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

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

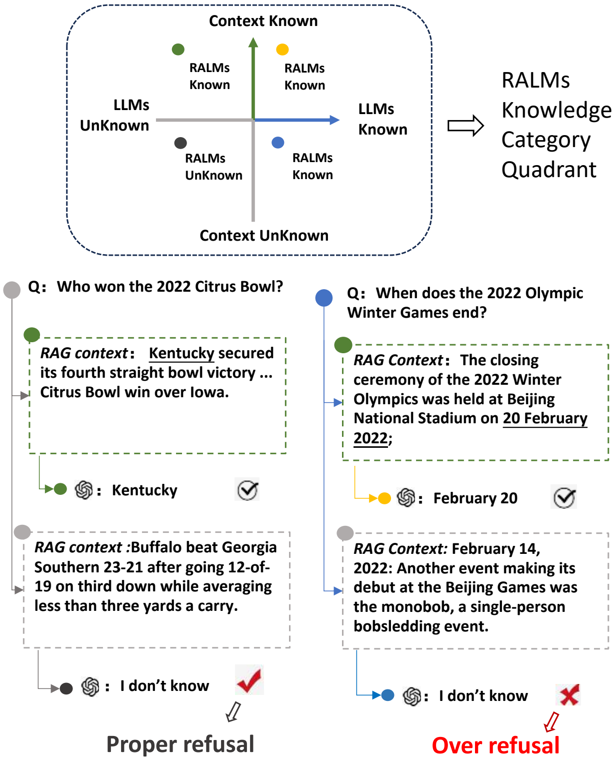
[1]Beijing Institute of Technology
[2]Singapore Management University
{ yczhou, hhy63, lyc2024, ruidai, wangxinglin, zxc2024, bjssm }@bit.edu.cn, ydeng@smu.edu.sg

Abstract

Existing Large Language Models (LLMs) occasionally generate plausible yet factually incorrect responses, known as hallucinations. Researchers are primarily using two approaches to mitigate hallucinations, namely Retrieval Augmented Language Models (RALMs) and refusal post-training. However, current research predominantly emphasizes their individual effectiveness while overlooking the evaluation of the refusal capability of RALMs. In this study, we ask the fundamental question: Do RALMs know when they don't know? Specifically, we ask three questions. First, are RALMs well-calibrated regarding different internal and external knowledge states? We examine the influence of various factors. Contrary to expectations, we find that LLMs exhibit significant **over-refusal** behavior. Then, how does refusal post-training affect the over-refusal issue? We investigate the Refusal-aware Instruction Tuning and In-Context Fine-tuning methods. Our results show that the over-refusal problem is mitigated by In-context fine-tuning. but magnified by R-tuning. However, we also find that the refusal ability may conflict with the quality of the answer. Finally, we develop a simple yet effective refusal method for refusal post-trained models to improve their overall answer quality in terms of refusal and correct answers. Our study provides a more comprehensive understanding of the influence of important factors on RALM systems.

Introduction

Existing Large Language Models (LLMs) have demonstrated remarkable performance across various challenging tasks. However, they occasionally generate plausible yet factually incorrect responses, a phenomenon commonly known as hallucination (Lewis et al. 2020; Huang et al. 2025). Prior research primarily addresses the hallucination issue in LLMs using 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 references, allowing retrieval-augmented language models (RALMs) to respond accurately to queries outside their internal knowledge. In contrast, refusal post-training aims to enhance the model's self-awareness, enabling it to proactively abstain from responding when uncertain.



Q: Who won the 2022 Citrus Bowl?

RAG context: Kentucky secured its fourth straight bowl victory ... Citrus Bowl win over Iowa.

: Kentucky

RAG context :Buffalo beat Georgia Southern 23-21 after going 12-of-19 on third down while averaging less than three yards a carry.

: I don't know

Q: When does the 2022 Olympic Winter Games end?

RAG Context: The closing ceremony of the 2022 Winter Olympics was held at Beijing National Stadium on 20 February 2022;

: February 20

RAG Context: February 14, 2022: Another event making its debut at the Beijing Games was the monobob, a single-person bobsledding event.

: I don't know

Figure 1: An example of the RALMs knowledge boundary and the corresponding answer correctness.

