Divide-Then-Align: Honest Alignment based on the Knowledge Boundary of RAG

Anonymous ACL submission

Abstract

Large language models (LLMs) augmented with retrieval systems have significantly advanced natural language processing tasks by integrating external knowledge sources, enabling more accurate and contextually rich responses. To improve the robustness of such systems against noisy retrievals, Retrieval-Augmented Fine-Tuning (RAFT) has emerged as a widely adopted method. However, RAFT conditions models to generate answers even in the absence of reliable knowledge. This behavior undermines their reliability in high-stakes domains, where acknowledging uncertainty is critical. To address this issue, we propose Divide-Then-Align (`DTA`), a post-training approach designed to endow RAG systems with the ability to respond with "I don't know" when the query is out of the knowledge boundary of both the retrieved passages and the model's internal knowledge. `DTA` divides data samples into four knowledge quadrants and constructs tailored preference data for each quadrant, resulting in a curated dataset for Direct Preference Optimization (DPO). Experimental results on three benchmark datasets demonstrate that `DTA` effectively balances accuracy with appropriate abstention, enhancing the reliability and trustworthiness of retrieval-augmented systems. Code is available at: https://anonymous.4open.science/r/Divide-Then-Align

1 Introduction

Large language models (LLMs) have achieved remarkable success across various NLP tasks (Radford et al., 2019; Brown et al., 2020; Bubeck et al., 2023; OpenAI, 2022). However, these models are constrained by their pretraining knowledge, which may become outdated or insufficient for domain-specific queries (Jiang et al., 2023; Shuster et al., 2021). Retrieval-Augmented Generation (RAG) (Izacard and Grave, 2021; Lewis et al., 2020) addresses this limitation by combining LLMs with

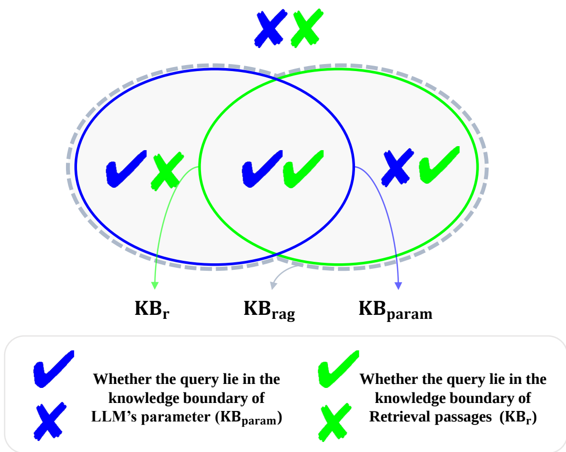

Figure 1: Knowledge Boundary of RAG. A query can be divided into four quadrants based on the model's parametric knowledge boundary ($\mathrm{KB}_{\mathrm{param}}$) and the knowledge boundary of the retrieval passages (KB_r). The queries that fall into ✗✗ should be answered with "I don't know" instead of generating potentially hallucinatory answers.

retrieval systems that access external knowledge sources (Pasca, 2019; Jin et al., 2019) to provide more accurate and contextually rich responses.

Despite its promise, RAG faces significant challenges due to the limitations of current retrieval systems. In practice, retrieval systems often fail to return entirely accurate passages, resulting in noisy contexts that can contain irrelevant, conflicting, or misleading information (Yoran et al., 2024; Fang et al., 2024; Cuconasu et al., 2024). Yoran et al. (2024); Fang et al. (2024); Liu et al. (2024b) propose Retrieval-Augmented Fine-Tuning (RAFT) to mitigate this issue, which involves fine-tuning LLMs with a combination of retrieved contexts, both relevant and noisy, encouraging the models to learn robustness to noisy inputs.

While RAFT has shown improvements in model performance, it introduces a critical drawback: **RAFT conditions the model to answer questions even when the retrieved contexts are entirely**

noisy. This behavior poses a significant risk for deploying LLMs in real-world applications, particularly in high-stakes domains like medical (Raja et al., 2024), legal (Reji et al., 2024), and financial (Yepes et al., 2024) fields. As shown in Figure 1, the knowledge boundary of RAG systems is the union of the model's parametric knowledge boundary and the retrieval knowledge boundary. **When faced with queries for which neither the model's parametric knowledge contains sufficient information to answer the query (✗), nor can useful information be found in the retrieved passages (✗), an ideal LLM should respond with "I don't know" instead of generating potentially hallucinatory answers**. However, our experiments reveal that RAFT models do not have this critical ability. Even when explicitly prompted to respond with "I don't know". In such scenarios, the models tend to overfit to the training paradigm and generate hallucinatory answers.

To address this limitation, we propose Divide-Then-Align (`DTA`), a systematic post-training approach to enhance RAFT models. `DTA` operates in two key stages: ❶ **Divide**: First, we divide data samples from three benchmark datasets (Natural Questions, TriviaQA, and WebQuestions) into four quadrants based on whether the answers lie within the LLM's parametric knowledge boundary and the retrieval knowledge boundary. This division is crucial as different knowledge quadrants require distinct strategies for preference data construction. ❷ **Align**: For each category, we carefully construct preference data by specifying appropriate chosen and rejected responses based on the knowledge boundary division. This results in a curated training set of 10,000 preference samples. We then employ Direct Preference Optimization (DPO) (Rafailov et al., 2024) to endow the model with the ability to acknowledge uncertainty with "I don't know" responses while maintaining the high accuracy achieved through RAFT training. To rigorously evaluate our approach, we develop **a comprehensive knowledge quadrants based evaluation framework with nine metrics** that assess both the model's overall performance and its ability to abstain from answering when queries fall outside both knowledge boundaries. Through careful analysis across different quadrants, we demonstrate the effectiveness of our approach in balancing accuracy with principled abstention behavior.

Our contributions can be summarized as follows:

❶ **Problem Identification**: We first divide the RAG samples into four quadrants based on whether the answers lie within the LLM's parametric knowledge boundary and the retrieval knowledge boundary. And we find that the RAFT model is not able to abstain from answering when the rag sample is out of both the LLM's parametric knowledge boundary and the retrieval knowledge boundary.

❷ **Proposed Solution**: We propose `DTA`, a systematic approach that constructs quadrant-specific preference data (10,000 samples) and leverages DPO to enable principled abstention behavior while preserving model performance.

❸ **Experimental Validation**: We evaluate our method on three widely used datasets, demonstrating its effectiveness in improving model reliability and trustworthiness.

2 Preliminary

2.1 Knowledge Boundary of RAG

Let $\mathcal{D}$ denote the knowledge corpus. Let $r : \mathcal{Q} \rightarrow \mathcal{P}$ be the retrieval function that maps a query q to relevant passages $P \subseteq \mathcal{D}$, where $\mathcal{Q}$ is the query space and $\mathcal{P}$ is the passage space. We use $M : \mathcal{Q} \times \mathcal{P} \rightarrow \mathcal{A}$ to represent the LLM function that takes both the query and passages as input and generates an answer from the answer space $\mathcal{A}$. Let golden : $\mathcal{Q} \rightarrow \mathcal{A}$ be the function that maps a query to its ground truth answer, which represents the correct response that should be generated for the query. Let $C(M(q, P))$ denote the correctness evaluation function.

