 2024)		✔
Attn Score (Sriramanan et al., 2024)		✔
Estimator (Ni et al., 2025)	✔	✔

Table 2: The uncertainty estimation methods considered in this paper. ✔indicates applicability in this scenario.

As shown in Table 2, computational methods such as PPL and entropy are applied in both prior and posterior scenarios, producing results based on either individual tokens or averaged across full token sequences. While other computational methods could also be extended to both scenarios, our implementation follows the setup described in the original papers. Among the considered methods, only the Estimator (Ni et al., 2025) requires additional training. We follow the original approach and train it using 2000 examples from the training set, with further details provided in Section 4.1.

3.3 Impact of Document

Prior works have largely focused on tasks such as hallucination detection within models, without accounting for the influence of external documents on the estimation results. This paper specifically examines the effect of external knowledge by comparing performance with and without retrieved documents. Following established practice (Ni et al., 2025; Vashurin et al., 2024; Xia et al., 2025), the ground truth confidence is set to 1 if the LLM's generation contains the correct answer, and 0 otherwise. We evaluate various uncertainty methods using Accuracy, F1 score, and AUROC. Accuracy offers a direct measure of estimation effectiveness,

	Method	Qwen2.5-7B						Qwen2.5-7B-Instruct					
		0doc			3doc			0doc			3doc		
		Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc
Prior	Perplexity	0.572	0.378	0.570	0.559	0.653	0.557	0.577	0.421	0.587	0.537	0.660	0.523
	Fst Prob	0.516	0.355	0.490	0.547	0.652	0.542	0.513	0.396	0.501	0.527	0.657	0.533
	Predictive Entropy	0.565	0.377	0.570	0.552	0.653	0.553	0.600	0.431	0.604	0.534	0.660	0.527
	Mink Entropy	0.565	0.377	0.571	0.551	0.652	0.546	0.535	0.406	0.502	0.529	0.659	0.606
	Attentional Entropy	0.573	0.373	0.572	0.543	0.653	0.532	0.594	0.429	0.609	0.592	0.670	0.537
	P(True)	0.603	**0.390**	0.631	**0.688**	**0.704**	0.729	0.645	0.472	0.673	0.699	**0.732**	**0.751**
	Internal Confidence	0.574	0.368	0.587	0.543	0.655	0.537	0.546	0.408	0.548	0.533	0.659	0.533
	Estimator	**0.782**	0.128	**0.651**	0.667	0.651	**0.737**	**0.717**	**0.483**	**0.742**	**0.688**	0.680	0.376
Posterior	Perplexity	0.501	0.348	0.453	0.506	0.652	0.434	0.581	0.430	0.553	0.502	0.657	0.624
	Fst Prob	0.605	0.378	0.623	**0.663**	**0.671**	0.687	0.505	0.396	0.498	0.505	0.658	0.623
	Min Prob	0.574	0.363	0.586	0.573	0.656	0.568	0.512	0.399	0.426	0.625	0.664	0.646
	Prod Prob	0.573	0.363	0.584	0.572	0.654	0.570	0.514	0.397	0.430	**0.626**	**0.672**	**0.655**
	Logprob	0.544	0.358	0.547	0.556	0.654	0.566	0.516	0.401	0.447	0.608	0.658	**0.655**
	Predictive Entropy	0.596	0.373	0.514	0.585	**0.671**	0.582	0.537	0.395	0.507	0.554	0.657	0.502
	Mink Entropy	0.615	0.399	0.554	0.593	0.654	0.578	0.546	0.403	0.537	0.625	0.663	0.503
	LogTokU	0.503	0.353	0.254	0.516	0.656	0.474	0.606	0.406	0.626	0.616	0.669	0.644
	Hidden Score	0.559	0.348	0.435	0.517	0.654	0.564	0.629	**0.454**	0.617	0.557	0.657	0.504
	Attn Score	0.703	**0.494**	**0.695**	0.561	0.654	**0.702**	0.613	0.436	0.612	0.599	0.657	0.472
	Estimator	**0.788**	0.000	0.635	0.661	0.628	0.565	**0.763**	0.213	**0.680**	0.535	0.652	0.581

Table 3: The impact of document on different uncertainty estimation methods in NQ dataset. If metrics improve after introducing document, cells are colored green; if metrics decline, they are colored red.

the F1 score balances precision and recall while guarding against extreme predictions, and AUROC assesses overall discriminative ability.

Based on the aforementioned datasets, we sample 1000 instances from each dataset and conduct our experiments on this subset. Results for the Qwen-series models on the NQ dataset are shown in Table 3. Results for other models and datasets are provided in Appendix B.1. In the table, cell color indicates the change in each metric after external documents are introduced. The results show that: (1) For most uncertainty estimation methods, the presence of external documents interferes with predictions, leading to a decline in both accuracy and AUROC. In contrast, the F1 score increases. This occurs because, in computing the F1 score, we treat the confident class (labeled 1) as positive. Thus, a drop in accuracy alongside a rise in F1 suggests that the model is reclassifying more originally unconfident examples as confident. The decrease in AUROC further reflects the model's reduced ability to discriminate between confident and unconfident samples, making it more likely to misclassify uncertain cases as confident. (2) Across methods, the Estimator demonstrates superior overall performance in both prior and posterior scenarios. This advantage is particularly evident in accuracy metrics, indicating that probing the internal states is more effective than probability- or distribution-based approaches. This finding aligns with conclusions from prior works (Ni et al., 2025; Ji et al., 2024a; Ashok and May, 2025b; Zeng et al., 2025; Zhou et al., 2025). However, we note that the Estimator yields a low F1 score when no documents are present, sometimes even reaching zero, which corresponds to all predictions being classified as zero. This occurs because the method is sensitive to the distributional shift between training and test data. We address this issue in the next section according to the specific requirements of our application. (3) In comparing prior and posterior scenarios, posterior methods generally demonstrate higher reliability. This aligns with the intuition that posterior methods, which assess complete generated responses, provide more comprehensive information. Among prior methods, P(True) remains a robust baseline that is largely unaffected by document interference. Between the two model types, the Instruct variant shows slightly higher F1 and AUROC scores without documents, and the impact of introducing documents is relatively minor. This suggests that fine-tuning may enhance the model's awareness of its knowledge boundaries, improving both its ability to distinguish known from unknown information and its overall response quality.

Overall, the introduction of external documents blurs the model's perception of the knowledge boundaries and ability to distinguish between known and unknown knowledge, making it more inclined to treat the unknown as known.

4 Method

In this section, we address the research question: *How can we enhance the perception and utilization of knowledge boundaries in RAG?*

4.1 Notations

We employ the Estimator method to assess the model's perception of its knowledge boundaries for the following reasons: (1) As shown in Table 3, the Estimator yields more reliable results. Probability- and distribution-based methods are highly sensitive to token distributions, which are significantly disrupted by external documents. In contrast, internal knowledge remains encoded in the model's internal states, making probing-based estimation more stable. (2) The method is applicable in both prior and posterior scenarios, offering broader utility. (3) Although additional training is required, the data volume is minimal, and different training strategies can be used to adjust the focus of the estimator.