Although both methods are widely adopted, current research predominantly emphasizes their individual effectiveness while overlooking the evaluation of refusal capability within RALMs. Specifically, refusal post-training techniques typically focus on boosting the LLM's confidence in its internal knowledge and its ability to articulate internal uncertainty. However, 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, we can divide the knowledge state of RALMs into four quadrants based on prior research (Sun et al. 2025). When RALMs face the question "*When does the 2022 Olympic Winter Game end?*", they can answer it correctly. However, when retrieved contexts contain misinformation, RALMs may confuse and refuse to answer. Currently, we don't know when this phenomenon occurs. To address this gap, we investigate how retrieved context influences the refusal capabilities of LLMs.

Specifically, in this work, as ask three critical research questions (RQs). First, *are retrieval-augmented language models well-calibrated regarding different internal and external knowledge states? (RQ1)* Ideally, if RALMs are well-calibrated (know when they don't known), they can refuse to answer or we can abstain question properly. We categorize knowledge states as shown in Figure 1 and apply uncertainty estimation techniques to quantify the knowledge confidence of RALMs. Contrary to expectations, we find that LLMs exhibit significant **over-refusal** behavior and decline in confidence, particularly when confronted with exclusively irrelevant contexts. However, we also find models demonstrate improved calibration when supportive context exists within negative context. This indicates a vulnerability to contextual distractions, impairing their ability to effectively distinguish between internal and external knowledge states.

As recent studies (Zhang et al. 2025; Zhu et al. 2025) primarily elicit the refusal capability of large language models through post-training. Then, given the over-refusal tendency observed in RALMs, we further examine the impact of refusal post-training on this behavior. Accordingly, we pose the second research question: *how does refusal post-training affect the over-refusal issue? (RQ2)* We implement 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). To provide a comprehensive evaluation, we assess these models using a broader set of metrics, including context utility and refusal confidence. Our results show **the over-refusal problem is mitigated by ICFT while magnified by R-tuning**. In addition, we also find that the refusal ability conflicts with the answer correctness when positive context exists, due to a degradation of context utility. This limits the reliability and practicality of existing refusal post-training methods for RALMs.

Lastly, considering the difficulty to balance proper refusal and context utilization based solely on LLMs themselves, we propose the third research question: *can we combine the refusal-aware RALMs and confidence-based answer abstention to mitigate over-refusal? (RQ3)* According to previous findings, we can leverage confidence and its variation to determine the knowledge state of RALMs. Then we can choose to answer a question with or without context, or further abstain.

Our contributions are as follow: 1) We investigate the confidence calibration of RALMs. We comprehensively examine important factors and identify the over-refusal problems. 2) We explore whether the current refusal instruction tuning will intensify LLMs' over-refusal and give further explanations. 3) We design a better refusal technique based for RALMs based on the above findings.

Related Works

Knowledge Boundary of LLMs. Identifying the knowledge boundary helps to recognize the limitations of knowledge. It is also referred to as "knowing what you don't know,"(Yin et al. 2023; Deng et al. 2024) a crucial aspect in determining their practical applicability. Li et al. (2024) formally category knowledge boundary according to prompt and model sensitivity. Gekhman et al. (2024) divide the knowledge type according to hallucination variation after knowledge finetuning. However, they mainly focus on the LLM's internal knowledge. Hallucination happens when users' requests are out of LLMs' knowledge boundary (Huang et al. 2025). The primary approach to mitigate hallucination includes retrieval-augmented generation (RAG). RAG (Lewis et al. 2020) is a convenient method for inference stage where the retrieved context fills the knowledge gap. recent RAG including adapt retrieval model for LLMs(Xu, Shi, and Choi 2024), external knowledge injection (Ovadia et al. 2024) and dynamic RAG (Su et al. 2024). Dynamic RAG leverage uncertainty estimation to determine when to retrieve. However, they ignore the refusal ability of RALMs and assume RALMs are well calibrated. The difference is they focusing on the dynamic generation process while we focus on the static knowledge state.