For honest alignment of RAG systems, it's crucial to determine whether a query q lies within or outside the system's knowledge boundary $\mathrm{KB}_{\mathrm{rag}}$. Ideally:

- If q $\in \mathrm{KB}_{\mathrm{rag}}$, the model should generate the correct answer other than IDK.
- If q $\notin \mathrm{KB}_{\mathrm{rag}}$, the model should abstain from answering.

2.2 Knowledge Quadrants

To better evaluate the knowledge boundary of RAG systems, we consider that $\mathrm{KB}_{\mathrm{rag}}$ is composed of two fundamental components: the parametric knowledge boundary of the LLM ($\mathrm{KB}_{\mathrm{param}}$) and the knowledge boundary of the retrieval passages (KB_r). Formally:

$$\mathrm{KB}_{\mathrm{param}} = \{q \in \mathcal{Q} \mid C(M(q, \emptyset)) = \mathrm{True}\} \quad (1)$$

$$\mathrm{KB}_r = \{q \in \mathcal{Q} \mid \exists p \in r(q) : \mathrm{contains}(p, \mathrm{golden}(q)) = \mathrm{True}\} \quad (2)$$

The overall knowledge boundary of the RAG system can be characterized as:

$$\mathrm{KB}_{\mathrm{rag}} = \mathrm{KB}_{\mathrm{param}} \cup \mathrm{KB}_r$$

This formulation captures that a query can be answered correctly if it falls within either the model's parametric knowledge or can be answered using retrieved information.

Then we can divide the samples into quadrants based on $\mathrm{KB}_{\mathrm{param}}$ and KB_r:

✔✔ : $q \in \mathrm{KB}_{\mathrm{param}} \cap \mathrm{KB}_r$

✔✘ : $q \in \mathrm{KB}_{\mathrm{param}} \setminus \mathrm{KB}_r$

✘✔ : $q \in \mathrm{KB}_r \setminus \mathrm{KB}_{\mathrm{param}}$

✘✘ : $q \notin \mathrm{KB}_{\mathrm{param}} \cup \mathrm{KB}_r$

The details of the description of the four quadrants can be found in the Appendix A.

3 Methodology

3.1 Knowledge Quadrants Division

To divide queries into the four knowledge quadrants defined in Section 2, we need to determine whether a query q belongs to $\mathrm{KB}_{\mathrm{param}}$ and/or KB_r. We use three widely-used question answering datasets: Natural Questions (Kwiatkowski et al., 2019a), TriviaQA (Joshi et al., 2017a), and WebQuestions (Berant et al., 2013a).

Determining $q \in \mathrm{KB}_{\mathrm{param}}$ To determine whether a query lies within the model's parametric knowledge boundary ($q \in \mathrm{KB}_{\mathrm{param}}$), we sample N answers $\{a_1, ..., a_N\}$ from the model without any retrieved context by evaluating $C(M(q, \emptyset))$ with different random seeds. If the proportion of correct answers in these N samples exceeds a threshold

$$\delta = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[C(a_i) = \mathrm{True}] > \delta$$

we consider $q \in \mathrm{KB}_{\mathrm{param}}$ (✔). Otherwise, we consider $q \notin \mathrm{KB}_{\mathrm{param}}$ (✘).

To determine whether a response is correct, we directly using lexical matching, which checks whether the golden answers appear in the responses gnerated by the model. According to the results shown in (Wang et al., 2024), applying lexical matching yields a consisitency rate of approximately 90% when compared to human evaluation. Therefore, we deem the lexical matching to be a good enough way to determine whether the response is correct.

Determining $q \in \mathrm{KB}_r$ To determine whether a query lies within the retrieval knowledge boundary ($q \in \mathrm{KB}_r$), we use GPT-4o (gpt-4o-2024-08-06) to evaluate whether the retrieved passages contain or directly imply the correct answer. We prompt GPT-4o with a specialized evaluation prompt (see Appendix F) that returns a binary score indicating whether the context sufficiently supports the answer. If GPT-4o determines the context contains or implies the correct answer (score = 1), we consider $q \in \mathrm{KB}_r$ (✔). Otherwise, we consider $q \notin \mathrm{KB}_r$ (✘).

3.2 Preference Data Construction

Based on the knowledge quadrants, we construct preference data for each quadrant as follows:

For ✔✔, we can directly use the ground truth as the chosen response and use IDK as the rejected response.

For ✔✘ samples, we select the ground truth as the chosen response, while constructing two types of rejected responses: (1) incorrect answers generated by the LLM when exposed to noisy context, demonstrating the model's vulnerability to noisy information; and (2) "I don't know" responses, which are overly conservative given the model's inherent knowledge.

For ✘✔ samples, the ground truth serves as the chosen response, paired with three categories of rejected responses: (1) incorrect answers resulting from the model's failure to utilize the golden information in the context; (2) incorrect answers generated by the LLM without any context to suppress the wrong parametric knowledge; and (3) "I don't know" responses, which indicate an inability to leverage available context.

For ✘✘ samples, where neither source contains reliable information, we designate "I don't know" as the chosen response. The rejected responses comprise: (1) incorrect answers generated by the LLM without any context, (2) incorrect answers

