Do Retrieval Augmented Language Models Know When They Don't Know?

Youchao Zhou[1,3*], Heyan Huang[1,3†], Yicheng Liu[1], Rui Dai[1], Xinglin Wang[1], Xingchen Zhang[1], Shumin Shi[1], Yang Deng[2]

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]School of Computing and Information Systems, Singapore Management University
[3]Southeast Academy of Information Technology, Beijing Institute of Technology
{ yczhou, hhy63, lyc2024, ruidai, wangxinglin, zxc2024, bjssm }@bit.edu.cn, ydeng@smu.edu.sg

Abstract

Existing large language models (LLMs) occasionally generate plausible yet factually incorrect responses, known as hallucinations. Two main approaches have been proposed to mitigate hallucinations: retrieval-augmented language models (RALMs) and refusal post-training. However, current research predominantly focuses on their individual effectiveness while overlooking the evaluation of the refusal capability of RALMs. Ideally, if RALMs know when they do not know, they should refuse to answer. In this study, we ask the fundamental question: Do RALMs know when they don't know? Specifically, we investigate three questions. First, are RALMs well calibrated with respect to different internal and external knowledge states? We examine the influence of various factors. Contrary to expectations, when all retrieved documents are irrelevant, RALMs still tend to refuse questions they could have answered correctly. Next, given the model's pronounced **over-refusal** behavior, we raise a second question: How does a RALM's refusal ability align with its calibration quality? Our results show that the over-refusal problem can be mitigated through in-context fine-tuning. However, we observe that improved refusal behavior does not necessarily imply better calibration or higher overall accuracy. Finally, we ask: Can we combine refusal-aware RALMs with uncertainty-based answer abstention to mitigate over-refusal? We develop a simple yet effective refusal mechanism for refusal-post-trained RALMs that improves their overall answer quality by balancing refusal and correct answers. Our study provides a more comprehensive understanding of the factors influencing RALM behavior. Meanwhile, we emphasize that uncertainty estimation for RALMs remains an open problem deserving deeper investigation.

Code — https://github.com/zuochao912/refusal-ability-of-retrieval-augmented-LLMs

Extended version — https://arxiv.org/abs/2509.01476

Introduction

Existing large language models (LLMs) have demonstrated remarkable performance across a wide range of challenging tasks. However, they occasionally generate plausible yet factually incorrect responses—a phenomenon commonly known as hallucinations (Lewis et al. 2020; Huang et al. 2025). Prior research has primarily addressed this issue through two approaches: retrieval-augmented generation (RAG) (Lewis et al. 2020; Ram et al. 2023) and refusal post-training (Zhang et al. 2024; Zhu et al. 2025). RAG leverages external knowledge sources to provide contextual grounding, enabling retrieval-augmented language models (RALMs) to answer queries beyond their internal (parametric) knowledge. In contrast, refusal post-training aims to enhance a model's ability to proactively abstain from answering when uncertain.

Although both methods are widely adopted, prior work has predominantly emphasized their individual effectiveness while overlooking systematic evaluation of the refusal capabilities of RALMs. Given that LLMs are sensitive to the quality and relevance of retrieval contexts (Park and Lee 2024; Cuconasu et al. 2024), a refusal-trained model might mishandle unreliable external information and become uncertain even when it internally possesses correct knowledge. As shown in Figure 1, RALMs may over-refuse questions that they would otherwise answer correctly when confronted with irrelevant documents. To address this gap, we pose the fundamental question: *Do RALMs know when they do not know?* Ideally, if RALMs know when they don't know (are well calibrated), they can refuse to answer, or users can post-hoc reject their answers based on model uncertainty.

Specifically, in this work, we study three critical research questions (RQs). First, *are RALMs well calibrated with respect to different internal and external knowledge states?* (**RQ1**) We categorize knowledge states as shown in Figure 1 and quantify the knowledge state of RALMs using uncertainty estimates. We also explicitly consider refusal behavior, which has been overlooked in prior work on uncertainty estimation. While models demonstrate improved calibration when a supportive document exists within otherwise irrelevant contexts, we find that RALMs exhibit significant **over-refusal** behavior, particularly when confronted with exclusively irrelevant contexts; that is, LLMs still tend to refuse questions they could have answered correctly.

Second, given the over-refusal tendency observed in RALMs, we pose our second research question: *How does a RALM's refusal ability align with its calibration quality?* (**RQ2**) We modify the refusal behavior of RALMs us-