Refusal Method of LLMs. Refusal behavior are mainly researched in post-training stages (Wen et al. 2025). It mainly considers the instruction tuning (Kapoor et al. 2024) and refusal alignments (Cheng et al. 2024; Sun et al. 2025). The former build static (Zhang et al. 2024) or dynamic (Zhu et al. 2025) refusal data according to unknwon knowledge of LLMs. The latter get refusal preference response, where a refusal is better than an answer for unknown question. They mainly focus on elicit the self-awareness of knowledge boundary and express by themselves. Another refusal method conducts at the inference stage (Feng et al. 2024). When user find the LLM be uncertain, they directly abstain the answer. As this method also assume RALMs are well-calibrated, the uncertainty estimation are the most important parts. However, most study (Zhao et al. 2024; Du et al. 2024) focus on the vanilla LLMs situation.

Uncertainty Estimation. It is crucial for LLMs to recognize their limitation and express their confidence when responding to users (Yin et al. 2023). Current research deems the uncertainty and confidence as antonyms (Lin, Trivedi, and Sun 2024). Namely the more the uncertain LLM, the less confidence the LLM is. Geng et al. (2024) divide uncertainty estimation(UE) of LLMs into white-box and black-box methods.The white-box suit for the open-sources LLMs where we can get the internal state (Kadavath et al. 2022). While the black-box methods only use the response for UE, thus have wider applications. In our work, we leverage black-box UE methods as we could potential leverage close-source LLMs to improve refusal quality.

Prelinminary

We briefly describe the concept of proper refusal and over refusal. According to (Feng et al. 2024), the questions could be divide into ”should refuse” and ”should an-

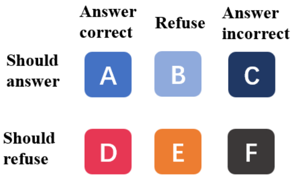


Figure 2: Refusal and answer confusion matrix.

swer". If LLMs tend to give false answer, then it should refuse the question, which means LLM do not entail the knowledge. We follow (Gekhman et al. 2024) to divide the internal knowledge into four categories, namely "highlyknown","maybeknown","weaklyknown" and "unknown". The LLMs obtain highlyknown knowledge which they can answer correctly with various prompt and greedy decoding. This means LLMs are confident to answer correctly. While unknown knowledge are those LLM can't answer correctly even once under temperature sampling. In our study, we deem "highly known" and "maybe known" to be "should answer" and "weakly known" and "unknown" to be "should reject". We illustrate the LLM response and the ground truth circumstance in Figure 2. Notice that the "C" and "D" part exists in our settings. In terms of "C" part, the "should answer" questions is tested on greedy sampling, and can mistake especially on "maybeknown" questions. In terms of "D", the weaklyknown part can be answer correctly ; especially for "unkown" questions, when LLM with even negative context will response with true answer. Thus the proper refusal is $\frac{E}{D+E+F}$ and the over-refusal is $\frac{B}{A+B+C}$. Besides, another metrics called mis-answer is $\frac{C}{A+B+C}$.

Methods

Uncertainty Estimation Methods

We mainly use the black-box uncertainty methods to evaluate confidence of the response of LLMs, because they are more universally applicable in previous research. They can be divided into the following three categories.

Verbalized based UE. This kind of methods leverage the LLM's self-awareness and express ability, which directly give the confidence of the corresponding answer. Following previous research, we design 4 different prompts following (Tian et al.) to express the confidence of answers. Details are in Appendix.

Consistency based UE. This kind of method has believe that more consistent of a answer reflects a higher belief of a models. Lyu et al. (2025a) use different approach to measure the uncertainty of LLMs and applied to decoding methods like self-consistency methods. We formalize three kinds of measurement calculation as follows. For a give input x and the correspond model $M(\cdot)$, we generate the multiple response $\{r_1, r_2, ..., r_m\}$ and decide a final answer by majority voting $\bar{r} = argmax \sum_{i=1}^{n} \mathbb{1}(r_i = r)$. The first measurement called **agreement** is calculated as below:

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

The second measurement $Ent(a)$ is entropy-based, which rescale the weights of each answer, it calculated as below:

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

where $\hat{r}$ is the duplicated answer set and the p_i is the probability of the unique answer r_i. The final measurement called FSD 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 kinds of methods consider the similarity of all responses. We use two feature of the similarity matrix following (Lin, Trivedi, and Sun 2024):

The first is U_{Eigv} ,which leverages the eigen values of Laplace matrix L to represent the uncertainty. Then We formalize the confidence C_{Eigv} as follows:

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \tag{4}$$

$$U_{Eigv} = \sum_{k=1}^{m} \max(0, 1 - \lambda_k), \tag{5}$$

$$C_{Eigv} = 1 - \frac{U_{Eigv} - 1}{m - 1}, \tag{6}$$

where W is the similarity matrix of responses and D is the degree matrix of W. λ_k is the eigen vectors of L, where $\lambda_k \in [0, +\infty)$. We normalize the value to $[0, 1]$ to match the scope with the above methods.