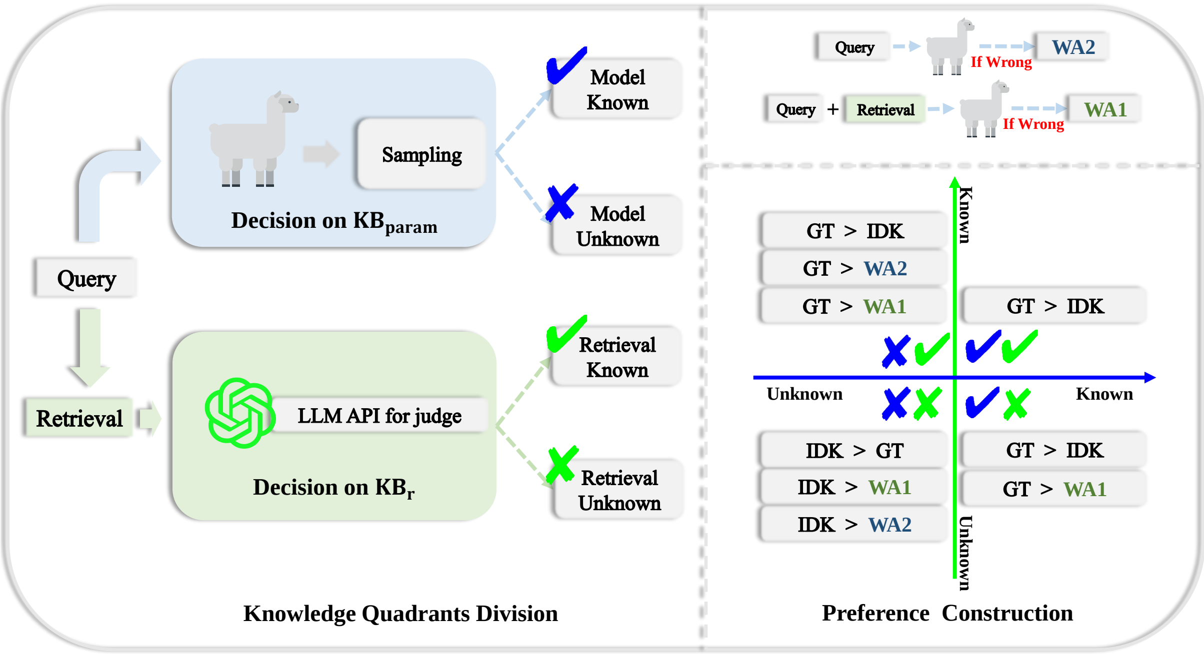

Figure 2: The pipeline of knowledge quadrants division and preference dataset construction. GT denotes the ground truth answer; IDK represents "I don't know" response; WA1 and WA2 are wrong answers generated by the LLM (WA = Wrong Answer); "If Wrong" indicates the condition where the model generates an incorrect response. The symbol ">" indicates a preference relationship where the left option is preferred over the right option. The preference construction (right) shows how different response types (GT, IDK, WA1, WA2) are ranked based on the knowledge quadrant the query falls into. $\mathrm{KB}_{\mathrm{param}}$ means the LLM's parametric knowledge boundary and KB_r means the retrieval knowledge boundary.

generated by the LLM with noisy context, and (3) the ground truth itself, as generating correct answers without supporting evidence may encourage unfounded speculation.

I don't know Response Our refusal to answer template is:

> This question is beyond the scope of my knowledge and the references. I don't know the answer.

We use "I don't know" to refer to this template in the paper.

3.3 Post training using DPO

In this section, we introduce how to post-train the RAFT model to enable it with the ability to abstain from answering.

After the preference data is constructed, we employ a multi-objective training approach combining three different losses.

DPO Loss We utilize the standard DPO loss to learn from preference pairs of chosen and rejected responses. This helps the model learn to distinguish between preferred and non-preferred outputs. Given a chosen response y_c and a rejected response y_r for a query q and retrieved context $r(q)$, the DPO loss is defined as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma(\tau(r_\theta(q, r(q), y_c) - r_\theta(q, r(q), y_r))) \quad (3)$$

where $r_\theta(q, r(q), y)$ represents the log probability of generating response y given query q and retrieved context $r(q)$ under the model parameters θ, τ is the temperature parameter, and σ is the sigmoid function. Note that this reward score is derived from the same language model being trained, eliminating the need for a separate reward model.

SFT Loss Our empirical observations show that DPO training tends to focus on reducing rejected response rewards rather than improving the quality of the chosen response. To address this limitation, we incorporate supervised fine-tuning loss on the chosen responses to explicitly enhance the model's ability to generate preferred outputs:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p_\theta(y_c^t | q, r(q), y_c^{<t}) \quad (4)$$

where y_c^t represents the t-th token of the chosen response, and T is the length of the response.

Knowledge Quadrant Classification Loss We add a value head on top of the last token's hidden state to predict which knowledge quadrant (0-3) a query belongs to. This classification task serves as an auxiliary objective that helps the model develop better awareness of its knowledge boundaries and improve its ability to determine when to abstain from answering. The classification loss is defined as:

$$\mathcal{L}_{\text{class}} = -\sum_{k=0}^{3} y_k \log p_\theta(k \mid q) \tag{5}$$

where y_k is the one-hot encoded ground truth label for the knowledge quadrant, and $p_\theta(k|q)$ is the predicted probability for quadrant k.

The final training objective is a weighted combination of these three losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DPO}} + \beta\mathcal{L}_{\text{SFT}} + \gamma\mathcal{L}_{\text{class}}, \tag{6}$$

where β, and γ are hyperparameters controlling the contribution of each loss component.

4 Experiments

4.1 Datasets

We evaluate our approach on three standard open-domain question answering datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019b), TriviaQA (Joshi et al., 2017b), and WebQuestions (WebQ) (Berant et al., 2013b). For each dataset, we follow the setting of (Fang et al., 2024) and employ the retrieval model DPR (Karpukhin et al., 2020) as our retriever, which retrieves 3 passages from wikipedia for each query.

To evaluate the model's ability to make appropriate abstentions, we also divide each sample in the test sets into four quadrants based on knowledge boundaries(✔✔, ✔✗, ✗✔, ✗✗). The details the test set are presented in Appendix D.

4.2 Baselines

We evaluate our approach against three categories of baselines: (1) RAFT models that focus on handling retrieval noise (**RAAT** (Fang et al., 2024), **Ret-Robust** (Yoran et al., 2024), **ChatQA** (Liu et al., 2024b)), (2) calibration-based methods that detect potential hallucinations (**P(True)** (Kadavath et al., 2022), **Logits** (Guerreiro et al., 2023)) and (3) two widely-used baselines like in-context learning (**ICL** (Wei et al., 2022)) and **self-Consistency** (Wang et al., 2022). Details of these baselines can be found in Appendix C and G.2.

4.3 Evaluation Metrics

To systematically evaluate the performance of our method, we propose a comprehensive evaluation framework based on the knowledge quadrant division. The framework consists of four main aspects: Overall Quality (OQ), Answer Quality (AQ), Retrieval Handling (RH), and Abstention Quality (AbQ). Across these aspects, we define 9 distinct metrics that thoroughly assess different dimensions of model performance. The details and formulations of these metrics are presented in Table 1.

4.4 Main Results

Main experimental results are shown in Table 2. Our post-training strategy `DTA` achieves the best performance on three llama architectures. Notably, it achieves Acc (64.1, 64.8, 65.5), F1 (64.6, 66.6 65.8), AF1(63.3, 59.9, 64.7), surpassing baseline methods by significant margins. Critically, `DTA` uniquely balances robust answer generation with precise abstention, addressing a key limitation of existing approaches.

While RAFT variants (RAAT, Ret-Robust, ChatQA) can improve answer quality of base model, they uniformly fail to abstain properly. As designed, RAFT models effectively enhance the model answer quality. In addition, following its training approach, RAAT did a good job of using golden contexts to generate correct answers. Ret-Robust can resist the most noisy retrieval and generate high-quality responses using model's knowledge. However, they all struggle with abstain quality. In both RAAT and Ret-Robust, none of the test queries can be abstained. ChatQA has the ability to refrain from some queries, but the quality is far from satisfactory. Post-hoc techniques, including two calibration methods (P(true), Logits) and consistency, are applied to RAFT models to enhance abstain quality but impair the ability to use model knowledge. And their answer quality is also affected, which is not good for the overall performance. ICL only improves the abstain quality when the RAFT model has the ability to abstain, but the improvement is not significant.

In stark contrast, `DTA` achieves highest AF1 without compromising answer quality. DTA did this by structurally aligning model behavior with knowledge boundaries, enabling reliable and self-aware QA systems. However, our method falls short in terms of the DR and CUR metrics, which is related to the trade-off with abstention. When appropri-

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

Table 1: Evaluation Metrics based on the knowledge quadrant division. Let ✔ denote correct answers, ✘ denote incorrect answers, and ○ denote abstentions ("I don't know" responses). For any category (e.g., ✔✘), $\lvert ✔ \cap ✔✘ \rvert$ represents the count of correct answers within the ✔✘ category.

ately enhancing the model's abstention capability to promote the growth of overall quality, a significant portion of the ✔✘ and ✘✔ data is also rejected. On the contrary, a significant reduction in the proportion of ✘✘ during training leads to a notable surge in both DR and CUR scores. Further discussion is shown in hyperparamter experiments.