Formally, following Ni et al. (2025), $\mathcal{E}$ is the confidence estimator that produce the estimation of model's confidence based on its hidden states. Specifically, for a given model M, a question q and corresponding documents D, the model generates a response along with internal states $I_{M,q,D}$. The confidence score $c_{M,q,D}$ is computed as follows:

$$c_{M,q,D} = \mathcal{E}(I_{M,q,D}) \tag{1}$$

The $\mathcal{E}$ is structured as a lightweight MLP (Multi-layer Perceptron) and perform binary classification to produce a confidence score. It is trained on an additional training dataset $D_{train} = \{(I^{train}_{M,q_i,D_i}, c^{train}_{M,q_i,D_i})_{i=1}^{N}\}$, where N is the number of training samples. The ground-truth confidence c^{train}_{M,q_i,D_i} is set to 1 if the model can answer the question q_i using its internal knowledge and the external documents D_i ; otherwise, it is set to 0. The training objective is the cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} c_i \log(P_i) + (1-c_i)\log(1-P_i) \tag{2}$$

$$P_i = P(\hat{c}_i = 1) = \sigma\left(\text{MLP}\left(I\right)\right) \tag{3}$$

where c_i is the ground truth confidence and $\hat{c}_i$ is the estimated confidence. σ is the sigmoid function and I is the hidden states vector transformed from I_{M,q_i,D_i}. Once $\mathcal{E}$ is learned, confidence estimation during inference is performed as:

$$\hat{c}_i = \arg\max_{y \in \{0,1\}} P(c_i = y). \tag{4}$$

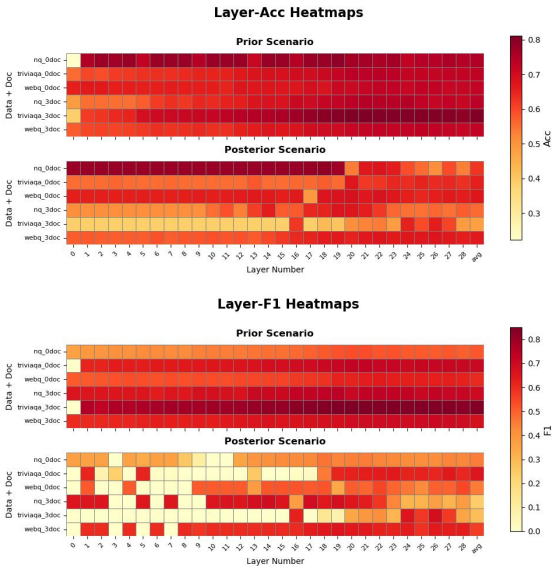

Figure 1: Performance of the Estimator when probing different layers for Qwen2.5-7B model. The metric in the upper figure is accuracy (acc), while the metric in the lower figure is F1 score.

We trained the estimator using only 2000 training samples, consistent with the Ni et al. (2025). Implementation details, including hyperparameters such as the number of layers, hidden units, learning rate, and optimizer, are also identical.

4.2 Improvement on Estimator

Several factors can affect estimator performance. We first examine how the choice of internal hidden state layers used for probing influences results.While Azaria and Mitchell (2023) and Ni et al. (2025) suggest that intermediate layer representations best reflect a model's factuality awareness, this finding may be task- and model-specific and may not generalize to all models or to RAG tasks. We therefore conduct experiments probing different layers across datasets in both prior and posterior settings. Results for the Qwen2.5-7B model are shown as heatmaps in Figure 1. Results for other models are provided in Appendix B.2.

The results indicate that: (1) Regarding layer selection, the intermediate layer does not show outstanding performance. Higher layers generally reflect the model's confidence better than lower layers, though exceptions exist. In contrast, probing based on the average across all layers (last column) yields more stable and robust performance. (2) In the posterior scenario on nq_0doc, we observe that probing the first 20 layers yields very high accuracy, yet the corresponding F1 scores remain low. This

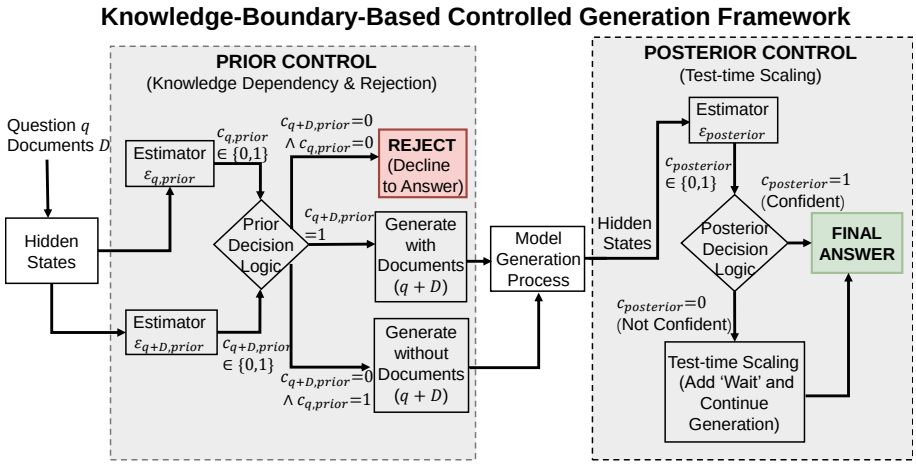

Figure 2: The overall framework.

suggests that the trained Estimator tends to produce extreme predictions, consistently outputting zero. This issue can be explained from two perspectives. First and most notably, the test data exhibits label imbalance, which encourages an accuracy-oriented Estimator to adopt simplistic predictive shortcuts. Second, it indicates that lower-layer hidden states do not encode sufficiently discriminative representations to support reliable confidence estimation. This effect is more pronounced in posterior settings, likely because task-relevant information becomes progressively embedded in higher layers during generation, making lower-layer activations less informative for factual confidence assessment. Probing based on the average across all layers helps mitigate this issue to some extent.

The estimator's tendency to produce all-zero predictions stems from its training process. Because the test data exhibits label imbalance (see Table 1), optimizing for accuracy can encourage this shortcut behavior. While machine-learning research offers solutions that address either training data or training objectives, we focus on refining the training objective. This approach avoids custom adjustments to the data distribution, which is especially suitable given the diversity of our test data and the variations across different models.

Since our downstream objective is for the model to answer selectively based on its knowledge boundaries, retaining confident responses while rejecting uncertain ones. For questions of moderate confidence, generating informative responses is preferable to outright refusal. Therefore, we place greater emphasis on training data labeled as 1, which also represent the minority class in the training distribution. We therefore employ a weighted cross-entropy loss during training:

$$\mathcal{L}_{\text{WCE}} = -\sum_{i=1}^{N} w_1 * c_i \log(P_i) + w_0 * (1-c_i) \log(1-P_i) \quad (5)$$

where w_1 is the weight for data labeled 1, and w_0 is the weight for data labeled 0. In the experiment, we set $w_1 = 2$ and $w_0 = 1$. Additionally, we shift the evaluation focus from accuracy-oriented to F1-oriented, which better aligns with our objectives.

4.3 Framework

Building on the refined Estimator, we construct a framework that enables the model to dynamically determine its reliance on knowledge according to perceived boundaries and to reject questions where confidence is low, as outlined in Section 3.1. The overall framework is illustrated in Figure 2. The framework adjusts the model generation process, functioning in both prior and posterior scenarios.

For prior scenarios, the estimator trained on prior settings can assess the model's internal confidence before the model performs generation. This prior awareness allows the model to operate within its perceived knowledge boundaries. In certain cases, relying on internal knowledge may yield better results due to possible interference from external knowledge (as Type ✔✘in Table 1). We therefore design two estimators for fine-grained prior awareness: $\mathcal{E}_{q+D,prior}$, trained using labels from inputs with external documents, which reflects the model's awareness of knowledge boundaries under the combined influence of external documents and internal knowledge; and $\mathcal{E}_{q,prior}$, trained using labels from inputs without documents, capturing