The second is C_{deg}, which uses the degree matrix. A node with high degree is well-connected to other nodes, suggesting that it lies in a confident region of the LLMs. The definition is as follows:

$$U_{deg}(x) = \frac{trace(mI - D)}{m^2} \tag{7}$$

$$C_{deg}(x) = 1 - U_{deg}, \tag{8}$$

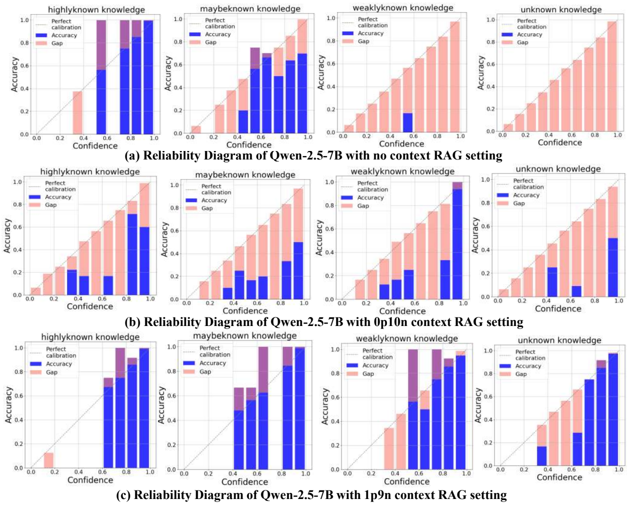
where $trace(mI - D) \in [0, m^2]$. The U_{deg} can be interpreted as the average pairwise distance.

Refusal Instruction Tuning

We mainly use two Refusal Instruction Tuning (RIFT) methods, including R-tuning and In-Context Fine-Tuning.

R-tuning. R-tuning (Zhang et al. 2024) is a simple but effective method to teach LLMs to express refusal. The workflow typically includes two stages. The first stage is to identify the knowledge boundary of LLMs, i.e. to find the knowledge that an LLM can't answer correctly. The second stage is to set the refusal expressions for those unknown questions and leverage the Instruction Tuning to enable LLMs to express "I don't known".

In-Context Fine-Tuning. (Zhu, Panigrahi, and Arora 2025; Lee, Lin, and Tan 2025) finds that inserting the positive context into prompt for Instruction Tuning will benefits the LLMs' accuracy. However, they generally append positive context and train to generate correct answer. In our work, we extend this method for refusal scenario. We insert not only the positive context, but also some negative context. We set answers of training samples to correct answer or refusal according to the knowledge category quadrant of RALMs as Figure 3. More details are in Appendix.



(a) Reliability Diagram of Qwen-2.5-7B with no context RAG setting

(b) Reliability Diagram of Qwen-2.5-7B with 0p10n context RAG setting

(c) Reliability Diagram of Qwen-2.5-7B with 1p9n context RAG setting

Figure 3: The reliability diagram under different internal and external knowledge states. We find that the negative documents lead to significant uncalibrated and over confidence, especially for highly known knowledge. The positive documents leads to under confidence.

Evaluation Experiments

We first give the experimental setup. Then we investigate the calibration error of RALMs to answer RQ1. Finally we investigate the calibration of Refusal Instruction-Tuning (RIFT) effect on RALMs to answer RQ2.

Experimental Setup

Datasets We explore the refusal ability in tasks like QA and RAG. In this work, we mainly use three prevalent 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). NQ and CURD are large scale QA/RAG datasets suitable for both training and test. RGB is a dataset particular developed for test, including refusal ability of RALMs. Details are described in Appendix A.