An interesting observation is that the Original LLM achieves remarkably high DR scores. While RAFT models are specifically trained to utilize context and rely more heavily on retrieved passages for generating answers, recent research (Tan et al., 2024; Bi et al., 2024) suggests that base models tend to prioritize their parametric knowledge while being less dependent on provided context. Since all contexts in the DR category are noisy, excessive reliance on context would only lead to degraded performance.

To better understand the impact of knowledge quadrant division, we conducted experiments using single knowledge boundaries (KB_r or $\mathrm{KB}_{\mathrm{param}}$) instead of the full quadrant approach. For these experiments, we used ground truth answers when queries fell within the knowledge boundary and abstention responses when queries fell outside it, while keeping all other hyperparameters identical to `DTA`. As shown in Table 4, using single knowledge boundaries led to notably worse performance across metrics, demonstrating the importance of our fine-grained quadrant-based approach for properly modeling RAG system knowledge boundaries.

4.5 Ablation Study

We conducted comprehensive ablation experiments to analyze the contribution of each component in our `DTA` framework based on the `DTA` results of RAAT. The results in Table 3 demonstrate the importance of each component from multiple aspects:

Training Objectives Without DPO loss, the model shows significantly degraded performance in answer quality (Rec drops from 63.7% to 38.8%) while maintaining high abstention rates (ARec: 78.6%). However, the abstain precision decreases substantially from 61.7% to 43.1%. This indicates that although the RAG system learns to abstain, it becomes overly cautious and lacks confidence in answering queries that it should be able to handle. Without SFT loss, the model exhibits a dramatic decline in overall quality (Acc drops from 63.7% to 38.8%) and severely degraded abstention quality (AF1 drops from 63.3% to 6.7%). These results validate our hypothesis that the SFT loss plays a crucial role in teaching the model how to make abstention. The removal of classification loss shows relatively minor impact across metrics, with slight decreases in both answer quality (F1 drops from 64.6% to 63.4%) and abstention quality (AF1 drops from 63.3% to 62.6%). This suggests that while knowledge quadrant classification serves as a helpful auxiliary task, it is not critical to the model's core capabilities.

Model Name	OQ	AQ			RH		AbQ		
	Acc	Rec	Prec	F1	DR	CUR	ARec	APrec	AF1
Llama-2-7b									
Original	42.2	64.1	42.2	50.9	**85.8**	49.9	0.00	0.00	0.00
RAAT	46.2	**70.2**	46.2	55.7	76.3	**61.7**	0.00	0.00	0.00
+ P(true)	45.0	65.0	46.0	53.8	68.9	57.4	6.71	32.1	11.0
+ Logits	49.2	58.8	50.5	54.3	69.8	47.0	30.9	45.1	36.6
+ Consistency	51.4	69.0	50.7	58.5	82.1	58.8	16.3	58.4	25.4
+ ICL	46.8	71.2	46.8	56.5	84.4	60.2	0.00	0.00	0.00
+ `DTA`	**64.1**	63.7	**65.5**	**64.6**	68.9	52.8	**65.0**	**61.7**	**63.3**
Llama-2-13b									
Original	48.1	66.3	48.1	55.8	82.1	40.7	0.00	0.00	0.00
Ret-Rrobust	51.6	71.0	51.6	59.8	**90.0**	44.5	0.00	0.00	0.00
+ P(true)	50.9	56.0	58.5	57.2	74.8	29.7	37.5	33.6	35.4
+ Logits	53.6	70.0	53.6	60.7	87.9	43.4	10.0	52.9	16.9
+ Consistency	53.9	**71.8**	54.0	61.7	89.6	**46.4**	6.30	52.5	11.2
+ ICL	52.0	71.6	52.0	60.3	89.1	46.6	0.00	0.00	0.00
+ `DTA`	**64.8**	67.9	**65.3**	**66.6**	76.8	45.5	**56.7**	**63.5**	**59.9**
Llama-3-8b									
Original	43.9	62.0	43.9	51.4	**76.0**	42.0	0.00	0.00	0.00
ChatQA	46.1	60.9	45.0	51.8	54.5	46.8	10.2	71.8	17.8
+ P(true)	50.1	45.2	55.6	49.9	46.2	29.1	61.9	42.6	50.5
+ Logits	46.6	57.8	46.8	51.7	51.0	44.8	19.3	44.9	27.0
+ Consistency	46.5	61.0	46.7	52.9	58.7	46.6	11.3	44.0	18.0
+ ICL	43.3	55.0	41.4	47.2	50.3	40.7	15.1	**75.4**	25.1
+ `DTA`	**65.5**	**64.5**	**67.2**	**65.8**	62.8	**48.9**	**67.9**	61.8	**64.7**

Table 2: Main results on the benchmark consisting of three datasets. OQ: Overall Quality (Acc: Accuracy); AQ: Answer Quality (Rec: Recall, Prec: Precision); RH: Retrieval Handling (DR: Denoise Rate, CUR: Context Utilization Rate); AbQ: Abstain Quality (ARec: Abstain Recall, APrec: Abstain Precision, AF1: Abstain F1).

Knowledge Boundary Components Removing ✔✔ samples from training leads to decreased performance across all metrics, particularly in context utilization (CUR drops to 43.9%), highlighting the importance of learning from samples where correct information is available in the context. Without ✔✘ samples, the model shows reduced ability to handle retrieved information (DR: 47.9%), indicating that exposure to noisy samples during training is crucial for developing robust retrieval handling capabilities. Without ✘✔ samples, the model shows an interesting trade-off: while the denoise rate (DR) improves to 72.1%, the context utilization rate (CUR) drops to 45.6%. This suggests that without training on samples where the model needs to rely on retrieved context, it becomes overly conservative with retrieval usage, preferring to rely on its parametric knowledge even when helpful context is available. This leads to degraded overall accuracy (58.6%), highlighting the importance of these samples for teaching the model when to effectively leverage retrieved information. Without ✘✘ samples, the model completely loses its abstention capability (AbQ metrics all 0.0) while showing artificially high recall (73.3%) and DR (84.5%), indicating that training with examples where abstention is appropriate is essential for developing proper abstention behavior.