[*]This work was done during an internship at SMU
[†]Corresponding Author

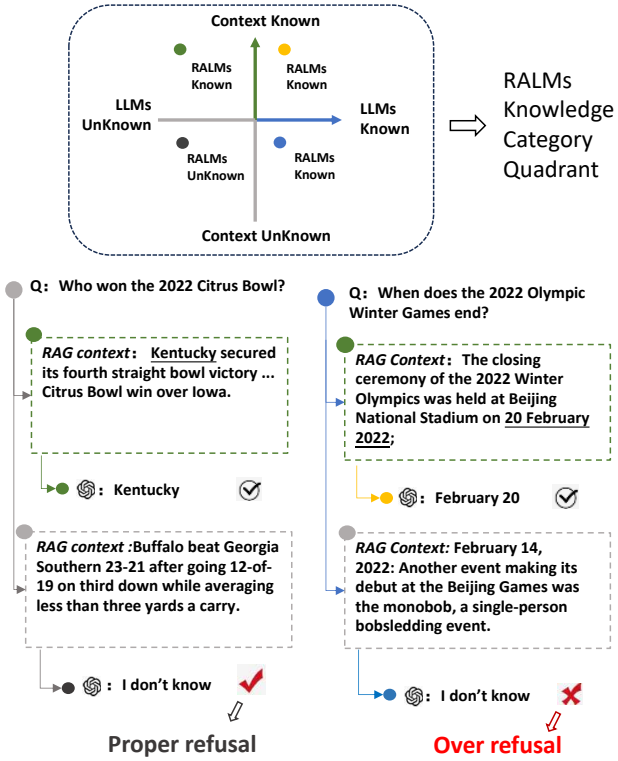

Figure 1: An illustration of the knowledge boundary of a RALM and the corresponding answer correctness. We divide the knowledge state into four quadrants based on the model's internal knowledge and the knowledge provided by external context. The question at the gray dot lies outside the model's knowledge boundary, whereas the question at the blue dot lies within it. However, given irrelevant context, the model may still refuse to answer the blue-dot question.

ing two instruction-tuning-based methods: Refusal-Aware Instruction Tuning (R-tuning) (Zhang et al. 2024) and In-Context Fine-Tuning (ICFT) (Lee, Lin, and Tan 2025; Zhu, Panigrahi, and Arora 2025). Our results show that the over-refusal problem is mitigated by ICFT but exacerbated by R-tuning. However, we observe that improved refusal performance does not necessarily imply better calibration or higher answer accuracy. We attribute these discrepancies to changes in robustness and contextual faithfulness.

Lastly, given the difficulty of balancing refusal and response competence based solely on the behavior of LLMs themselves, we investigate our third research question: *Can we combine refusal-aware RALMs with uncertainty-based answer abstention to mitigate over-refusal?* (**RQ3**) Building on our previous findings, we leverage uncertainty and its variation to infer the knowledge state of RALMs, and then decide whether to answer a question with or without retrieved context, or to abstain altogether.

Our contributions are threefold: 1) We investigate the uncertainty calibration of RALMs and conduct a comprehensive analysis of key factors, including context variation and different knowledge states (internal vs. external knowledge). 2) We identify and characterize the over-refusal problem, and then examine the relationship between refusal behavior and calibration. In particular, we study whether existing refusal tuning exacerbates over-refusal in LLMs and provide further explanations. 3) We design a simple yet effective refusal method for RALMs, informed by the above findings.

Related Works

Knowledge Boundary of LLMs. Identifying the knowledge boundary of an LLM helps delineate the limits of its knowledge (Deng et al. 2025). This notion is also described as "knowing what you don't know" (Yin et al. 2023; Deng et al. 2024), which is crucial for assessing the practical applicability of LLMs. Li et al. (2025) formally categorizes the knowledge boundary with respect to prompt and model sensitivity. However, these works mainly focus on the LLMs' internal knowledge. Hallucinations typically occur when users' requests fall outside the LLM knowledge boundary (Huang et al. 2025). The primary approach to mitigating hallucinations is retrieval-augmented generation (RAG) (Lewis et al. 2020). More advanced RAG variants leverage LLM self-generated rationales (Wei, Chen, and Meng 2024), perform post-retrieval knowledge selection (Xu, Shi, and Choi 2024; Li et al. 2024), or adopt dynamic retrieval strategies (Jeong et al. 2024). Recent dynamic RAG methods (Asai et al. 2024; Su et al. 2024) still rely on confidence and manually chosen thresholds to decide when retrieval is necessary; even though the system's knowledge may evolve dynamically, these thresholds remain static. This implicitly assumes that the model is always well calibrated. To the best of our knowledge, no prior work has systematically analyzed the factors that influence the uncertainty of RALMs, and our study fills this gap.

Refusal Method of LLMs. Refusal behavior has predominantly been studied at the post-training stage (Wen et al. 2025). Existing work mainly focuses on instruction tuning (Zhang et al. 2024; Zhu et al. 2025; Kapoor et al. 2024) and refusal-alignment training (Cheng et al. 2024; Sun et al. 2025). In these setups, instances where the model is uncertain or produces incorrect answers are typically labeled as "should-refuse" examples. Another line of work controls refusal at inference time (Feng et al. 2024), where uncertainty estimates are used to abstain from answering by thresholds.

Uncertainty Estimation. Current research typically treats uncertainty and confidence as opposite quantities (Lin, Trivedi, and Sun 2024); that is, the higher the uncertainty of an LLM, the lower its confidence. Geng et al. (2024) divide uncertainty estimation (UE) methods for LLMs into white-box and black-box approaches. White-box methods are suitable for open-source LLMs, where internal states are accessible (Kadavath et al. 2022). By contrast, black-box methods rely solely on model responses for UE and therefore have broader applicability. Recent work discusses the UE of RALMs (Moskvoretskii et al. 2025) and Language Reasoning Models (Mei et al. 2025; Soudani, Zamani, and Hasibi 2025). However, these studies do not construct controlled experimental settings to analyze the influence of specific factors, and they neglect the model's refusal behavior.

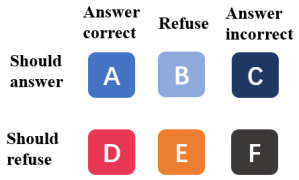

Figure 2: Refusal and answer confusion matrix. "Should answer/refuse" is the ground truth label while "answer correct/incorrect", refuse is the response situation.