RALM models We adopt two prevalent families of open-source LLMs, including Qwen and LLaMA. Though the current LLMs are multilingual, we find that Qwen-2.5 models perform far less effectively than LLaMA-3.1 on English datasets and so do LLaMA on Chineses datasets. To better utilize the knowledge of LLMs, we test Qwen[1] model on Chinese datasets and LLaMA [2] on English datasets. For retrieval parts, we perform hybrid search and rerank to include high quality negatives. This considers both semantic and lexical similarity. Details are described in Appendix B.

LLM sampling hyper-parameters All the generation sampling temperature is set to 0.5 following (Lyu et al. 2025a). The sampling times are set to 16. Other generation hyper-parameters are set to default for corresponding LLMs.

Answer judgment We apply a strict answer decision workflow following (Sun et al. 2025), including LLM-as-judge, Exact Match and reject words filter. Details are described in Appendix C.

Evaluation metrics Evaluation metrics include accuracy-based and confidence calibration metrics (Feng et al. 2024; Sun et al. 2025). We briefly describe them as follows:

- **Accuracy-based Metrics**: The answer ability of RALMs is multi-dimension, reflecting the answer quality, refusal quality and retrieval handling ability.
  - Answer Quality (AQ): Including the answer precision (Pre), recall (Rec) and F1 of the correct answer and mis-answer rate (MR).
  - Refusal Quality (RQ): We measure the refusal precision (RPrec), recall (RRec) and F1(RF1). Moreover, we measure the refusal rate (RR), and especially the over-refusal rate (OR).
  - Overall Quality (OQ): We test the overall accuracy (OAcc) which includes the ratio of proper refusal and correctly answers.
  - Retrieval Handling (RH): Including the context utilization rate (CU) and the denoise rate (DR), which measure the ability to utilize positive information and ignore noisy retrieval correspondingly.
- **Confidence Calibration Metrics**: We mainly use **Brier Score** to measure whether the answer confidence measure the answers. In addition, we test the refusal confidence error.

More details are in Appendix D.

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

In this section, we address RQ1. We specifically design different RAG settings with various positive and negative context combinations.

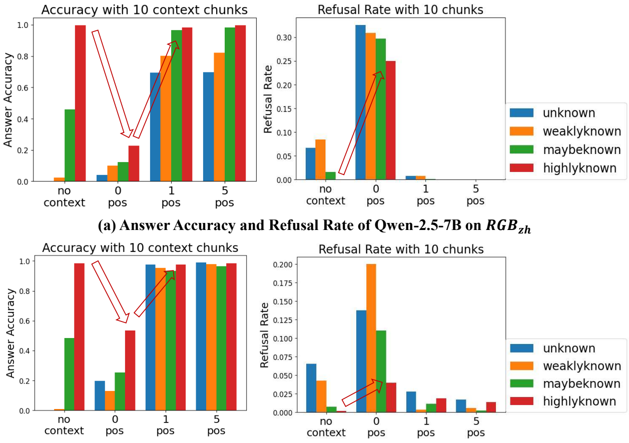
Calibration error of RALMs We first use different uncertainty estimation methods to measure the confidence of RALMs. Then we evaluate the calibration error with Brier score. Notice we exclude those refusal results, as we find the recent LLMs can actively refuse to answer. Results are in Table 1. Though the calibration error varies under different RAG settings, the RALMs become extremely well-calibrated when positive documents exist, especially for verbalize and agreement-based UE methods. This indicates that the UE methods are also suitable for RALMs settings. As the consistency-based methods perform best among generally, we take the results for further explanation. We find that when no positive context exits (10n), the calibration error becomes worse. And when we insert 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 do not significantly change. This means that RALMs can sensitively perceive the availability of knowledge. We provide more context settings in "Appendix E" to

[1] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[2] https://github.com/meta-llama/llama3

UE type	UE name	RGB_{en} no context	RGB_{en} 0p10n	RGB_{en} 1p9n	RGB_{en} 5p5n	RGB_{en} 1p19n	RGB_{zh} no context	RGB_{zh} 0p10n	RGB_{zh} 1p9n	RGB_{zh} 5p5n	RGB_{zh} 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
	Agreement	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
SimilarityMatrix-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 documents and B negative documents for RAG context settings.