Wrong Answer Types The impact of removing wrong answer types (w/o WA1, w/o WA2) reveals an interesting trade-off in model behavior. Without the suppression of wrong answers, the model becomes more inclined to generate responses rather than abstain, leading to higher recall (68.8% for

	OQ	AQ			RH		AbQ		
Model Name	Acc	Rec	Prec	F1	DR	CUR	ARec	APrec	AF1
`DTA`	64.1	63.7	65.5	64.6	68.9	52.8	65.0	61.7	63.3
w/o DPO	52.4	38.8	67.8	49.4	52.1	28.7	78.6	43.1	55.7
w/o SFT	37.1	54.6	36.5	43.8	58.9	45.2	3.50	76.6	6.7
w/o CLS	63.1	63.5	63.3	63.4	63.9	53.6	62.4	62.7	62.6
w/o ✔✔	57.0	54.6	59.0	56.7	57.1	43.9	61.5	53.9	57.5
w/o ✔✘	61.7	53.4	67.3	59.5	47.9	44.5	77.7	55.7	64.9
w/o ✘✔	58.6	58.5	59.8	59.1	72.1	45.6	58.7	56.5	57.6
w/o ✘✘	48.2	73.3	48.2	58.2	84.5	64.0	0.00	0.00	0.00
w/o WA1	61.8	68.8	59.0	63.5	75.3	58.8	48.4	71.2	57.6
w/o WA2	61.5	66.2	59.1	62.4	68.5	56.4	52.4	68.2	59.3
w/o WA1∪WA2	58.2	68.5	53.8	60.3	71.7	59.4	38.5	80.6	52.1

Table 3: Ablation results.

	OQ	AQ			RH		AbQ		
Knowledge Boundary	Acc	Rec	Prec	F1	DR	CUR	ARec	APrec	AF1
`DTA`	64.1	63.7	65.5	64.6	68.9	52.8	65.0	61.7	63.3
KB_r	58.9	49.4	62.9	55.3	43.4	41.7	77.3	54.7	64.1
$\mathrm{KB}_{\mathrm{param}}$	45.8	32.6	42.6	36.9	39.3	23.1	71.1	49.0	58.0

Table 4: Experimental results on different knowledge boundary.

w/o WA1, 66.2% for w/o WA2) and improved retrieval handling metrics. However, this increased response rate comes at the cost of precision, dropping from 65.5% to around 59%, as the total number of attempted answers grows significantly. The model's abstention capability is also compromised, with lower abstention recall but higher abstention precision, indicating more conservative use of "I don't know" responses. These results demonstrate that wrong answer samples play a crucial role in training by helping the model establish appropriate decision boundaries between answering and abstaining, ultimately contributing to better overall performance when both types are included.

4.6 Hyperparameter

Experiments are conducted on preference dataset size, multi-objective loss weights and IDK-ratio for the preference dataset. The experimental results are shown in Appendix E.

5 Conclusion

In this paper, we propose a novel framework for honest alignment of retrieval-augmented language models based on knowledge boundary quadrants. We first identify that the knowledge boundary of RAG systems consists of two fundamental components: the parametric knowledge boundary ($\mathrm{KB}_{\mathrm{param}}$) and the retrieval knowledge boundary (KB_r). Based on this insight, we divide RAG samples into four knowledge quadrants. To address the critical limitation of RAFT models regarding their inability to abstain from answering when queries fall outside both knowledge boundaries (✘✘), we construct a comprehensive preference dataset that captures the desired behavior for each quadrant. Using this dataset, we employ DPO training with a multi-objective approach combining DPO loss, SFT loss, and knowledge quadrant classification loss to align the model's behavior with the knowledge boundary constraints. Furthermore, we introduce a systematic evaluation framework with 9 metrics to assess both response quality and abstention capabilities. Experiments conducted on three benchmark datasets demonstrate that our approach effectively improves the model's ability to make appropriate abstention decisions while maintaining strong performance on answerable queries.

Limitations

While our work presents a promising approach for honest alignment of RAG systems, following limitations should be noted:

Knowledge Boundary Determination : Our method for determining whether a query belongs to $\mathrm{KB}_{\mathrm{param}}$ relies on sampling from the base model without context, which is used by a lot of previous works (Xu et al., 2024a; Cheng et al., 2024). However, this approach may not perfectly capture the true parametric knowledge boundary, as model performance can vary across different prompting strategies. And we think this is a potential research direction for future work.

Specific Domain : Our evaluation focuses on three general-domain open QA datasets (NQ, TriviaQA, WebQ). While these datasets provide a good foundation for testing, they may not fully represent the challenges and nuances specific to specialized domain applications. The effectiveness of our approach in highly specialized domains requires further investigation.

Ethical Considerations

Our work improves the refusal capability of RAG systems to reduce the risk of generating harmful or incorrect information. Nevertheless, the model may still produce low-quality or hallucinated responses, when faced with ambiguous or out-of-distribution queries. Additionally, since our model has not undergone safety alignment, it may still generate inappropriate content when faced with adversarial or malicious queries.

References

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013a. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013b. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. 2024. Context-dpo: Aligning language models for context-faithfulness. *arXiv preprint arXiv:2412.15280*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? In *Forty-first International Conference on Machine Learning*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 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*, pages 719–729.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 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)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang Zhang, and Lichao Sun. 2024. The best of both worlds: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *ArXiv preprint*, abs/2403.14403.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *ArXiv preprint*, abs/2401.04088.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017a. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017b. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019b. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Chatqa: Surpassing gpt-4 on conversational qa and rag. In *NeurIPS*.

Meta-AI. 2024. Llama 3 model card.

OpenAI. 2022. Introducing ChatGPT.

Marius Pasca. 2019. Wikipedia as a resource for text analysis and retrieval. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 24, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Mahimai Raja, E Yuvaraajan, et al. 2024. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1128–1133. IEEE.

Sneha Ann Reji, Reshma Sheik, A Sharon, Avisha Rai, and S Jaya Nirmala. 2024. Enhancing llm performance on legal textual entailment with few-shot cot-based rag. In *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pages 1–6. IEEE.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2024. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *arXiv preprint arXiv:2409.11242*.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. “knowing when you don’t know”: A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Hongshen Xu, Zichen Zhu, Da Ma, Situo Zhang, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *arXiv preprint arXiv:2403.18349*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024b. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

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

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *arXiv preprint arXiv:2402.05131*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024b. RAFT: Adapting language model to domain specific RAG. In *COLM*.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-reflection (FaR) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8702–8718.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A The Details of the Knowledge Quadrants

✔✔ represents the most ideal but trivial scenario, where both the model's parametric knowledge and retrieved passages contain the correct information.

✔✘ occurs when $q \in \mathrm{KB}_{\mathrm{param}}$ but $q \notin \mathrm{KB}_r$, indicating that while the model has the necessary parametric knowledge, the retriever fails to find relevant passages. In such cases, retrieval is unnecessary and the model should rely on its parametric knowledge. Many adaptive RAG methods (Jeong et al., 2024; Asai et al., 2024) focus on identifying and handling this scenario.

✘✔ represents the core scenario that RAG systems are designed to handle, where $q \in \mathrm{KB}_r$ but $q \notin \mathrm{KB}_{\mathrm{param}}$. Here, while the model lacks the necessary parametric knowledge, the retrieved passages contain the correct information. However, even with the correct information present in the retrieved passages, the model may fail to utilize it effectively due to issues such as "lost in the middle" (Liu et al., 2024a).

RAFT acctually enhances the RAG system's answer accuracy across both ✔✘ and ✘✔ scenarios by addressing their distinct challenges: For ✔✘: RAFT teaches the model to rely on its parametric knowledge when retrieved passages are noisy. For ✘✔: RAFT helps the model better utilize information from retrieved passages. So the RAFT get some emprical success in a some previous work (Fang et al., 2024; Yoran et al., 2024; Zhang et al., 2024b; Liu et al., 2024b).