Method	Qwen2.5-7B			Qwen2.5-7B-Instruct			Llama-3.1-8B			Llama-3.1-8B-Instruct		
	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1
Vanilla-G	0.552	0.592	0.361	0.668	0.716	0.552	0.557	0.632	0.005	0.670	0.736	0.345
Vanilla-R	0.267	0.000	0.418	0.267	0.000	0.418	0.267	0.000	0.418	0.267	0.000	0.418
Verb-Prior	0.587	0.592	0.575	0.540	0.558	0.515	0.439	0.500	0.305	0.577	0.642	0.426
Verb-Post	0.570	0.581	0.539	0.367	0.295	0.426	0.485	0.554	0.294	0.591	0.647	0.447
COT	0.578	0.609	0.451	0.620	0.633	**0.590**	0.548	0.620	0.013	0.679	0.744	0.391
CAD	0.473	0.533	0.109	0.668	0.724	0.533	0.557	0.631	0.003	0.649	0.720	0.270
CKPLUG	0.528	0.579	0.264	0.669	**0.742**	0.532	0.554	0.628	0.004	0.677	0.752	0.339
LogtokU	0.544	0.581	0.404	0.569	0.570	0.559	0.508	0.550	0.375	0.461	0.493	0.383
P(True)	0.573	0.567	**0.583**	0.593	0.618	0.548	0.492	0.529	0.413	0.596	0.651	0.479
Internal Confidence	0.526	0.523	0.528	0.598	0.632	0.539	0.451	0.483	0.387	0.536	0.595	0.416
Our	**0.614**	**0.628**	0.567	**0.670**	0.723	0.549	**0.621**	**0.650**	**0.507**	**0.714**	**0.777**	**0.508**

Table 4: The average results across NQ, TriviaQA and WebQ datasets. Acc, Ans-F1, Abs-F1 denote Accuracy, Answer-F1, Abstain-F1 respectively .

only the model's awareness under internal knowledge. The outputs $c_{q+D,prior}$ and $c_{q,prior}$ jointly determine the model's behavior. If both are 0, the model lacks confidence and declines to answer. If $c_{q+D,prior}$ is 1, the model is confident in generating output based on documents and will utilize them for generation. If $c_{q+D,prior}$ is 0 and $c_{q,prior}$ is 1, internal knowledge is considered more reliable, and generation proceeds without external documents.

This selective generation, grounded in fine-grained prior awareness, enables the model to manage knowledge dependencies and reject questions where it lacks confidence. For Types ✘✘and ✔✘, prior control can be effective and reduce inference costs. However, for Types ✘✔and ✔✔,it does not enhance the ability to utilize documents. We therefore employ posterior control to address this limitation. Posterior awareness reflects the model's confidence in generated answers. We design a posterior Estimator $\mathcal{E}_{posterior}$: when its output $c_{posterior}$ is 1, the model is confident in the generated result and returns it directly. When $c_{posterior}$ is 0, confidence is low, suggesting the model may have missed crucial document information. In this case, we leverage the model's reasoning capability to continue generation by appending the string "Wait" to the output. This encourages the model to reflect on its reasoning trace and proceed further, implementing a form of test-time scaling inspired by Muennighoff et al. (2025). While this process could be repeated, we perform it only once in subsequent experiments for cost efficiency.

Overall, our framework employs knowledge-boundary-based controlled generation, where prior awareness determines rejection and knowledge dependencies, while posterior awareness decides whether to utilize test-time scaling to further capture document information.

5 Experiments

5.1 Baselines and Metrics

In this section, we compare the proposed framework with the following baselines: (1) **Vanilla-G**: Direct generation of the answer. (2) **Vanilla-R**: Direct refusal to answer. (3) Prompt-based methods: These methods elicit the model's verbal confidence via prompting to decide whether to refuse or generate a response, categorized into prior (**Verb-Prior**) and posterior (**Verb-Post**) variants (Ren et al., 2023). Moreover, **COT** (Kojima et al., 2022) is also included as a prompt-based reasoning enhancement approach. (4) Decoding-based methods: These methods adjust the model's reliance on internal versus external knowledge during decoding using contrastive strategies, such as **CAD** (Shi et al., 2024b) and **CKPLUG** (Bi et al., 2025b). (5) Confidence-based methods: **LogTokU** (Ma et al., 2025) proposes a combined aleatoric and epistemic uncertainty metric to determine refusal. Additionally, **P(True)** (Kadavath et al., 2022) and **Internal Confidence** (Chen and Varoquaux, 2025) have proven to be robust baselines in previous experiments. Prompt templates used for each baseline are provided in Appendix C.

Sun et al. (2025) introduce a comprehensive evaluation framework based on the knowledge quadrant. Our analysis focuses on the balance between responses and refusals, as well as the quality of the final output. Therefore, we employ three metrics: Accuracy, Answer-F1 and Abstain-F1. Accuracy measures the proportion of correct answers plus appropriate abstentions relative to the total number of queries, reflecting overall performance. Answer-F1 and Abstain-F1 represent the harmonic mean

Dataset	Method	✗✗	✗✓	✓✗	✓✓
NQ	Vanilla-G	0.001	0.603	0.630	0.894
	Our	0.039	0.540	0.788	0.908
TriviaQA	Vanilla-G	0.003	0.580	0.831	0.955
	Our	0.457	0.524	0.849	0.944
WebQ	Vanilla-G	0.004	0.433	0.720	0.869
	Our	0.319	0.356	0.714	0.870

Table 5: The accuracy for data within each quadrant of Llama-3.1-8B model.

of precision and recall for responses and refusals, respectively. Further details regarding metric calculations are provided in Appendix D.

5.2 Main Results

The results are shown in Table 4. Due to space constraints, the table report average results across the three datasets. Detailed results per dataset are provided in Appendix B.3. The results indicate that our approach consistently improves generation quality across all models. The concurrent increase in both Answer-F1 and Abstain-F1 suggests that our method does not sacrifice answer generation capability in order to learn rejection, but rather enhances both capabilities simultaneously.

Beyond overall performance improvements, we assess performance gains according to the four-quadrant classification of knowledge boundaries. Using Llama-3.1-8B as an example, we compute accuracy within each quadrant. Specifically, we calculate the refusal rate for Type ✗✗data and the correct response rate for the remaining types. The results are presented in Table 5. The results further demonstrate that our approach enhances rejection rates while preserving the original capability for correct responses. Moreover, the performance gains observed on ✓✗data confirm the effectiveness of our knowledge dependency control. We believe that further enhancing the accuracy of perceiving the boundaries of knowledge will yield additional gains.

5.3 Ablation Experiments

In this section, we perform ablation studies on the proposed framework, which consists of two components: prior control and posterior control. Using Llama-3.1-8B on TriviaQA dataset as a representative case, we evaluate the effects of removing each component, with results shown in Table 6.

Method	Estimator	Acc	Ans-F1	Abs-F1
Vanilla-G		0.664	0.736	0.005
Prior+ Posterior	avg-based + weighted	0.739	0.775	0.567
Only Prior	mid-based	0.709	0.763	0.557
	avg-based + weighted	0.717	0.768	0.573
Only Posterior	avg-based + weighted	0.684	0.741	0.021

Table 6: Ablation studies of prior and posterior controls and training strategies for estimator.

The results show that the complete framework, integrating both prior and posterior control, achieved the best performance. Removing either component led to a decline. Prior control contributed more significantly to generation quality than posterior control. We further conduct an ablation study comparing different training strategies for estimator in prior-only setup: our proposed method versus the previous accuracy-oriented intermediate-layer probing approach. This ablation confirms that the strategy introduced in Section 4.2 outperforms original implementation.

The main experiment is conducted in a three-document setting, though our method can be extended to other numbers of documents. Results for experiments with more (10doc) and fewer (0doc) documents are presented in the Appendix E.1. Notably, in the zero-document setting, prior control relies on a single estimator. To verify the scalability of our framework, we also conduct experiments on Qwen3-14B (Yang et al., 2025) model, with the results presented in the Appendix E.2.

6 Conclusion

In this paper, we focus on the model's internal perception of knowledge boundaries. We first investigate how the presence of external documents influences this boundary perception. Experiments reveal that introducing external knowledge disrupts the model's ability to distinguish between known and unknown information, causing it to treat the unknown as known. To enhance the model's discrimination capability, we optimize the training strategy for the estimator. Building upon this, we propose a framework incorporating both prior and posterior controls, enabling the model to learn to dynamically adjust its knowledge dependencies and reject unknown queries. Experiments demonstrate that our framework significantly enhances generation quality, enabling the model to generate when appropriate and reject when necessary.