(a) Answer Accuracy and Refusal Rate of Qwen-2.5-7B on RGB_{zh}

(b) Answer Accuracy and Refusal Rate of LLaMA-3.1-8B on RGB_{en}

Figure 4: The answer accuracy and refusal rate vary according to the internal and external knowledge states. The negative documents leads to significant decrease of accuracy on highly known knowledge and increase of refusal. The phenomenon is referred to as over-refusal. However, even a single positive document can lead to significant increase of answer accuracy and eliminate over-refusal.

further demonstrate this findings. We systematically investigate how prompt variants, positive context position, context quality and quantity affect the model performance. As we find the key factor is the positive context existence, the following settings use 10 context as default.

Over-confident or under-confident As the original LLMs are known to be over-confident (Li et al. 2024), we further investigate the phenomenon for RALMs. We utilize the reliable diagram to illustrate the situations. Results are shown in Figure 3. Under the no context setting, the highlyknown knowledge are under-confident while the others are over-confident. The highly known type have high confidence, while the others are dispersed. However, under the all negative context setting, the RALMs become over-confident and the confidence scores also dispersed. Surprisingly, the weakly known knowledge achieve better accuracy where negative contexts have unexpected effects. Even when one positive context exists, RALMs are under-confident except for the unknown knowledge. This indicates that the accuracy and confidence are greatly influenced by noisy contexts.

Accuracy and confidence of RALMs The calibration error measures the gap between model accuracy and confidence, We first investigate the answer accuracy and the refusal rate. In terms of accuracy, the answer accuracy decreases for highly known and maybe known questions, while increase for weakly and unknown questions. While positive context exists, the accuracy significantly increases, especially for unknown and weakly known knowledge. This indicates that the negative context extremely harm the answer quality. In terms of refusal rate, we observe an significant increase on all the knowledge types. However, considering the LLMs can correctly answer highly known questions, refusal on those questions are not correct. We identify this phenomenon as **over-refusal**, which are not observed in previously research.

Summary In this section, we empirically reveal that RALMs genrally “know they don’t know” under no context or positive context. However, become over-confident when confronted with negative context and may over refusal the knowledge they actually know.

Does Refusal Instruction Tuning Enhance the Refusal Quality of RALMs? (RQ2)

In this section, we address the RQ2. We test the R-tuning and In-context Fine-tuning variants. Considering the knowledge quadrants of Figure 1, we set four In-context Fine-tuning variants as follows:

- ICFT(n) : We append the negative contexts for LLMs. The answer of training samples depend on the internal state of LLMs. If internal knowledge type is ”known”, the answer is original ground truth;else the answer is ”I don’t known”.
- ICFT(p) : We append the 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 useful context and we want to test how negative context influence the RALMs in training stages.

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

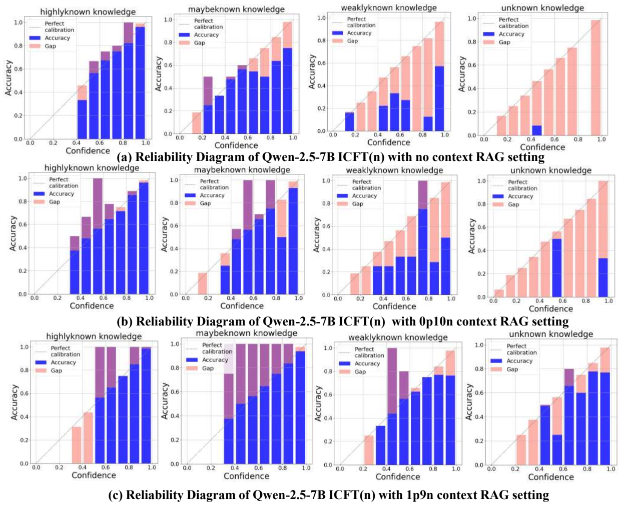
Method name	DR		CU	
	no context	10n	10p0n	1p9n
Vanilla	0.579	0.191	0.759	**0.738**
R-tuning	0.444	0.138	0.750	0.682
ICFT (n)	0.734	0.632	0.750	0.591
ICFT (p)	0.750	0.658	**0.824**	0.723
ICFT (pn)	**0.757**	**0.691**	0.777	0.696
ICFT (w)	0.704	0.684	0.770	0.703

Table 3: Results of denoise rate and context utilization.

- ICFT(w): We include both the ICFT(n) and ICFT(pn) training samples.