Preliminary

We briefly describe the concept of proper refusal and over-refusal. We illustrate the refusal and answer and their correctness situation as in Figure 2. According to (Feng et al. 2024), the questions could be divided into "should refuse" and "should answer". If LLMs tend to give false answers, which means that LLMs do not entail knowledge, then they should refuse the question. Thus the proper refusal rate is $\frac{E}{D+E+F}$ and the over-refusal rate is $\frac{B}{A+B+C}$. Notice that the "C" and "D" parts exist in our settings. This arises from the threshold used under repeated sampling and the model's prompt sensitivity.

Methodology

Uncertainty Estimation Methods

We primarily adopt black-box UE methods to quantify the confidence of LLM responses, as they are more broadly applicable. Following (Moskvoretskii et al. 2025), we select three categories of well-performing UE methods.

Verbalization-based UE This class of methods leverages the LLM's self-awareness and expressive ability by eliciting explicit confidence estimates for its answers via prompting. We design four different prompts following (Tian et al. 2023). These prompt variants mainly differ in (i) whether the answer and its uncertainty estimate are produced within the same conversation turn, and (ii) the number of generations elicited. Detailed prompt descriptions are provided in Appendix A.

Consistency-based UE This class of methods is based on the assumption that more consistent answers indicate higher model confidence. Lyu et al. (2025a) propose an alternative approach to quantifying the uncertainty of LLMs and apply it to decoding strategies such as self-consistency. We formalize three types of consistency-based measures as follows. For a given input x and an LLM $M(\cdot)$, we generate m responses $\{r_1, r_2, \ldots, r_m\}$ and decide the final answer via majority voting:

$$\bar{r} = \arg\max_r \sum_{i=1}^{m} \mathbb{1}I(r_i = r),$$

where $\mathbb{1}I(\cdot)$ is the indicator function.

The first measurement $Agree(\cdot)$ is based on agreement among answers:

$$Agree(\bar{r}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(r_i = \bar{r}), \tag{1}$$

where the agreement indicator could be implemented as semantic or lexical agreement, or LLM-as-judge.

The second measurement $Ent(\cdot)$ is entropy-based and rescales the weights of each answer. It is computed as:

$$Ent(r) = 1 - (-\frac{1}{log|\bar{r}|} \sum_{i=1}^{|\hat{r}|} p_i log(p_i)), \tag{2}$$

where $\hat{r}$ is the set of duplicated answers, p_i is the probability of the unique answer r_i.

The final measurement $FSD(\cdot)$ balances the two ways, which is based on the top two most-voted responses $\bar{r}$ and $\bar{\bar{r}}$:

$$FSD(r) = Agree(\bar{r}) - Agree(\bar{\bar{r}})). \tag{3}$$

Similarity Matrix based UE This kind of methods consider the similarity of all responses. We use two features,including degree and eigenvalue of the similarity matrix following (Lin, Trivedi, and Sun 2024). The formulations are in the Appendix A.

Refusal Post-Training Methods

We aim to adjust the proactive refusal behavior of RALMs. We adopt two refusal instruction tuning (RIFT) methods, namely R-tuning and in-context fine-tuning (ICFT), due to their broad adoption. Further implementation details are provided in Appendix A.

R-tuning. R-tuning (Zhang et al. 2024) is a simple yet effective method for teaching LLMs to issue appropriate refusals. Its workflow typically consists of two stages. In the first stage, the questions that the LLM cannot answer correctly are detected. In the second stage, training data are constructed and instruction tuning is performed. For questions outside the model's knowledge boundary, we assign refusal targets such as "I don't know".

In-Context Fine-Tuning. Zhu, Panigrahi, and Arora (2025); Lee, Lin, and Tan (2025) find that inserting positive context into prompts during instruction tuning improves LLM accuracy. However, they generally append only positive context and train the model to generate correct answers. Fang et al. (2024); Yoran et al. (2024) adopt a similar strategy but optimize a corresponding training objective to enhance robustness and faithfulness. In our work, we extend this framework to the refusal setting. For each training example, we insert not only positive context but also negative context. We set the training targets to either a correct answer or a refusal expression according to the knowledge-state quadrant of the RALM, as illustrated in Figure 1. When the knowledge is unknown to the RALM, we set the answer to a refusal expression.

Experiments

Experimental Setup

To focus on the model's knowledge capacity while minimizing the influence of reasoning, we primarily consider simple factual questions with short answers. These questions typically require only a single evidence document to be answered correctly, for which single-step retrieval is sufficient. Additional details are described in Appendix B.

UE type	UE name	RGB_{en}					RGB_{zh}				
		no context	0p10n	1p9n	5p5n	1p19n	no context	0p10n	1p9n	5p5n	1p19n
Verbalize	Verb-1s-1	0.445	**0.139**	0.208	0.023	0.042	0.477	0.441	0.119	0.242	0.124
	Verb-1s-5	0.253	0.186	0.182	0.160	0.179	**0.173**	0.170	0.182	0.170	0.198
	Verb-2s-1	0.339	0.190	0.183	0.013	0.040	0.448	0.338	0.122	0.210	0.125
	Verb-2s-5	0.225	0.190	0.176	0.124	0.178	0.204	**0.165**	0.412	0.240	0.442
Consistency	Ent	0.126	0.305	0.030	**0.009**	0.033	0.253	0.256	0.093	0.148	0.082
	Agree	0.127	0.192	**0.026**	0.010	**0.028**	0.250	0.261	**0.078**	0.150	**0.075**
	FSD	**0.104**	0.162	0.041	0.014	0.048	0.201	0.182	0.083	**0.122**	0.086
Similarity Matrix-based	Eigv	0.202	0.232	0.289	0.271	0.260	0.247	0.282	0.299	0.271	0.284
	Deg	0.200	0.229	0.292	0.275	0.262	0.236	0.277	0.297	0.268	0.283

Table 1: The Brier score (lower score indicates better calibration) of different UE methods on different RAG settings and datasets.The "ApBn" means A positive chunks and B negative chunks for RAG context settings.