Limitations

There are several limitations of our current work that we plan to address in the future:

(1) The core of this paper lies in the model's perception of its knowledge boundaries. Due to resource constraints, we have only considered uncertainty estimation methods based on latent information. We have not explored consistency-based and semantic clustering-based approaches that may be potentially more accurate. The impact of external knowledge on these methods remains for subsequent research.

(2) We employ an estimator based on hidden states as our foundation, which necessitates the downstream model being a white-box model. This approach is not applicable to black-box models where intermediate states cannot be accessed.

(3) The estimator requires additional training data for training. The training labels employed in this paper are binary labels indicating whether a question can be answered. The labeling scheme could be refined further, for instance by utilizing semantic entropy as labels or constructing them based on question difficulty. Moreover, the framework design could be made more sophisticated, such as incorporating dynamic retrieval or implementing multi-round iterations within posterior control. Our approach still holds potential for improvement.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your LLM: iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6416–6432. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511.

Dhananjay Ashok and Jonathan May. 2025a. Language models can predict their own behavior. *ArXiv*, abs/2502.13329.

Dhananjay Ashok and Jonathan May. 2025b. Language models can predict their own behavior. *CoRR*, abs/2502.13329.

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 967–976. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. 2025a. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. 2025b. Parameters vs. context: Fine-grained control of knowledge reliance in language models. *ArXiv*, abs/2503.15888.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024. Teaching large language models to express knowledge boundary from their own signals. *CoRR*, abs/2406.10881.

Lihu Chen and Gaël Varoquaux. 2025. Query-level uncertainty in large language models. *ArXiv*, abs/2506.09669.

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, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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

2024, Bangkok, Thailand, August 11-16, 2024, pages 5050–5063. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:539–555.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. A survey of uncertainty estimation in llms: Theory meets practice. *ArXiv*, abs/2410.15326.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. 2025. PIP-KAG: mitigating knowledge conflicts in knowledge-augmented generation via parametric pruning. *CoRR*, abs/2502.15543.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024a. LLM internal states reveal hallucination risk faced with a query. *CoRR*, abs/2407.03282.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024b. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, and Zhijiang Guo. 2024. DVD: dynamic contrastive decoding for knowledge amplification in multi-document question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4624–4637. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Katie Kang, Eric Wallace, Claire J. Tomlin, Aviral Kumar, and Sergey Levine. 2025. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3600–3612. Association for Computational Linguistics.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *CoRR*, abs/2506.09038.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. 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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5131–5157. Association for Computational Linguistics.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, Jie Zhou, Yujiu Yang, Ngai Wong, Xixin Wu, and Wai Lam. 2024. A survey on the honesty of large language models. *Trans. Mach. Learn. Res.*, 2025.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 19260–19268. AAAI Press.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. 2025. Estimating llm uncertainty with evidence.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *CoRR*, abs/2501.19393.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24315–24329. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching LLMs credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8452–8472. Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *International Conference on Computational Linguistics*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Huai hsin Chi, Nathanael Scharli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024a. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems 37*.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11461–11480. Association for Computational Linguistics.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.

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. *CoRR*, abs/2307.03987.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Lyudmila Rvanova, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Trans. Assoc. Comput. Linguistics*, 13:220–248.

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. *Trans. Assoc. Comput. Linguistics*, 13:529–556.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling*.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8541–8565. Association for Computational Linguistics.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6002–6024. 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 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Yue Xing, Monica Xiao Cheng, and Jiliang Tang. 2024. Towards knowledge checking in retrieval-augmented generation: A representation perspective. In *North American Chapter of the Association for Computational Linguistics*.

Shenglai Zeng, Jiankun Zhang, Bingheng Li, Yuping Lin, Tianqi Zheng, Dante Everaert, Hanqing Lu, Hui Liu, Yue Xing, Monica Xiao Cheng, and Jiliang Tang. 2025. Towards knowledge checking in retrieval-augmented generation: A representation perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2952–2969. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7113–7139. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 8908–8949. Association for Computational Linguistics.

Tianyi Zhou, Johanne Medina, and Sanjay Chawla. 2025. Can llms detect their confabulations? estimating reliability in uncertainty-aware language models. *CoRR*, abs/2508.08139.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14544–14556. Association for Computational Linguistics.

A Uncertainty Estimation Methods

In this section, we briefly outline the various uncertainty estimation methods under consideration.

- Probability-level: According to Xia et al. (2025), some studies directly employ the predicted probability of response tokens as a confidence score. Other work proposes using the negative log likelihood of response tokens, whether averaged or maximized across tokens, as an uncertainty measure (Vashurin et al., 2024). The average negative log likelihood across tokens is equivalent to perplexity (Fomicheva et al., 2020). In contrast, Kadavath et al. (2022) introduce a method called P(True), which prompts the model to evaluate its answers by responding true or false and uses the latent probability associated with "True" as the confidence score. Additionally, Chen and Varoquaux (2025) propose a training free approach termed Internal Confidence that leverages self evaluations across layers and tokens to provide a reliable uncertainty signal.

- Distribution-level: Malinin and Gales (2021); Shi et al. (2024a) compute the entropy of the probability distribution for each generated token and use either the mean or maximum entropy as the uncertainty measure. Extending this line of work, Duan et al. (2024) incorporate attention mechanisms into entropy calculation and propose Attentional Entropy. In a different direction, Ma et al. (2025) introduce Logits induced token uncertainty (LogTokU), a framework for estimating decoupled token uncertainty in LLMs that enables real time uncertainty estimation without requiring multiple sampling processes.

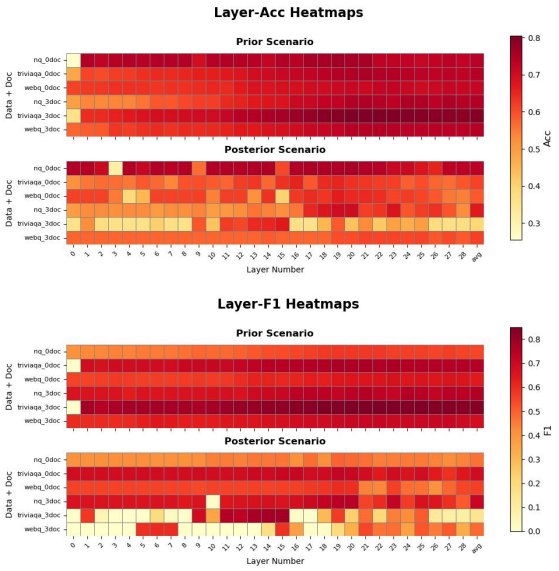

Figure 3: Performance of the Estimator when probing different layers for Qwen2.5-7B-Instruct model. The metric in the upper figure is accuracy (acc), while the metric in the lower figure is F1 score.

- Hidden states-level: Sriramanan et al. (2024) propose a method to identify hallucinations within individual responses in both white-box and black-box settings by analyzing internal hidden states, attention maps, and output prediction probabilities to compute LLM-Check scores. Probing intermediate states is also a widely adopted practice (Ni et al., 2025; Ji et al., 2024b; Ashok and May, 2025a; Zeng et al., 2024; Zhou et al., 2025).

B Full Results

Due to space constraints, we are unable to present the results of all models across all datasets within the main text. These are presented in full in this section.

B.1 Uncertainty Estimation

The complete results of uncertainty estimation methods are shown in Table 7 and 8.

B.2 Probing Layers

The results of Qwen2.5-7B-Instruct, Llama-3.1-8B, Llama-3.1-8B-Instruct model are shown in Figure 3, 4 and 5 correspondingly.

B.3 Main Results of Each Dataset

The individual results for each dataset are shown in Table 9.