The evaluation metics are calculated following sections . Moreover, we use the the same query, only different context and answers to ensure the training fairness.

Response quality of RIFT models The response quality of refusal trained RALMs is in multi-dimension, including the overall quality (OQ), answer quality (AQ) and refusal quality (RQ). As shown in Table 2, we observe the model performance vary according to RALMs settings. In the no context settings, the ICFT(n) perform best in terms of OAcc(OQ) and ICFT(p) for F1(AQ). The R-tuning model perform best in terms of RF1(RQ), and the ICFT(n) the second. This may because R-tuning scenario matches with testing, which has higher refusal rate (RR) and moderate refusal precision (RPre). However, the **over-refusal rate (OR) increases**, which means R-tuning potentially harm the self-awareness. The decrease of answer precision(Pre), and increase of mis-answer rate(MA) could support this findings. We will validate the conference change in the following part.



(a) Reliability Diagram of Qwen-2.5-7B ICFT(n) with no context RAG setting

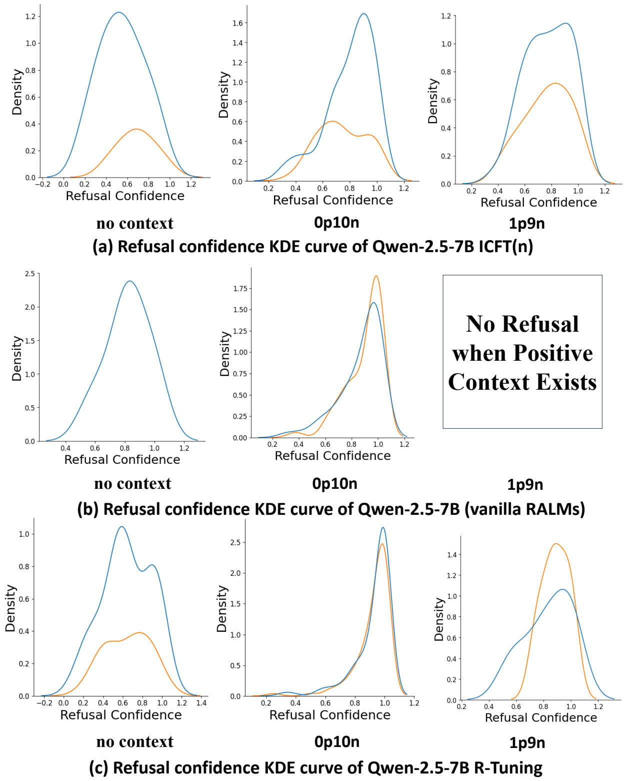
(b) Reliability Diagram of Qwen-2.5-7B ICFT(n) with 0p10n context RAG setting

(c) Reliability Diagram of Qwen-2.5-7B ICFT(n) with 1p9n context RAG setting

Figure 5: The reliability diagram of ICFT models under different RALMs context settings.

In the all negative (0p10n) settings, the ICFT(n) perform far better than other models in terms of OAcc(OQ), F1(OQ), RF1(RQ). Though the **over-refusal rate (OR)** of R-tuning is the worst, the ICFT(n) improve the phenomena and perform better than the vanilla RALMs. Moreover, we find ICFT methods with positive context significantly eliminate the over refusal phenomena and ensure a moderate OAcc(OQ).

Surprisingly, when the positive context exists, the vanilla RALMs perform best in terms of OAcc(OQ) and F1(AQ). The refusal quality(RQ) is not suitable here, as the RALMs can directly answer according to context. The metrics only consider the internal knowledge of RALMs. In terms of RQ, the ICFT(n) actually performs the worst. Though it achieve



no context 0p10n 1p9n

(a) Refusal confidence KDE curve of Qwen-2.5-7B ICFT(n)

no context 0p10n 1p9n

(b) Refusal confidence KDE curve of Qwen-2.5-7B (vanilla RALMs)

no context 0p10n 1p9n

(c) Refusal confidence KDE curve of Qwen-2.5-7B R-Tuning

Figure 6: The KDE curve of refusal confidence of different RIFT models. The blue and orange curve is for correct and incorrect refusal correspondingly.

the highest RF1(RQ) and it refuse nearly half of the internal unknown questions.