RALM Models We adopt two prevalent families of open-source LLMs, Qwen and LLaMA. Although modern LLMs are multilingual, We find that Qwen has stronger knowledge in Chinese, whereas LLaMA performs better on English knowledge. To better exploit the knowledge of each model family, we evaluate Qwen on Chinese datasets and LLaMA on English datasets. *In the main text, we mainly report results for models with approximately 7B parameters.* For the retrieval component, document chunks and positive ground-truth passages are provided by the original datasets. We perform hybrid search and re-ranking using Milvus to construct high-quality negative examples, taking both semantic and lexical similarity into account to provide sufficient difficulty.

Hyper-Parameters The generation temperature is set to 0.5, and the number of sampled generations is set to 16, following (Lyu et al. 2025a). Other generation hyperparameters are kept at the default values for the corresponding LLMs.

Datasets We explore the RALMs' performance in open-domain QA tasks, using three prevalent fact-oriented single-hop question datasets to evaluate the performance of LLMs, including two RAG datasets, CRUD (Lyu et al. 2025b) and RGB (Chen et al. 2024), and an QA dataset, NQ (Kwiatkowski et al. 2019). Covering both Chinese and English, the datasets are well-suited for testing Qwen and LLaMA series.

Answer Judgment We first assign a knowledge state to each question based on both temperature-sampled and greedy-decoding results, following (Gekhman et al. 2024). This yields four categories: "highlyknown", "maybeknown", "weaklyknown", and "unknown". We treat the former two categories as "should answer" and the latter two as "should refuse" according to the precision analysis in Section of RQ1. Following (Sun et al. 2025), we then apply a strict answer-decision workflow to determine whether a model output should be regarded as a refusal or a correct answer, including an LLM-as-a-judge step, exact-match checking, and a refusal-word filter.

Evaluation Metrics Evaluation metrics include accuracy-based and confidence-calibration measures (Feng et al. 2024; Sun et al. 2025). The formal definitions of all metrics are given in Appendix B, and we briefly summarize them as follows:

- **Accuracy-based metrics**: The answering ability of RALMs is multi-dimensional, reflecting both answer quality and refusal quality.
  - Answer Quality (AQ): We report answer precision (Pre), recall (Rec), and F1 for correct answers.
  - Refusal Quality (RQ): We measure the refusal rate(RR), refusal precision (RPrec), recall (RRec) and F1(RF1).
  - Overall Quality (OQ): We report overall accuracy (OAcc), defined as the proportion of outputs that are either correct answers or proper refusals.
- **Confidence calibration metrics**: We mainly use Brier Score to measure whether the answer confidence measure precision.

Do RALMs Know When They Don't Know? (RQ1)

We systematically investigate how prompt variants, positive context position, context quality, and quantity affect the model performance.[1] We heuristically varied the numbers of positive and negative examples and examined their impact on the results. In this section, we first examine the calibration error with different UE methods to choose the best one for the following analysis. We then analyze confidence and accuracy in turn as they contribute to the calibration results.

Calibration error of RALMs. We exclude refusals for UE, since they are outcome-level decisions co-equal with answering, not comparable to specific answer content. Results are in Table 1. The calibration error varies under different RAG settings, and no single method performs best across all scenarios. This aligns with (Moskvoretskii et al. 2025). However, the RALMs become extremely well-calibrated when positive documents exist, especially for verbalize and consistency-based UE methods. This indicates that the UE methods are also acceptable for RALMs. **As the consistency-based methods perform best generally**, we take their results for further explanation. We contrast the presence versus the absence of context. We find that when

[1]Detailed discussions are in Appendix C

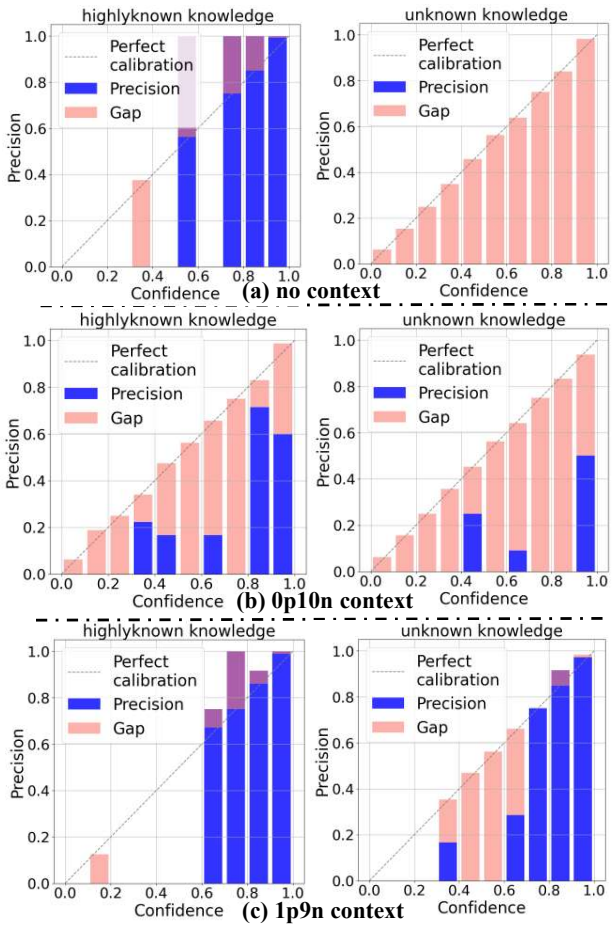

Figure 3: The reliability diagram under different internal and external knowledge states. The blue bar is the answer precision. The pink bar indicates the over-confident gap, and the purple bar indicates the under-confident gap.