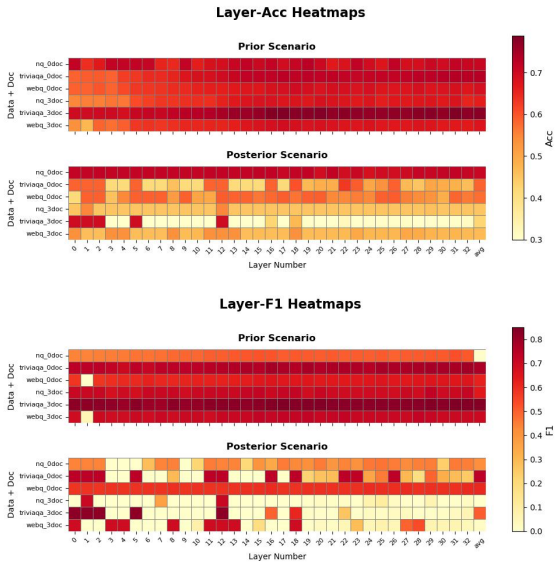

Figure 4: Performance of the Estimator when probing different layers for Llama-3.1-8B model. The metric in the upper figure is accuracy (acc), while the metric in the lower figure is F1 score.

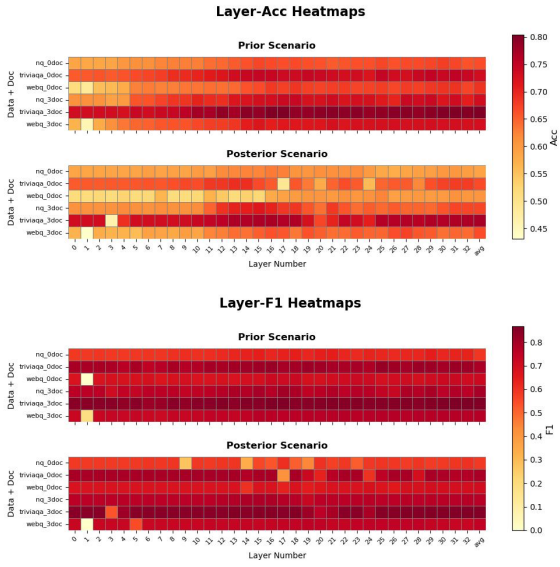

Figure 5: Performance of the Estimator when probing different layers for Llama-3.1-8B-Instruct model. The metric in the upper figure is accuracy (acc), while the metric in the lower figure is F1 score.

C Templates

The prompt templates used for each baseline in our experiments are as shown in Table 10. Most approaches, including our framework, employ the default template.

D Metrics

The details on the calculation of the metrics are shown in Table 11.

E Ablation Experiments

E.1 Number of Documents

Experiments under different numbers of documents are shown in Table 12. For undocumented scenarios, decoding methods such as CAD and CKPLUG are not applicable. Furthermore, within our framework, undocumented cases are handled using a single Estimator for prior control.

E.2 Model Scale

We present the results for Qwen3-14B in Table 13. Due to time and equipment constraints, we selected only one thousand samples from each dataset for experimentation. The discrepancy with Table 9 stems not only from the model but also from the sample selection: previous models were tested using the full test dataset, whereas Qwen3-14B was evaluated using only one thousand samples. Moreover, we achieved superiority over other baselines using only prior control, without employing posterior control.

	Method	Qwen2.5-7B						Qwen2.5-7B-Instruct					
		0doc			3doc			0doc			3doc		
		Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc
						TriviaQA							
prior	Perplexity	0.506	0.586	0.473	0.567	0.721	0.547	0.535	0.663	0.494	0.573	0.751	0.513
	Fst Prob	0.542	0.586	0.534	0.549	0.721	0.564	0.533	0.661	0.514	0.522	0.751	0.578
	Predictive Entropy	0.522	0.587	0.493	0.568	0.721	0.552	0.577	0.668	0.591	0.576	0.751	0.514
	Mink Entropy	0.515	0.585	0.493	0.547	0.722	0.570	0.531	0.664	0.530	0.530	0.751	0.651
	Attentional Entropy	0.507	0.586	0.480	0.565	0.722	0.572	0.535	0.664	0.515	0.645	0.752	0.582
	P(True)	0.602	0.588	0.630	0.743	0.796	**0.790**	0.645	0.681	0.690	0.748	0.806	**0.804**
	Internal Confidence	0.592	0.591	0.621	0.746	0.798	0.784	0.617	0.681	0.647	0.737	0.790	0.791
	Estimator	**0.707**	**0.647**	**0.761**	**0.762**	**0.804**	0.432	**0.704**	**0.702**	**0.770**	**0.766**	**0.815**	0.326
posterior	Perplexity	0.513	0.586	0.468	0.511	0.721	0.568	0.566	0.661	0.550	0.501	0.751	0.674
	Fst Prob	0.650	0.602	0.685	0.713	0.751	0.566	0.524	0.667	0.524	0.511	0.752	0.677
	Min Prob	0.585	0.592	0.596	0.592	0.722	0.604	0.515	0.662	0.463	0.688	0.755	0.713
	Prod Prob	0.587	0.590	0.595	0.595	0.722	0.605	0.528	0.661	0.465	0.691	0.779	0.734
	Logprob	0.538	0.587	0.532	0.575	0.723	0.605	0.500	0.661	0.450	0.664	0.752	0.734
	Predictive Entropy	0.680	0.617	0.614	0.648	0.736	**0.749**	0.502	0.662	0.402	0.579	0.751	0.510
	Mink Entropy	0.702	0.654	0.662	0.677	0.743	0.500	0.572	**0.677**	0.529	0.693	0.782	0.505
	LogTokU	0.501	0.590	0.218	0.503	0.726	0.660	**0.679**	0.694	0.709	**0.696**	**0.792**	**0.753**
	Hidden Score	0.501	0.585	0.246	0.503	0.722	0.648	0.568	0.661	0.537	0.530	0.751	0.524
	Attn Score	**0.749**	**0.712**	**0.735**	0.652	**0.778**	0.299	0.646	0.675	0.644	0.618	0.753	0.334
	Estimator	0.587	0.000	0.597	0.572	0.728	0.595	0.584	**0.695**	**0.747**	0.644	0.766	0.600
						WebQ							
prior	Perplexity	0.531	0.534	0.526	0.535	0.609	0.511	0.576	0.566	0.566	0.552	0.610	0.509
	Fst Prob	0.518	0.529	0.518	0.521	0.609	0.525	0.521	0.565	0.516	0.520	0.609	0.536
	Predictive Entropy	0.535	0.534	0.531	0.529	0.609	0.512	0.585	0.566	0.592	0.549	0.609	0.506
	Mink Entropy	0.534	0.534	0.530	0.521	0.609	0.552	0.518	0.565	0.513	0.515	0.609	0.588
	Attentional Entropy	0.533	**0.535**	0.517	0.553	0.612	0.532	0.592	0.567	0.599	0.571	0.622	0.544
	P(True)	0.593	0.545	0.612	0.626	0.654	0.660	0.630	0.600	0.667	0.635	0.652	**0.675**
	Internal Confidence	0.584	0.544	0.599	0.631	**0.660**	**0.663**	0.626	**0.606**	0.652	0.642	**0.656**	0.673
	Estimator	**0.661**	0.426	**0.669**	**0.665**	0.640	0.451	**0.689**	0.571	**0.735**	**0.678**	0.619	0.429
posterior	Perplexity	0.552	0.531	0.531	0.512	0.609	0.549	0.577	0.566	0.548	0.507	0.609	0.571
	Fst Prob	0.587	0.528	0.583	**0.602**	**0.615**	0.549	0.504	0.564	0.501	0.517	0.613	0.569
	Min Prob	0.519	0.527	0.493	0.551	0.611	0.547	0.516	0.565	0.444	0.574	0.615	0.582
	Prod Prob	0.518	0.528	0.490	0.551	0.614	0.547	0.512	0.563	0.444	0.576	**0.622**	**0.589**
	Logprob	0.509	0.528	0.469	0.550	0.609	0.547	0.513	0.564	0.452	0.572	0.610	**0.589**
	Predictive Entropy	0.573	0.527	0.531	0.582	0.614	**0.617**	0.507	0.564	0.423	0.533	0.610	0.516
	Mink Entropy	0.585	0.527	0.533	0.580	0.612	0.499	0.509	0.563	0.469	0.582	0.609	0.488
	LogTokU	0.502	0.521	0.309	0.526	0.611	0.573	0.592	0.535	0.572	0.565	0.615	0.588
	Hidden Score	0.553	0.528	0.455	0.527	0.611	0.577	0.590	**0.573**	0.594	0.541	0.611	0.516
	Attn Score	0.624	**0.549**	**0.635**	0.549	0.610	0.491	0.582	0.563	0.583	0.568	0.611	0.486
	Estimator	**0.646**	0.161	0.611	0.569	0.356	0.537	**0.644**	0.364	**0.637**	**0.586**	0.324	0.554

Table 7: The impact of document on different uncertainty estimation methods in TriviaQA and WebQ dataset for Qwen-series models.