In the ✘✘ case ($q \notin \mathrm{KB}_{\mathrm{param}} \cup \mathrm{KB}_r$), neither the model's parametric knowledge nor the retrieved passages contain the correct information. In such case, the model should ideally abstain from answering to maintain faithfulness and avoid hallucination. However, current RAFT-trained models are conditioned to always generate an answer, even when the query is out of $\mathrm{KB}_{\mathrm{rag}}$. This leads to an overly aggressive response pattern that prioritizes answer generation over honesty, potentially producing misleading or entirely fabricated responses. While RAFT approaches may improve surface-level metrics like answer accuracy, it fundamentally compromises the system's reliability and trustworthiness. In this work, we specifically focus on addressing this critical gap by developing methods that enable models to recognize when a query falls outside of $\mathrm{KB}_{\mathrm{rag}}$ and appropriately respond with "I don't know". This capability is essential for deploying RAG systems in high-stakes applications where the cost of hallucination and misinformation can be severe.

B Related works

B.1 Retrieval-Augmented Generation

RAG (Lewis et al., 2020; Borgeaud et al., 2022; Izacard and Grave, 2021; Zhang et al., 2024b) is a widely adopted paradigm for augmenting large language models (LLMs) with external knowledge.

By integrating a retrieval system, RAG enables models to access and utilize external knowledge sources during generation, overcoming the limitations of static, parameterized knowledge in LLMs. This approach has shown significant promise in tasks requiring factual accuracy, domain-specific knowledge (Zhang et al., 2024b), and up-to-date information (Li et al., 2024). Despite its advantages, the effectiveness of RAG heavily depends on the quality of the retrieved passages. Current retrieval systems often fail to guarantee complete relevance, introducing noisy contexts into the generation process. To address this challenge, Retrieval-Augmented Fine-Tuning (RAFT) (Zhang et al., 2024b; Fang et al., 2024; Liu et al., 2024b) has been proposed. RAFT fine-tunes models with a mixture of retrieved contexts, including both clean and noisy passages, encouraging robustness to imperfect retrieval results.

However, RAFT-trained models exhibit a critical limitation: they are conditioned to answer queries even when provided with entirely noisy contexts. This over-reliance on retrieved information increases the risk of generating hallucinated or misleading responses, especially in high-stakes applications. Our work builds on this understanding by addressing the overlooked issue of enabling RAFT-trained models to acknowledge uncertainty and respond with "I don't know" when appropriate.

B.2 Honest Alignment in Large Language Models

Honesty is a foundational principle in aligning large language models (LLMs) with human values. It requires models to accurately express their knowledge, recognize their limitations, and avoid misleading users when uncertain. Honesty encompasses two critical components: self-knowledge and self-expression. Self-Knowledge refers to the model's ability to discern what it knows and doesn't know, enabling it to explicitly admit uncertainty (e.g., responding "I don't know") when necessary. This capability is crucial for mitigating hallucinations and ensuring model reliability in high-stakes applications. Current methods to improve self-knowledge include: Training-free approaches: These leverage predictive probabilities (Duan et al., 2024), prompting strategies (Zhou et al., 2023; Kadavath et al., 2022; Zhao et al., 2024) (e.g., Chain-of-Thought reasoning), and sampling/aggregation techniques to elicit calibrated confidence from models (Tian et al., 2023; Guo et al., 2017; Xiong et al., 2024). While effective in some contexts, these approaches often struggle with free-form generation and require significant computational overhead. Training-based approaches: Methods such as supervised fine-tuning and reinforcement learning aim to teach models to abstain from answering uncertain queries or provide confidence scores alongside responses (Yang et al., 2023; Zhang et al., 2024a; Jiang et al., 2024; Zhou et al., 2023; Gao et al., 2024; Xu et al., 2024b; Stengel-Eskin et al., 2024). However, these works only consider the LLM's parametric knowledge boundary, and ignore the knowledge boundary of the retrieval system.

Our work builds on these foundations, endowing the retrieval-augmented models with the ability to acknowledge uncertainty under noisy contexts based on the preference training on the four knowledge quadrants.

Comparison with the existing works: Most of the current raft work (Yoran et al., 2024; Fang et al., 2024; Liu et al., 2024b) and rag work (Asai et al., 2023; Lewis et al., 2020) try to improve the model's ability on the accuracy of response and ignore the faithfulness of the response. And we have shown that the success of the current raft work is built on the sacrifice of the faithfulness of the response. The model actually becomes an aggressively omniscient model. Cheng et al. (2024); Feng et al. (2024); Xu et al. (2024a) align the model to abstain when the model can not answer the query. These work actually only focus on the knowledge boundary of the LLM itself. But in the RAG scenario, the knowledge boundary is actually the combination of the LLM knowledge boundary and the retrieval knowledge boundary. Song et al. (2024); Thakur et al. (2024) align the model to refuse answer when the retrieved passages are noisy. But they ignore the knowledge boundary of the LLM itself. **Our work is the first work that simultaneously considers the knowledge boundary of the LLM itself and the retrieval knowledge boundary** and aligns the model to refuse answer only when the query is out of the both knowledge boundaries.

C Baseline Methods

We compare our approach against several state-of-the-art baselines and corresponding Llama family base models.

Base Models:

- **Llama2-7B** (Touvron et al., 2023): A member of Llama2 family with 7 billion parameters, which is released in July 2023.

- **Llama2-13B** (Touvron et al., 2023): A member of Llama2 family with 13 billion parameters, which is released in July 2023.

- **Llama3-8B** (Meta-AI, 2024): A member of Llama3 family with 8 billion parameters, which is released in April 2024.

RAFT Models:

- **RAAT** (Fang et al., 2024): A model that employs adaptive adversarial training to handle three types of retrieval noises (relevant, irrelevant, and counterfactual). During training, it dynamically selects the most challenging noise type based on the model's current performance and uses multi-task learning to enhance noise awareness.

- **Ret-Robust** (Yoran et al., 2024): A model that trains with a mixture of relevant and irrelevant retrieved contexts. For each training example, it retrieves either top-1, low-ranked, or random passages with equal probability to teach the model when to use or ignore retrieved information.

- **ChatQA** (Liu et al., 2024b): A two-stage instruction tuning approach that outperforms GPT-4 on retrieval-augmented generation and conversational QA tasks. It first performs supervised fine-tuning to enhance basic instruction following capabilities, then conducts context-enhanced instruction tuning specifically for dialogue QA and RAG tasks.

Calibration Methods: These methods use post-hoc techniques to predict whether the retrieved passages are relevant to the question or if the model is likely to hallucinate, which can trigger a refusal to answer.

- **P(True)** (Kadavath et al., 2022): Uses prompt-based evaluation to assess the correctness of model generations, leveraging the observation that LLMs are relatively well-calibrated in self-evaluation tasks.

- **Logits**: Implements various methods from previous studies (Guerreiro et al., 2023; Kadavath et al., 2022; Varshney et al., 2023; Huang et al., 2023) that aggregate output token probabilities or logits to score LLM confidence for error detection.