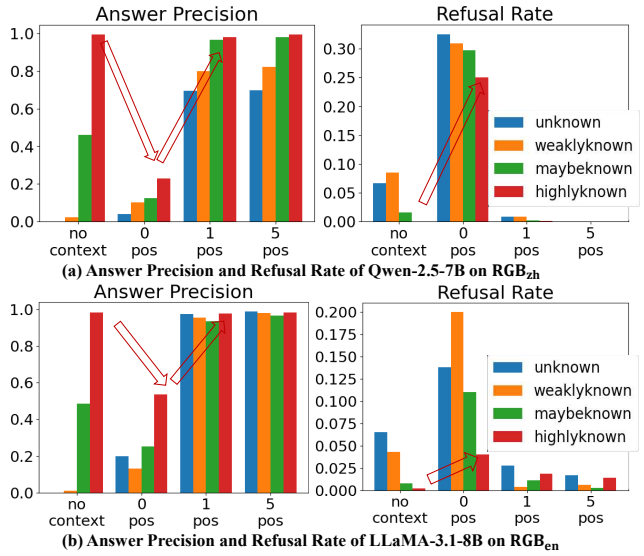

Figure 4: The answer precision and refusal rate vary according to the internal/external knowledge states, where context chunks are 10. The whole negative context (0 pos) leads to significant increase of refusal on "highlyknown" questions.

no positive context exists (0p10n), the calibration error becomes worse. And when we insert a single positive context (1p9n), the model becomes extremely calibrated. If we insert more positive context (5p5n), the trend of calibration error vary, become better on RGB_{en} and worser on RGB_{zh}. And if we insert more negative context (1p19n), the calibration error does not significantly change. This means that RALMs can sensitively perceive the availability of knowledge. **As we find the key factor is the positive context existence**, the following settings use 10 context chunks as the default.

Over-confident or under-confident. In this section, we examine how confidence scores vary as shown in Figure 3, given that base LLMs are known to be over-confident (Li et al. 2025). *In the no-context setting*, the "highlyknown" type is slightly under-confident, whereas the other types are over-confident. However, *in the all-negative-context setting*, RALMs become strongly over-confident and the confidence scores become dispersed. For "highlyknown" questions, the LLM could answer correctly without retrieval, yet the observed accuracy is noticeably worse. This indicates that both accuracy and confidence are substantially affected by noisy contexts. Interestingly, "weaklyknown" questions achieve higher accuracy under negative contexts, suggesting that the injected noise can have unexpected effects. This finding is consistent with Cuconasu et al. (2024), while we further delineate how this effect depends on specific knowledge categories. Finally, even *when one positive context is provided*, RALMs tend to be under-confident for most knowledge types, except for the "unknown" category. Across knowledge types, the model attains high accuracy and more concentrated confidence distributions, indicating that RALMs can effectively detect and exploit helpful information. In summary, these observations explain the calibration trends in Table 1: with all-negative context, accuracy generally decreases and confidence becomes more diffuse, whereas with positive context, accuracy improves and confidence becomes more concentrated.

Precision and refusal rate. We begin by analyzing how answer correctness varies. *In the all-negative(0 pos) setting*, we observe a decline on "highlyknown" and "maybeknown" questions and a gain on "weaklyknown" and "unknown" ones compared to the no-context setting. When a positive context exists, the precision significantly increases, especially for unknown and weakly known knowledge. Increasing the count of positives yields no significant gains in precision. This indicates that LLMs are sensitive to both harmful and supportive contexts. While increasing the number of positive and negative examples does not significantly alter the model's response for fact-oriented questions in this kind of shorter context. *Then we analyze refusal rate.* In the all-negative (0 pos) setting, we observe an significant increase on all the knowledge types. Considering the LLMs can correctly answer "highlyknown" questions on their own,