	Method	Llama-3.1-8B						Llama-3.1-8B-Instruct					
		0doc			3doc			0doc			3doc		
		Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc	Acc	F1	Auroc
							NQ						
prior	Perplexity	0.608	0.470	0.630	0.590	0.706	0.513	0.571	**0.613**	0.600	0.618	0.759	0.528
	Fst Prob	0.526	0.438	0.526	0.529	0.698	**0.612**	0.576	0.599	0.558	0.540	0.748	0.655
	Predictive Entropy	0.613	**0.478**	**0.635**	0.590	0.704	0.521	0.566	0.602	0.577	0.621	0.759	0.555
	Mink Entropy	0.561	0.444	0.562	0.529	0.698	0.604	0.563	0.603	0.555	0.549	0.749	0.654
	Attentional Entropy	0.609	0.455	0.619	0.583	**0.710**	0.607	0.568	0.608	0.582	0.620	0.764	0.650
	P(True)	0.554	0.442	0.550	0.544	0.700	0.551	0.590	0.609	0.599	0.678	0.773	**0.727**
	Internal Confidence	0.567	0.454	0.577	0.520	0.699	0.507	0.546	0.598	0.549	0.546	0.748	0.524
	Estimator	**0.725**	0.000	0.473	**0.675**	0.700	0.451	**0.653**	0.588	**0.703**	**0.746**	**0.798**	0.455
posterior	Perplexity	0.544	0.431	0.509	0.509	0.698	0.549	0.509	0.597	0.474	0.506	0.748	0.545
	Fst Prob	0.569	0.453	0.568	0.546	0.699	0.549	0.535	0.597	0.530	0.545	0.749	0.542
	Min Prob	0.537	0.432	0.498	0.552	0.701	0.545	0.504	0.598	0.481	0.551	0.749	0.552
	Prod Prob	0.542	0.433	0.494	0.558	0.702	0.553	0.518	0.598	0.497	0.542	0.748	0.545
	Logprob	0.542	0.432	0.491	0.553	0.702	0.553	0.544	0.598	0.526	0.542	0.748	0.545
	Predictive Entropy	0.582	0.461	0.555	0.546	0.716	0.543	0.524	0.598	0.508	0.526	0.749	0.544
	Mink Entropy	**0.633**	**0.497**	0.637	**0.568**	**0.718**	0.543	0.532	0.599	0.532	0.566	0.749	0.544
	LogTokU	0.518	0.429	0.446	0.561	0.681	**0.588**	0.527	0.544	0.498	0.557	0.751	0.561
	Hidden Score	0.511	0.433	0.419	0.502	0.716	0.549	0.576	0.621	0.563	0.557	0.749	0.515
	Attn Score	0.559	0.436	0.541	0.528	0.716	0.472	0.596	**0.625**	0.589	0.582	0.748	0.539
	Estimator	**0.725**	0.000	0.473	0.464	0.000	0.518	**0.612**	0.281	**0.653**	**0.699**	**0.783**	**0.593**
							TriviaQA						
prior	Perplexity	0.586	0.720	0.598	0.617	0.800	0.507	0.594	**0.814**	0.613	0.611	0.826	0.490
	Fst Prob	0.523	0.713	0.520	0.528	0.793	0.636	0.526	0.807	0.514	0.508	0.823	0.630
	Predictive Entropy	0.588	0.720	0.616	0.608	0.799	0.529	0.612	0.808	0.635	0.606	0.824	0.508
	Mink Entropy	0.557	0.713	0.553	0.532	0.793	**0.668**	0.574	0.807	0.571	0.516	0.823	0.674
	Attentional Entropy	0.591	0.722	0.599	0.633	0.807	0.642	0.586	0.812	0.610	0.643	0.834	0.629
	P(True)	0.551	0.715	0.542	0.565	0.794	0.575	0.568	0.809	0.573	0.702	0.829	**0.750**
	Internal Confidence	0.558	0.716	0.568	0.583	0.793	0.585	0.567	0.808	0.568	0.695	0.828	0.727
	Estimator	**0.734**	**0.761**	**0.786**	**0.774**	**0.825**	0.501	**0.730**	0.795	**0.759**	**0.797**	**0.848**	0.435
posterior	Perplexity	0.500	0.713	0.454	0.546	0.793	0.499	0.514	0.807	0.491	0.502	0.823	0.565
	Fst Prob	0.583	0.722	0.585	0.557	0.794	0.498	0.543	0.807	0.540	0.547	0.823	0.571
	Min Prob	0.539	0.714	0.541	0.507	0.793	0.459	0.517	0.807	0.510	0.549	0.823	0.544
	Prod Prob	0.539	0.714	0.534	0.509	0.794	0.458	0.517	0.807	0.501	0.557	0.823	0.565
	Logprob	0.542	0.718	0.546	0.539	0.799	0.458	0.519	0.807	0.509	0.557	0.824	0.565
	Predictive Entropy	0.549	0.727	0.509	0.524	0.811	0.557	0.527	0.807	0.506	0.592	0.823	0.546
	Mink Entropy	**0.633**	**0.738**	**0.632**	0.601	**0.813**	0.523	0.537	0.807	0.537	0.525	0.823	0.599
	LogTokU	0.534	0.718	0.430	0.542	0.780	**0.620**	0.601	0.801	0.569	0.609	0.823	0.485
	Hidden Score	0.512	0.720	0.360	0.515	0.809	0.522	0.526	**0.809**	0.508	0.582	0.824	**0.599**
	Attn Score	0.545	0.718	0.504	0.540	0.809	0.483	0.540	**0.809**	0.524	0.588	0.823	0.575
	Estimator	0.559	0.717	0.536	**0.666**	0.800	0.536	**0.646**	0.747	**0.639**	**0.756**	**0.841**	0.596
							WebQ						
prior	Perplexity	0.535	0.590	0.520	0.583	0.704	0.511	0.524	0.692	0.513	0.569	0.731	0.544
	Fst Prob	0.521	0.587	0.515	0.525	0.701	0.586	0.528	0.690	0.531	0.546	0.730	0.580
	Predictive Entropy	0.566	**0.596**	0.574	0.568	0.704	0.532	0.557	0.695	0.564	0.571	0.734	0.548
	Mink Entropy	0.550	0.593	0.543	0.535	0.702	0.592	0.532	0.691	0.519	0.554	0.735	0.581
	Attentional Entropy	0.517	0.587	0.488	0.576	0.703	**0.593**	0.519	0.690	0.495	0.572	0.732	0.578
	P(True)	0.538	0.591	0.541	0.563	0.706	0.566	0.550	0.692	0.570	0.645	**0.747**	**0.672**
	Internal Confidence	0.513	0.585	0.486	0.532	0.701	0.512	0.540	0.691	0.526	0.609	0.742	0.638
	Estimator	**0.704**	0.595	**0.754**	**0.703**	**0.732**	0.461	**0.700**	**0.732**	**0.759**	**0.709**	0.745	0.455
posterior	Perplexity	0.515	0.585	0.497	0.507	0.701	0.539	0.505	0.689	0.469	0.507	0.730	0.545
	Fst Prob	0.527	0.586	0.524	0.518	0.701	0.542	0.521	0.692	0.522	0.542	0.732	0.552
	Min Prob	0.525	0.587	0.518	0.514	0.702	0.490	0.527	0.691	0.520	0.535	0.732	0.520
	Prod Prob	0.526	0.586	0.501	0.530	0.702	0.511	0.529	0.689	0.518	0.548	0.730	0.545
	Logprob	0.527	0.586	0.503	**0.545**	0.705	0.511	0.534	0.689	0.531	0.549	0.731	0.545
	Predictive Entropy	**0.609**	0.595	0.603	0.537	0.712	0.512	0.546	0.689	0.515	0.571	0.730	0.543
	Mink Entropy	0.571	**0.599**	0.567	0.523	0.711	0.543	0.549	0.692	0.546	0.530	0.732	0.567
	LogTokU	0.534	0.596	0.497	0.542	0.686	0.510	0.515	0.664	0.481	0.544	0.724	0.508
	Hidden Score	0.550	0.590	0.548	0.533	**0.713**	**0.548**	0.547	0.697	0.545	0.549	0.731	0.563
	Attn Score	0.556	0.585	0.559	0.544	0.712	0.504	0.566	**0.701**	0.572	0.564	0.731	0.549
	Estimator	0.583	0.491	**0.624**	0.533	0.695	0.519	**0.585**	0.582	**0.618**	**0.662**	**0.733**	**0.568**