Dataset	✔✔	✔✘	✘✔	✘✘
LLaMA-2-7B				
NQ	204	40	2,125	1,241
TriviaQA	2,225	1,109	4,391	3,588
WebQ	202	76	882	872
LLaMA-2-13B				
NQ	451	105	1,877	1,172
TriviaQA	3,669	1,978	2,809	2,652
WebQ	258	105	826	843
LLaMA-3-8B				
NQ	442	122	1,887	1,159
TriviaQA	3,229	1,721	3,387	2,976
WebQ	224	94	860	854

Table 5: Statistics of the test set across different model architectures and datasets. The columns show the distribution of samples across the four knowledge quadrants.

We also include two widely-used baseline approaches:

- **ICL**: We implement in-context learning using a prompt template with three carefully curated demonstration examples: one showing appropriate abstention for out-of-knowledge-boundary queries, and two showcasing correct answer generation for in-boundary queries. This balanced demonstration set helps the model learn both when to answer and when to abstain.

- **Consistency** (Wang et al., 2022): Uses the consistency of the model's responses to determine whether it should abstain from answering.

D Dataset Statistics

We determine whether a query belongs to the LLM's parametric knowledge ($\mathrm{KB}_{\mathrm{param}}$) based on the performance of vanilla model (LLaMA-2-7b, etal.), and evaluate retrieval knowledge (KB_r) based on whether the top-3 retrieved passages contain the correct answer. This division approach allows us to analyze both the RAFT model's improvements over the base model across different knowledge quadrants and its abstention capabilities.

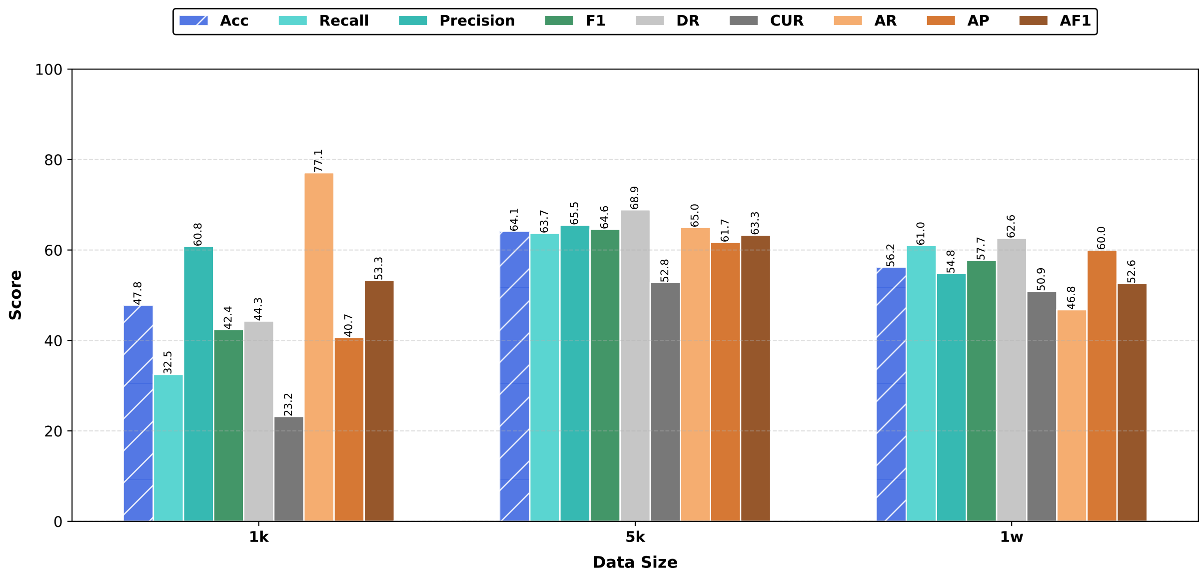

Figure 3: Experiments across DPO data size. (IDK ratio=0.7, loss weights β=1.0, γ=0.5)

After division, we randomly select 3000 queries from three datasets to evaluate all methods.

To balance the model's ability to answer questions and abstain when appropriate, we introduce a hyperparameter called IDK-ratio, which controls the proportion of training examples where the preferred response is "I don't know" (IDK). Specifically, IDK-ratio determines the fraction of ✗✗ samples in the training set. Importantly, we maintain the natural distribution of queries across all four quadrants in the test set without any manipulation, ensuring evaluation reflects real-world conditions and provides a more generalizable assessment of model performance.

Table 5 shows the distribution of test queries across the four knowledge quadrants. A substantial portion of queries fall into the ✗✗ quadrant. This represents a critical scenario where models should abstain from answering, yet traditional RAFT approaches force a response. The distribution highlights why defining $\mathrm{KB}_{\mathrm{rag}}$ through the combination of both $\mathrm{KB}_{\mathrm{param}}$ and KB_r is crucial. Relying solely on KB_r (Liu et al., 2024b; Song et al., 2024) would incorrectly exclude ✓✗ queries from the model's knowledge boundary (for example, 1,978 TriviaQA queries for LLaMA-2-13B where the model has parametric knowledge). Similarly, using only $\mathrm{KB}_{\mathrm{param}}$ (Cheng et al., 2024; Feng et al., 2024; Xu et al., 2024a) would mistakenly omit ✗✓ queries (such as 2,125 NQ queries for LLaMA-2-7B) that RAG systems can effectively handle through retrieval. Our dual-boundary approach enables more precise identification of true knowledge gaps (✗✗ cases) where abstention is warranted, while allowing optimal knowledge source selection in other cases.

E Hyper-parameter experiments

Multi-Objective Loss Adjusting the weights of the multi-objective loss significantly impacts model's overall quality. As shown in Figure 4, increasing the weight of the SFT loss generally leads to steady improvements, which is in line with our hypothesis. The experiments confirm that SFT effectively assists in aligning with the chosen data, demonstrating strong auxiliary alignment effects. Meanwhile, the classification loss (CLS) is not without merit; it plays a critical role when combined with the SFT loss, achieving optimal performance within the weight range of 0.5 to 0.7. This highlights the synergistic interplay between the two loss components under balanced configurations.

Data Size Statistics in Figure 3 show that 5k DPO preference data achieves competitive performance in terms of overall quality(OQ Acc), answer quality(AQ F1), and abstain quality(AbQ F1). Reducing data to 20% sharply degrades the outcomes, which indicates the significance of sufficient training data. However, when data size grows to 10k, it seems increased noise-potentially introduced by scaling without rigorous quality control-lead to performance degradation. This pattern emphasizes the importance of the quality of data in preference optimization.