RALMs test setting	Method name	CalErr	OQ	AQ				RQ				
		OaBs (↓)	OAcc(↑)	Pre(↑)	Rec(↑)	F1(↑)	MA(↓)	RR	OR(↓)	RPre(↑)	RRec(↑)	RF1(↑)
Qwen-2.5-7B												
no context	Vanilla	0.245	0.427	0.411	1.000	0.583	**0.217**	0.027	0.000	**1.000**	0.044	0.085
	R-tuning	0.191	0.457	0.395	0.857	0.541	0.336	**0.190**	**0.105**	0.719	**0.218**	**0.335**
	ICFT (n)	0.226	**0.487**	**0.450**	0.953	0.611	0.250	0.103	0.039	0.806	0.145	0.245
	ICFT (p)	0.169	0.443	0.443	1.000	**0.614**	0.250	0.000	0.000	0.000	0.000	0.000
	ICFT (pn)	**0.167**	0.440	0.440	1.000	0.611	0.243	0.000	0.000	0.000	0.000	0.000
	ICFT (w)	0.181	0.423	0.414	1.000	0.585	0.296	0.017	0.000	1.000	0.028	0.055
0p10n	Vanilla	0.325	0.290	0.168	0.372	0.231	0.500	0.363	0.355	0.505	0.257	0.341
	R-tuning	0.408	0.457	0.294	0.195	0.235	0.184	**0.717**	**0.678**	0.521	**0.651**	0.579
	ICFT (n)	0.216	**0.620**	**0.578**	0.709	**0.637**	**0.158**	0.423	0.270	**0.677**	0.541	**0.601**
	ICFT (p)	0.204	0.400	0.400	1.000	0.571	0.342	0.000	0.000	0.000	0.000	0.000
	ICFT (pn)	**0.189**	0.430	0.430	1.000	0.601	0.309	0.000	0.000	0.000	0.000	0.000
	ICFT (w)	0.217	0.460	0.436	0.976	0.603	0.296	0.060	0.020	0.833	0.086	0.156
1p9n	Vanilla	0.079	**0.863**	**0.863**	1.000	**0.927**	**0.013**	0.000	0.000	0.000	0.000	0.000
	R-tuning	0.127	0.830	0.853	0.960	0.903	0.033	0.070	0.066	0.524	0.212	0.301
	ICFT (n)	0.164	0.787	0.835	0.881	0.858	0.033	**0.230**	**0.171**	**0.623**	**0.531**	**0.573**
	ICFT (p)	**0.068**	0.827	0.827	1.000	0.905	0.072	0.000	0.000	0.000	0.000	0.000
	ICFT (pn)	0.085	0.820	0.820	1.000	0.901	0.059	0.000	0.000	0.000	0.000	0.000
	ICFT (w)	0.094	0.827	0.827	1.000	0.905	0.053	0.000	0.000	0.000	0.000	0.000

Table 2: Evaluation of refusal trained models under different settings. (↑) indicates a higher score is better, and (↓) vice versa. If no arrow is marked, then the score have no directionality. The best result under a RALMs test settings is marked bold and we do not mark those "1.000" scores. The over-refusal score (OR) which is marked in red indicates the worst case.

refusal on those questions are not correct. We identify this phenomenon as **over-refusal**, which are not observed in previously research. Likewise, the presence of positive chunk markedly reduces refusal. This is consistent with the pattern of accuracy changes.

Summary. In this section, we empirically show that RALMs generally "know they don't know" under no-context and positive-context settings. However, they become over-confident when confronted with negative context and may over-refuse questions whose answers they actually know.

How Does RALMs' Refusal Ability Align With Its Calibration Quality? (RQ2)

We adjust refusal ability though the R-tuning and In-context Fine-tuning variants. Considering the knowledge quadrants of Figure 1, we set four ICFT variants as follows:

- ICFT(n) : We append only negative contexts for LLMs, thus the answer of training samples depend on the internal state of LLMs. If internal knowledge entail the question, the answer is original ground truth; else the answer is "I don't known".
- ICFT(p) : We append only positive contexts for LLMs. The answers are all set to original ground truth.
- ICFT(pn): We append both positive and negative contexts for LLMs and the answers are all set to original ground truth. This is because the LLMs can distinguish the positive context and we want to enhance this ability.
- ICFT(w): We include both the ICFT(n) and ICFT(pn) training samples.

We use the training query, only different context and answers to ensure the training fairness. Training and model selection details are in Appendix D.

Response quality of RIFT models The response quality of refusal-trained RALMs is multi-dimensional. As shown in Table 2, model performance varies across different RALM settings. *In the no-context setting*, ICFT(n) achieves the best overall accuracy (OAcc, OQ), while ICFT(p) performs best in terms of F1 (AQ). The R-tuning model obtains the highest RF1 (RQ), with ICFT(n) ranking second. This may be because the R-tuning training scenario closely matches the test setting, leading to a higher refusal rate (RR) and moderate refusal precision (RPrec). However, the **over-refusal rate (OR) also increases**, suggesting that R-tuning may harm the model's self-awareness. The decrease in answer precision (Pre) and the increase in mis-answer rate (MR) support this finding. We will further examine the corresponding change in confidence calibration in the following subsection. *In the all-negative (0p10n) setting*, ICFT(n) performs substantially better than the other models in terms of OAcc (OQ), F1 (OQ), and RF1 (RQ). Although the over-refusal rate (OR) of R-tuning is the worst, ICFT(n) alleviates this issue and performs better than the vanilla RALMs. Moreover, we find that ICFT variants with positive context substantially reduce over-refusal while maintaining competitive overall accuracy (OAcc, OQ). Surprisingly, *when positive context is available*, the vanilla RALMs achieve the best OAcc (OQ) and F1 (AQ). From the perspective of RQ, ICFT(n) actually appears to perform the best. However, we emphasize that RQ in this positive-context setting should be interpreted with caution, as we do not relabel the "should-answer" set in order to remain consistent with the previous two settings.

Refusal Confidence of RIFT models In RQ1 we do not consider the refusal part, we check the overall brier score (OaBs) as in Table 2. We notice that the performance of calibration error do not align with overall,answer, or refusal quality. Surprisingly, ICFT with positive context(p/pn) get best calibration performance, though their refusal perfor-

Method name	DR		CU	
	no context	0p10n	10p0n	1p9n
Vanilla	0.579	0.191	0.759	**0.738**
R-tuning	0.444	0.138	0.750	0.682
ICFT (n)	0.734	0.632	0.750	0.591
ICFT (p)	0.750	0.658	**0.824**	0.723
ICFT (pn)	**0.757**	**0.691**	0.777	0.696
ICFT (w)	0.704	0.684	0.770	0.703

Table 3: Results of DR and CU. DR of "no context" measures answer precision of the LLM internally known questions.