Table 8: The impact of document on different uncertainty estimation methods in NQ, TriviaQA and WebQ dataset for Llama-series models.

Method	Qwen2.5-7B			Qwen2.5-7B-Instruct			Llama3.1-8B			Llama-3.1-8B-Instruct		
	acc	ans-f1	abs-f1	acc	ans-f1	abs-f1	acc	ans-f1	abs-f1	acc	ans-f1	abs-f1
NQ												
Vanilla-G	0.521	0.567	0.291	0.668	0.701	0.593	0.516	0.593	0.002	0.665	0.723	0.421
Vanilla-R	0.274	0.000	0.426	0.274	0.000	0.426	0.274	0.000	0.426	0.274	0.000	0.426
Verb-Prior	**0.595**	0.587	**0.618**	0.539	0.538	0.540	0.427	0.476	0.327	0.579	0.628	0.478
Verb-Post	0.564	0.563	0.569	0.391	0.313	0.457	0.466	0.520	0.339	0.584	0.631	0.480
COT	0.557	0.588	0.419	0.614	0.610	**0.621**	0.507	0.582	0.011	0.674	0.730	0.456
CAD	0.460	0.523	0.076	**0.681**	0.724	0.585	0.538	**0.618**	0.004	0.659	0.728	0.339
CKPLUG	0.498	0.553	0.201	0.676	**0.743**	0.550	0.503	0.577	0.006	0.667	0.732	0.404
LogtokU	0.519	0.563	0.338	0.568	0.562	0.580	0.493	0.524	0.412	0.475	0.506	0.427
P(True)	0.576	0.555	0.615	0.619	0.634	0.594	0.478	0.496	0.446	0.639	0.676	**0.554**
Internal Confidence	0.430	0.433	0.426	0.609	0.640	0.555	0.445	0.478	0.383	0.481	0.523	0.415
Our	0.584	**0.592**	0.557	0.661	0.696	0.582	**0.578**	0.588	**0.559**	**0.714**	**0.766**	0.543
TriviaQA												
Vanilla-G	0.671	0.710	0.495	**0.771**	**0.858**	0.555	0.664	0.736	0.005	0.756	0.837	0.321
Vanilla-R	0.246	0.000	0.391	0.246	0.000	0.391	0.246	0.000	0.391	0.246	0.000	0.391
Verb-Prior	0.672	0.693	**0.620**	0.585	0.646	0.495	0.482	0.571	0.277	0.618	0.717	0.387
Verb-Post	0.658	0.687	0.587	0.369	0.346	0.389	0.560	0.648	0.296	0.641	0.724	0.426
COT	0.688	0.718	0.569	0.718	0.760	**0.616**	0.654	0.724	0.019	0.769	0.850	0.389
CAD	0.566	0.625	0.207	0.758	0.846	0.547	0.636	0.704	0.002	0.720	0.793	0.247
CKPLUG	0.632	0.682	0.375	0.765	0.855	0.544	0.669	0.741	0.006	0.768	0.865	0.308
LogtokU	0.655	0.696	0.501	0.680	0.712	0.595	0.573	0.649	0.298	0.579	0.680	0.358
P(True)	0.661	0.681	0.616	0.668	0.738	0.541	0.519	0.587	0.374	0.627	0.716	0.434
Internal Confidence	0.662	0.683	0.616	0.667	0.738	0.539	0.508	0.579	0.365	0.585	0.678	0.404
Our	**0.708**	**0.741**	0.592	0.764	0.850	0.535	**0.739**	**0.775**	**0.567**	**0.799**	**0.881**	**0.513**
WebQ												
Vanilla-G	0.465	0.500	0.297	0.564	0.590	0.508	0.492	0.565	0.007	0.588	0.648	0.292
Vanilla-R	0.281	0.000	0.436	0.281	0.000	0.436	0.281	0.000	**0.436**	0.281	0.000	0.436
Verb-Prior	0.494	0.497	0.487	0.497	0.489	0.509	0.407	0.452	0.310	0.533	0.583	0.412
Verb-Post	0.486	0.494	0.461	0.340	0.227	0.432	0.428	0.495	0.248	0.547	0.587	0.435
COT	0.490	0.519	0.364	0.529	0.527	0.533	0.481	0.553	0.011	0.594	0.651	0.327
CAD	0.392	0.451	0.044	0.564	0.601	0.468	0.498	0.572	0.004	0.569	0.639	0.224
CKPLUG	0.455	0.503	0.214	0.569	**0.629**	0.503	0.492	0.566	0.000	0.595	0.658	0.305
LogtokU	0.458	0.484	0.373	0.459	0.436	0.502	0.456	0.476	0.413	0.329	0.294	0.365
P(True)	0.481	0.463	0.516	0.494	0.482	0.510	0.479	0.505	0.420	0.524	0.561	0.448
Internal Confidence	0.485	0.454	0.540	0.519	0.516	0.523	0.401	0.392	0.413	0.540	0.585	0.430
Our	**0.552**	**0.552**	**0.550**	**0.587**	0.622	**0.531**	**0.546**	**0.587**	0.395	**0.629**	**0.683**	**0.470**

Table 9: The average results of NQ, TriviaQA and WebQ datasets.

Method	Prompt
Default	You need to complete the question-and-answer pair. The answers should be short phrases or entities, not full sentences. The following contexts will help you complete the question-and-answer pair. If you don't know the answer and the following contexts do not contain the necessay information to answer the question, respond with 'This question is beyond the scope of my knowledge and the references, I don't know the answer'. {Documents} Question:{question} Answer:
COT	You need to complete the question-and-answer pair. The answers should be short phrases or entities, not full sentences. The following contexts will help you complete the question-and-answer pair. If you don't know the answer and the following contexts do not contain the necessay information to answer the question, respond with 'This question is beyond the scope of my knowledge and the references, I don't know the answer'.Please think step by step. {Documents} Question:{question} Answer:
Verb Prior	Respond only with 'Yes' or 'No' to indicate whether you are capable of answering the question accurately based on your knowledge or given documents. The provided reference contexts are as follows: {Documents} Question:{question} Answer Yes or No:
Verb Post	Respond only with 'Yes' or 'No' to indicate whether the provided text can answer the question based on the given documents or your internal knowledge. {Documents} Question:{question} Provided text: {output} Answer Yes or No:

Table 10: The prompt templates used in our experiments.