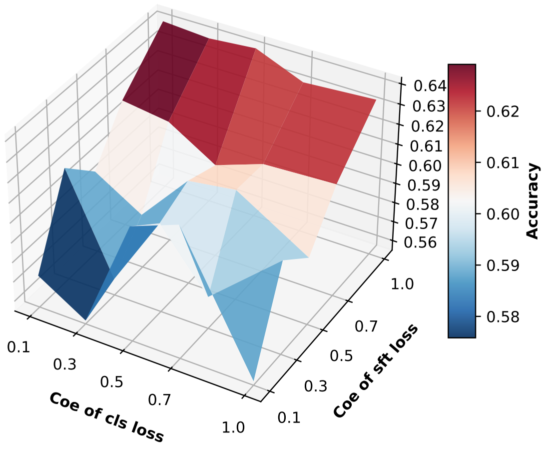

Figure 4: Experiments across multi-objective loss weights. (DPO data size=5k, IDK ratio=0.7)

IDK Ratio Varying the ratio of IDK-labeled data reveals a nuanced and interesting trade-off. Higher ratios (0.1-0.7) intuitively improve AbQ F1 as the model learns to master the ability to abstain. However, too much IDK chosen data can lead to overly abstention resulting in decrease in overall abstain quality. Answer quality increases in sync with abstain quality showing an interesting balance. As the IDK ratio increases, the quality of correct responses does not decline significantly compared to the sharp rise in the model's refusal to answer. While the recall decreases as a result of fewer correctly answered questions, this way improves the precision of correct responses, ultimately enhancing the overall F1. However, when the model begins to overuse IDK (e.g., at extremely high ratio), this strategy ceases to work, as excessive abstention undermines correct answer coverage and utility. In addition, both DR and CUR scores consistently decrease as the IDK ratio increases, primarily due to the reduction in the proportion of ✔✘ and ✘✔ training data. The results suggest that moderate IDK ratios strike an optimal balance between precision and robustness, while aggressive reliance on IDK triggers diminishing returns.

F Prompts

F.1 Context Evaluation Prompt

The following prompt is used to evaluate whether a context contains or implies the correct answer to a query:

You are an expert at evaluating whether a context contains the correct answer to a question. You should:
1. Check if the given answer can be found or directly implied by the context
2. Return a score of 1 if the context contains or directly implies the answer
3. Return a score of 0 if the context does not contain or support the answer
4. Provide a brief explanation for your decision
Respond in the following JSON format:
{
"score": 0 or 1,
"explanation": "your explanation here"
}

G Implementation Details

G.1 Our Method Implementation

For our proposed approach, we train the model for 3 epochs using a cosine learning rate scheduler with an initial learning rate of 5e-5 and a warmup ratio of 0.1. The β and γ are set to 1.0 and 0.5 respectively for all experiments. The training process employs the Paged AdamW optimizer with 32-bit precision and a weight decay of 0.05. To balance computational efficiency and memory constraints, we set the batch size to 16 per device with 2 gradient accumulation steps, allowing for effective training on larger datasets while maintaining memory efficiency. The threshold δ used for $\mathrm{KB}_{\mathrm{param}}$ to sample $N(=10)$ responses is 1.0. Moreover, experiments are conducted on NVIDIA A100 GPUs with 80G of memory. Fixed random seed of 0 is used and the experimental results are reported within a single run. The versions of the libraries used in this work are as follows: accelerate version 0.34.2, transformers version 4.46.3, trl version 0.12.1 and vllm version 0.6.1.post2. And the dpo training process costs approximately 6 GPU hours.

G.2 Baselines Implementation

We implement several baseline detection methods for comparison:

- **P(True)**: Following Kadavath et al. (2022), we prompt the LLM to evaluate the correctness of its own answer. The prompt presents the original question and the model's proposed answer, asking for a binary True/False classification. We experiment with multiple

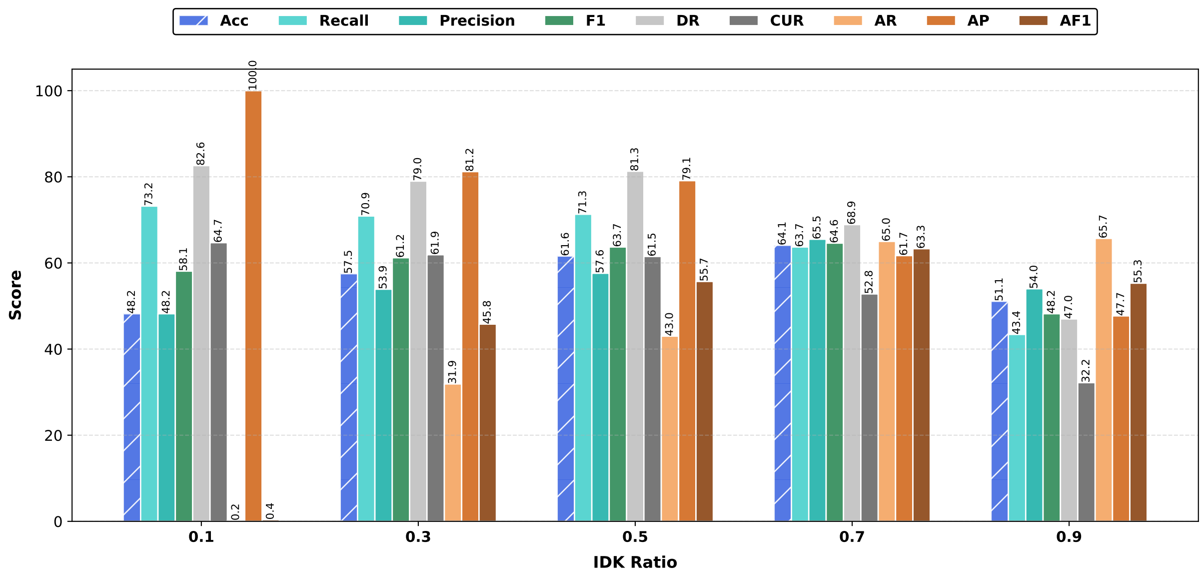

Figure 5: Experiments across IDK ratio. (DPO data size=5k, loss weights β=1.0, γ=0.5)

confidence thresholds (0.3, 0.5, 0.7, 0.9) to determine the optimal cutoff for each experimental setting.

> Question: [Question]
> Proposed Answer: [Predictions]
> Is the proposed answer:
> (A) True
> (B) False
> The proposed answer is:

- **Logits**: We implement three baselines using different logprob statistics of the output tokens: minimum (Min), mean (Mean), and last token (Last). The Min baseline, which uses the minimum logprob across all output tokens, is the only one reported in the paper as the other two approaches proved ineffective at enabling model abstention. We experiment with multiple logtis thresholds (-2.0, -1.0, 0.0) to determine the optimal cutoff for each experimental setting.

- **Self-Consistency**: We generate multiple responses (n=10) for each question and measure consistency among the generated answers. The system proceeds with answering if the most frequent response receives more than 5 votes; otherwise, it abstains. This approach helps identify cases where the model exhibits high uncertainty through response variability.

- **ICL**: We implement in-context learning using a prompt template with three carefully curated demonstration examples: one showing appropriate abstention for out-of-knowledge-boundary queries, and two showcasing correct answer generation for in-boundary queries. This balanced demonstration set helps the model learn both when to answer and when to abstain.

H Licensing

Llama2-7B and Llama2-13B are released under the Meta Llama 2 Community License Agreement. Llama3-8B is released under the Meta Llama 3 Community License Agreement. All of them are accessible for academic usage and consistent with their intended use.

And three open-domain QA datasets, Natural Questions (NQ), TriviaQA, and WebQuestions (WebQ) are publicly available for academic research purposes, which is also consistent with their intended use.