Retrieval handling of RIFT models Intuitively,ICFT(n) will enhance the denoise ability of RALMs, as models are trained to ignore the negative context to answer questions when internal known and external unknown. Besides, ICFT(p) will enhances the context utilization ability as models are trained to utilize context when they internal unknown. We check these ability as shown in Table 3. In terms of denoise ability, all ICFT models are better than vanilla models. However, the R-tuning methods are worse than vanilla models. Though the R-tuning methods perform better than vanilla model in terms of OAcc(OQ) and RF1(RQ), which indicates R-tuning mainly encourage models to refuse according internal states. In terms of context utilization, we find that the ICFT(p) leads to better results, while including negative context lead to worse result in all positive context (10p) settings. However, we surprisingly find all the refusal finetuned model perform worse than vanilla RALMs. The ICFT(n) perform even worse than R-tuning models. This well explains why these models perform poorly in scenarios with positive samples. As they tend to refuse the internal unknown questions while ignore the positive context.

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.

Refusal Confidence of RIFT models Considering the over refusal differs for R-tuning and ICFT, we examine their distribution of refusal confidence. As shown in Figure 6, the best performed ICFT(n) can distinguish the correct abstain while the worst R-tuning methods have large overlap between correct and incorrect abstain. More illustrations of answer confidence are in Appendix F.

Summary In this section, our results show that the over-refusal problem is mitigated by In-context fine-tuning. but magnified by R-tuning. However, we also find that the refusal ability may conflict with the quality of the answer.

Mitigating the Over-refusal Issue in RALMs (RQ3)

We try to leverage the confidence of LLMs to improve the refusal ability of RALMs, especially to mitigate the over-refusal problems. Post-refusal techniques for QA is simple: evaluate the answer confidence and set the refusal threshold, we refuse the answer once the confidence is low. This only considers the internal knowledge of LLMs. We further develop a simple refusal technique. It needs to detect internal and external knowledge state first, then determine the context usage and further post refusal. Results are shown in Table 4. The post refusal methods achieve higher overall acc compared to counterpart in Table 2, while get achieve much higher over-refusal rate. As the uncalibrated state of RALMs under all negative context, the correspond searched threshold is ”0.58”. This results in a high refusal rate including both uncertain known answer and unknown answer as shown in Figure 6. However, by first determining the knowledge state of LLM itself, LLMs can choose to leverage its own knowledge. We can get a more calibrated confidence and perform further refusal and avoid to use the harmful negative contexts. Real RAG testing details are in appendix G.

Conclusions

In this work, we mainly investigate whether RALMs know they don’t know, i.e. whether they are calibrated. We find the calibration state of RALMs greatly influenced by the external contexts. We mainly identify the exclusive negative context strongly harm the calibration and cause over-refusal problem. We further investigate how refusal instruction tuning affect calibration and refusal quality of RALMs. We find the refusal-aware RALMs are hard to manage different RAG settings, due to the confused internal state and decline of context utilization. Finally, we combine refusal ability of LLMs and post refusal method to balance overall quality while mitigate over-refusal problem.

References

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*, volume 38, 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.; 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.

Du, K.; Snæbjarnarson, V.; Stoehr, N.; White, J.; Schein, A.; and Cotterell, R. 2024. Context versus Prior Knowledge in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13211–13235.

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.

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. Finetune-RAG: 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.-K.; Chua, T.-S.; and Deng, Y. 2024. Knowledge boundary of large language models: A survey. *arXiv preprint arXiv:2412.12472*.

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*, volume 39, 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.

Ovadia, O.; Brief, M.; Mishaeli, M.; and Elisha, O. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 237–250.

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.

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, W.; Wang, Z.; and Wang, L. 2025. Divide-Then-Align: Honest Alignment based on the Knowledge Boundary of RAG. *arXiv preprint arXiv:2505.20871*.

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

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.

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, R.; Xu, Y.; Xiao, Y.; Zhu, R.; Jiang, X.; Chu, X.; Zhao, J.; and Wang, Y. 2025. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25895–25903.

Zhao, Y.; Yan, L.; Sun, W.; Xing, G.; Meng, C.; Wang, S.; Cheng, Z.; Ren, Z.; and Yin, D. 2024. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7044–7056.

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*, volume 39, 26157–26165.

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

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) NA

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) NA

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/-partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/-partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) partial

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) partial