Category	**Metric**	**Formula**	**Description**
Overall Quality	Accuracy	$\frac{\lvert ✔ \cap (✔✔ \cup ✔✘ \cup ✘✔) \rvert + \lvert ○ \cap ✘✘ \rvert}{\lvert ✔✔ \cup ✔✘ \cup ✘✔ \cup ✘✘ \rvert}$	Ratio of correct answers plus proper abstentions to total queries
Answer Quality	Recall	$\frac{\lvert ✔ \cap (✔✔ \cup ✔✘ \cup ✘✔) \rvert}{\lvert ✔✔ \cup ✔✘ \cup ✘✔ \rvert}$	Ratio of correct answers to all queries in $\mathrm{KB}_{\mathrm{rag}}$
	Precision	$\frac{\lvert ✔ \cap (✔✔ \cup ✔✘ \cup ✘✔) \rvert}{\lvert ✔ \rvert + \lvert ✘ \rvert}$	Ratio of correct answers to attempted answers
	F1	$\frac{2 \cdot \mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}$	The harmonic mean of precision and recall
Abstain Quality	Abstain Recall	$\frac{\lvert ○ \cap ✘✘ \rvert}{\lvert ✘✘ \rvert}$	Ratio of correct abstentions to all queries in ✘✘
	Abstain Precision	$\frac{\lvert ○ \cap ✘✘ \rvert}{\lvert ○ \rvert}$	Ratio of correct abstentions to all abstentions
	Abstain F1	$\frac{2 \cdot \mathrm{AbPrec} \cdot \mathrm{AbRec}}{\mathrm{AbPrec} + \mathrm{AbRec}}$	The harmonic mean of abstain precision and abstain recall

Table 11: Evaluation Metrics based on the knowledge quadrant division. Let ✔ denote correct answers, ✘ denote incorrect answers, and ○ denote abstentions ("I don't know" responses). For any category (e.g., ✔✘), $\lvert ✔ \cap ✔✘ \rvert$ represents the count of correct answers within the ✔✘ category.

Method	Qwen2.5-7B			Qwen2.5-7B-Instruct			Qwen2.5-7B			Qwen2.5-7B-Instruct		
	0doc						10doc					
	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1
NQ												
Vanilla-G	0.588	0.523	0.637	0.690	0.699	0.686	0.459	0.516	0.129	0.596	0.638	0.493
Vanilla-R	0.274	0.000	0.426	0.274	0.000	0.426	0.274	0.000	0.426	0.274	0.000	0.426
Verb-Prior	0.715	0.450	0.808	0.730	0.506	0.799	0.494	0.518	0.403	0.516	0.519	0.512
Verb-Post	0.631	0.529	0.696	0.705	0.431	0.794	0.507	0.523	0.457	0.400	0.336	0.458
COT	0.294	0.292	0.297	0.725	0.496	0.812	0.487	0.536	0.239	0.519	0.511	0.538
CAD							0.400	0.462	0.010	0.585	0.633	0.445
CKPLUG							0.452	0.510	0.107	0.616	**0.683**	0.481
LogtokU	0.574	0.493	0.634	0.715	0.529	0.793	0.462	0.517	0.172	0.502	0.501	0.506
P(True)	0.712	0.516	0.795	0.766	0.639	0.808	0.495	0.483	**0.518**	0.571	0.594	0.533
Internal Confidence	0.748	0.325	**0.844**	0.719	0.677	0.738	0.437	0.470	0.348	0.529	0.556	0.485
Our	**0.753**	**0.562**	0.830	**0.797**	**0.728**	**0.821**	**0.545**	**0.553**	0.518	**0.630**	0.666	**0.558**
TriviaQA												
Vanilla-G	0.762	**0.792**	0.724	0.726	**0.860**	0.520	0.597	0.656	0.255	0.719	0.805	0.469
Vanilla-R	0.246	0.000	0.391	0.246	0.000	0.391	0.246	0.000	0.391	0.246	0.000	0.391
Verb-Prior	0.629	0.466	0.716	0.713	0.800	0.615	0.627	0.659	0.528	0.561	0.630	0.463
Verb-Post	0.778	0.783	0.773	0.578	0.494	0.635	0.623	0.663	0.490	0.427	0.447	0.405
COT	0.437	0.469	0.377	0.683	0.681	0.686	0.630	0.674	0.426	0.685	0.725	**0.564**
CAD							0.459	0.520	0.029	0.695	0.768	0.456
CKPLUG							0.559	0.620	0.181	**0.742**	**0.832**	0.481
LogtokU	0.733	0.757	0.703	0.701	0.714	0.687	0.591	0.650	0.294	0.631	0.676	0.514
P(True)	0.671	0.567	0.735	0.672	0.670	0.672	0.586	0.613	0.532	0.641	0.719	0.497
Internal Confidence	0.725	0.680	0.758	0.666	0.684	0.651	0.599	0.627	0.538	0.643	0.723	0.492
Our	**0.794**	0.780	**0.808**	**0.777**	0.856	**0.691**	**0.693**	**0.714**	**0.626**	**0.742**	0.829	0.496
WebQ												
Vanilla-G	0.593	0.629	0.549	0.726	**0.858**	0.620	0.411	0.460	0.148	0.523	0.563	0.424
Vanilla-R	0.281	0.000	0.436	0.281	0.000	0.436	0.281	0.000	0.436	0.281	0.000	0.436
Verb-Prior	0.633	0.485	0.718	0.682	0.715	0.660	0.441	0.464	0.357	0.483	0.492	0.468
Verb-Post	0.616	0.626	0.605	0.636	0.403	0.730	0.434	0.460	0.335	0.395	0.338	0.449
COT	0.334	0.389	0.244	0.624	0.506	0.699	0.442	0.485	0.234	0.493	0.507	0.453
CAD							0.339	0.393	0.007	0.505	0.537	0.406
CKPLUG							0.417	0.469	0.133	0.566	0.625	0.447
LogtokU	0.572	0.589	0.552	0.593	0.513	0.657	0.415	0.439	0.346	0.399	0.361	0.452
P(True)	0.673	0.553	0.742	0.729	0.765	0.708	0.438	0.440	0.434	0.482	0.490	0.469
Internal Confidence	0.675	0.575	0.736	0.730	0.751	0.719	0.452	0.434	0.487	0.496	0.509	0.474
Our	**0.732**	**0.652**	**0.785**	**0.767**	0.800	**0.748**	**0.526**	**0.528**	**0.522**	**0.581**	**0.630**	**0.501**

Table 12: The results across NQ, TriviaQA and WebQ datasets under 0doc and 10doc. CAD and CKPLUG are not applicable for 0doc setting.

Method	NQ			TriviaQA			WebQ		
	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1	Acc	Ans-f1	Abs-f1
Vanilla-G	0.475	0.521	0.250	0.473	0.549	0.179	0.395	0.433	0.213
Vanilla-R	0.274	0.000	0.426	0.246	0.000	0.391	0.281	0.000	0.436
Verb-Prior	0.550	0.549	0.554	0.603	0.591	0.632	0.449	0.448	0.449
Verb-Post	0.559	0.554	0.571	0.599	0.583	0.623	0.478	0.465	0.507
COT	0.418	0.458	0.218	0.394	0.458	0.135	0.301	0.322	0.195
CAD	0.536	**0.569**	0.402	0.542	0.610	0.315	0.440	0.471	0.300
CKPLUG	0.422	0.465	0.208	0.448	0.521	0.158	0.337	0.373	0.152
LogtokU	0.500	0.530	0.389	0.516	0.564	0.387	0.411	0.434	0.323
P(True)	0.559	0.544	**0.664**	0.570	0.621	**0.744**	0.485	0.437	**0.585**
Internal Confidence	0.421	0.394	0.458	0.547	0.546	0.549	0.420	0.385	0.488
Our	**0.562**	0.532	0.614	**0.654**	**0.623**	0.709	**0.510**	**0.480**	0.562

Table 13: The results on NQ, TriviaQA and WebQ datasets for Qwen3-14B model.Only prior control is employed without posterior control due to time constraints.