Refusal method	Method Name	OQ	AQ		RQ	
		OAcc	MA	AF1	OR	RF1
0p10n						
Post refusal	Vanilla	0.437	0.145	0.167	0.770	0.570
	ICFT(n)	0.673	**0.098**	0.655	0.462	**0.690**
	ICFT(p)	**0.683**	0.240	**0.672**	**0.243**	0.682
Ours	Vanilla	0.523	0.104	0.240	0.282	0.590
	ICFT(n)	**0.729**	**0.059**	**0.707**	0.176	**0.731**
	ICFT(p)	0.697	0.178	0.691	**0.106**	0.698

Table 4: RALMs knowledge state aware refusal technique.

mance is not good as ICFT(n). This provides support for jointly considering active and passive refusals. We provide confidence distribution illustration Appendix D.

Retrieval handling of RIFT models Because a single calibration-error metric cannot fully reflect refusal quality, we introduce retrieval-handling metrics to further explain the results. Intuitively, a model that is more robust to noise is more likely to rely on its internal knowledge. While some methods (Zhang et al. 2025; Bi et al. 2025) explicitly emphasize the context faithfulness of RALMs. We evaluate these abilities using the denoising rate (DR) and the positive context utilization rate (CU), as reported in Table 3. In terms of denoising ability, all ICFT models perform better than the vanilla models, whereas the R-tuning models perform worse than the vanilla baseline. Although the R-tuning methods outperform the vanilla models in OAcc (OQ) and RF1 (RQ), the R-tuning approach appears to sacrifice the model's underlying knowledge competence in exchange for a stronger ability to articulate refusals, according to its worse DR performance in no context settings. In terms of context utilization, we find that ICFT(p) yields better results, while including negative context leads to worse performance in the all-positive (10p) setting. Surprisingly, however, all refusal fine-tuned models perform worse than the vanilla RALMs. This explains why these models perform poorly in scenarios with positive evidence: they tend to refuse internally unknown questions while ignoring the positive context.

Summary In this section, our results show that the over-refusal problem is mitigated by In-context fine-tuning, but magnified by R-tuning. The system's performance should be assessed by jointly considering the model's confidence, robustness, and context faithfulness. However, we also find that improved refusal performance does not necessarily imply better calibration or higher answer accuracy.

Mitigating the Over-Refusal Issue in RALMs (RQ3)

Although some refusal-aware RALM models do not support appropriate abstention by themselves, their confidence profiles can still distinguish correct refusals from incorrect ones. To validate whether we can distinguish different knowledge states and enable more appropriate refusals, we first study a simple threshold-based post-refusal technique. Concretely, we follow the thresholds-based refusal at inference stage.

To reduce the negative effects introduced by noisy contexts, we further develop a two-stage refusal technique. In the first stage, we apply a threshold T_s to U_{LLM} (the uncertainty of the base LLM) to detect whether the answer can be supported by internal knowledge, and a threshold on $\Delta U = U_{\text{RALM}} - U_{\text{LLM}}$ (where U_{RALM} is the uncertainty of the RALM, which incorporates context) to infer the knowledge state. In the second stage, we apply a refusal threshold in the same way as the baseline, but only when the RALM is classified as "unknown". All threshold values are selected via grid search on the development set. To better isolate the effect of knowledge on refusal, we compare these methods under an idealized but challenging (0p10n) context configuration. The results are summarized in Table 4. The post-refusal methods achieve higher overall accuracy than their counterparts in Table 2, but they also exhibit a substantially higher over-refusal rate. By first determining the knowledge state of the LLM itself, the model can choose when to rely on its own knowledge, yielding more calibrated confidence estimates and enabling further refusals without overusing harmful negative contexts, especially for ICFT(p) which show better calibration but less tendency to refuse on its own. Finally, we note that Wang et al. (2025) adopts similar information-gain-based method to detect context utility. This further supports our findings, while we provide a more explicit analysis of how knowledge states influence refusal behavior. Additional experiments of realistic RAG are provided in Appendix E.

Conclusions

In this work, we investigate whether RALMs "know when they don't know". We find that the calibration state of RALMs is greatly influenced by external contexts. In particular, we identify that purely negative contexts severely harm calibration and induce an over-refusal problem. We further study how the refusal quality of RALMs aligns with their calibration and observe that refusal-aware RALMs struggle to handle different RAG settings, due to entangled internal knowledge states and reduced context utilization. Finally, we combine the refusal ability of LLMs with post-refusal methods to balance overall response quality while mitigating over-refusal. Our study offers insights that underscore the need for improved calibration methods and the explicit modeling of dynamically evolving knowledge.

Acknowledgments

The authors thank all the reviewers for their suggestions and comments. This work is supported by National Natural Science Foundation of China (No.U21B2009). It is also supported by scholarship under the State Scholarship Fund and a visiting to Singapore Management University organized by the China Scholarship Council (CSC). The authors also acknowledge the material support by *Boston Meditech Group* and *Hangzhou Kangyi Health Management Limited Partnership*.

References

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.

Bi, B.; Huang, S.; Wang, Y.; Yang, T.; Zhang, Z.; Huang, H.; Mei, L.; Fang, J.; Li, Z.; Wei, F.; et al. 2025. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, 10280–10300.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17754–17762.

Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; He, Z.; Chen, K.; and Qiu, X. 2024. Can AI Assistants Know What They Don't Know? In *International Conference on Machine Learning*, 8184–8202. PMLR.

Cuconasu, F.; Trappolini, G.; Siciliano, F.; Filice, S.; Campagnano, C.; Maarek, Y.; Tonellotto, N.; and Silvestri, F. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 719–729.

Deng, Y.; Li, M.; Pang, L.; Zhang, W.; and Lam, W. 2025. Unveiling Knowledge Boundary of Large Language Models for Trustworthy Information Access. In *SIGIR 2025*, 4086–4089. ACM.

Deng, Y.; Zhao, Y.; Li, M.; Ng, S. K.; and Chua, T.-S. 2024. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13652–13673.

Fang, F.; Bai, Y.; Ni, S.; Yang, M.; Chen, X.; and Xu, R. 2024. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. In *ACL (1)*.

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14664–14690.

Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; and Herzig, J. 2024. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7765–7784.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 6577–6595. Association for Computational Linguistics.

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

Jeong, S.; Baek, J.; Cho, S.; Hwang, S. J.; and Park, J. C. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7029–7043.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; Showk, S. E.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022. Language Models (Mostly) Know What They Know. *CoRR*, abs/2207.05221.

Kapoor, S.; Gruver, N.; Roberts, M.; Collins, K.; Pal, A.; Bhatt, U.; Weller, A.; Dooley, S.; Goldblum, M.; and Wilson, A. G. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37: 85932–85972.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Lee, Z. P.; Lin, A.; and Tan, C. 2025. FinetuneRAG: Fine-Tuning Language Models to Resist Hallucination in Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.10792*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Li, M.; Zhao, Y.; Zhang, W.; Li, S.; Xie, W.; Ng, S.; Chua, T.; and Deng, Y. 2025. Knowledge Boundary of Large Language Models: A Survey. In *ACL 2025*.

Li, Z.; Hu, X.; Liu, A.; Zheng, K.; Huang, S.; and Xiong, H. 2024. Refiner: Restructure Retrieved Content Efficiently to Advance Question-Answering Capabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8548–8572.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Trans. Mach. Learn. Res.*, 2024.

Lyu, Q.; Shridhar, K.; Malaviya, C.; Zhang, L.; Elazar, Y.; Tandon, N.; Apidianaki, M.; Sachan, M.; and Callison-Burch, C. 2025a. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 19260–19268.

Lyu, Y.; Li, Z.; Niu, S.; Xiong, F.; Tang, B.; Wang, W.; Wu, H.; Liu, H.; Xu, T.; and Chen, E. 2025b. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2): 1–32.

Mei, Z.; Zhang, C.; Yin, T.; Lidard, J.; Shorinwa, O.; and Majumdar, A. 2025. Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know? *arXiv preprint arXiv:2506.18183*.

Moskvoretskii, V.; Marina, M.; Salnikov, M.; Ivanov, N.; Pletenev, S.; Galimzianova, D.; Krayko, N.; Konovalov, V.; Nikishina, I.; and Panchenko, A. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6355–6384.

Park, S.-I.; and Lee, J.-Y. 2024. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12: 1686–1702.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11: 1316–1331.

Soudani, H.; Zamani, H.; and Hasibi, F. 2025. Uncertainty Quantification for Retrieval-Augmented Reasoning. *arXiv preprint arXiv:2510.11483*.

Su, W.; Tang, Y.; Ai, Q.; Wu, Z.; and Liu, Y. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12991–13013.

Sun, X.; Xie, J.; Chen, Z.; Liu, Q.; Wu, S.; Chen, Y.; Song, B.; Wang, Z.; Wang, W.; and Wang, L. 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)*, 11461–11480.

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

Wang, Z.; Liang, Z.; Shao, Z.; Ma, Y.; Dai, H.; Chen, B.; Mao, L.; Lei, C.; Ding, Y.; and Li, H. 2025. InfoGain-RAG: Boosting Retrieval-Augmented Generation through Document Information Gain-based Reranking and Filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 7201–7215.

Wei, Z.; Chen, W.-L.; and Meng, Y. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*.

Wen, B.; Yao, J.; Feng, S.; Xu, C.; Tsvetkov, Y.; Howe, B.; and Wang, L. L. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13: 529–556.

Xu, F.; Shi, W.; and Choi, E. 2024. RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X.-J. 2023. Do Large Language Models Know What They Don't Know? In *Findings of the Association for Computational Linguistics: ACL 2023*, 8653–8665.

Yoran, O.; Wolfson, T.; Ram, O.; and Berant, J. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*.

Zhang, H.; Diao, S.; Lin, Y.; Fung, Y.; Lian, Q.; Wang, X.; Chen, Y.; Ji, H.; and Zhang, T. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7106–7132.

Zhang, Q.; Xiang, Z.; Xiao, Y.; Wang, L.; Li, J.; Wang, X.; and Su, J. 2025. FaithfulRAG: Fact-Level Conflict Modeling for Context-Faithful Retrieval-Augmented Generation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 21863–21882. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhu, R.; Ma, Z.; Wu, J.; Gao, J.; Wang, J.; Lin, D.; and He, C. 2025. Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 26157–26165.

Zhu, X.; Panigrahi, A.; and Arora, S. 2025. On the power of context-enhanced learning in llms. *arXiv preprint arXiv:2503.01821*.