All That Glitters is Not Gold: Improving Robust Retrieval-Augmented Language Models with Fact-Centric Preference Alignment

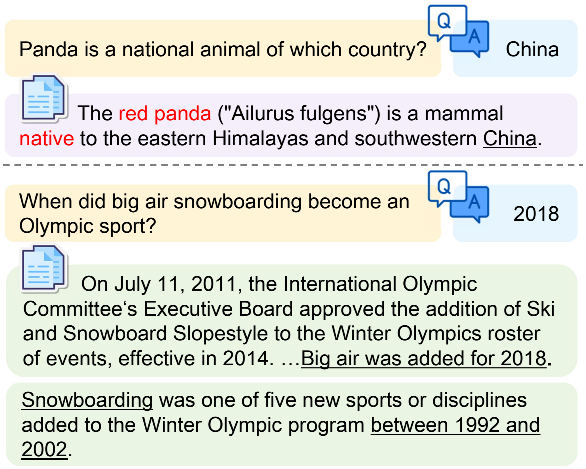
Jia Hao, Chunhong Zhang, Jiarun Liu, Haiyu Zhao, Zhiqiang Zhan, Zheng Hu[*]
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{haojia, zhangch, liujiarun01, zhaohaiyu, zqzhan, huzheng}@bupt.edu.cn

Abstract

Retrieval-augmented language model (RALM) relies on retrieved external knowledge to generate responses, resulting in vulnerability in the face of retrieval results with noisy documents. Previous works integrate additional filters or finetune Large Language Models (LLMs) to learn adaptive retrieval to reduce the performance damage of noisy documents. However, prior noise filtering may lead to the loss of crucial information, and these methods do not focus on distracting documents with high semantic relevance, which is the most challenging problem. In this study, we propose a training method for fact-centric preference alignment (FPA) to improve the ability of LLMs to directly extract useful information from noisy retrieval results without prior filtering. Our method performs positive document mining based on factual consistency and uses LLMs self-generated synthetic data as training data without manual annotation. We evaluate our FPA on four question answering benchmarks, and the experimental results demonstrate that our method achieves significant improvement with a small scale of training data.[1]

1 Introduction

Although Large Language Models (LLMs) show excellent performance in many tasks such as text comprehension and reasoning (Brown et al., 2020; Wei et al., 2022; OpenAI, 2023), they still face challenges such as lack of knowledge and hallucinations in knowledge-intensive tasks (Elazar et al., 2021; Ji et al., 2023), etc. Retrieval-Augmented Language Models (RALMs) alleviate this problem by providing retrieved external information to LLMs (Guu et al., 2020; Lewis et al., 2020b; Shuster et al., 2021; Shi et al., 2024). However, retrieval results cannot be perfect, and relevant information is often mixed with irrelevant or misleading context, leading LLMs to generate incorrect responses (BehnamGhader et al., 2023; Shi et al., 2023; Wu et al., 2024; Cuconasu et al., 2024).



Panda is a national animal of which country?

China

The red panda ("Ailurus fulgens") is a mammal native to the eastern Himalayas and southwestern China.

When did big air snowboarding become an Olympic sport?

2018

On July 11, 2011, the International Olympic Committee's Executive Board approved the addition of Ski and Snowboard Slopestyle to the Winter Olympics roster of events, effective in 2014. …Big air was added for 2018.

Snowboarding was one of five new sports or disciplines added to the Winter Olympic program between 1992 and 2002.

Figure 1: Examples of irrelevant documents in the retrieval results. The first example (top) shows that a document containing the ground truth may still be irrelevant. The second example (bottom) shows that two highly semantically related documents actually contain different facts.

To solve this problem, previous works add a post-retrieval process to granularly select necessary information and filter out noise by re-ranking (Bai et al., 2023), compression (Jiang et al., 2023; Zhu et al., 2024; Yu et al., 2024), and other methods (Ke et al., 2024). However, these methods may suffer from loss of critical information due to excessive compression or reranking bias, leading to performance degradation. Some methods directly fine-tune the LLMs and teach LLMs to adaptively decide when to retrieve, reducing the potential harm of irrelevant information by avoiding unnecessary retrieval (Asai et al., 2023; Yoran et al., 2024; Jeong et al., 2024). However, these methods do not focus on improving the ability of LLMs to distinguish distracting documents, and may still be

[*] Corresponding author.

[1] Our code is available at https://github.com/haojia-hj/FPA.

Findings of the Association for Computational Linguistics: ACL 2025, pages 11277–11292
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

misled when providing retrieved documents.

In this work, we enhance the noise robustness of RALMs by improving the ability of LLMs to distinguish between relevant and noisy documents (especially distracting documents with high semantic similarity), rather than adding additional noise filters. We expect LLM to (1) refer to useful information as comprehensively as possible while avoiding referring to any irrelevant information, and (2) make use of its internal knowledge to answer the given question when the retrieval results do not contain enough useful information. To achieve this, we propose a new training method for fact-centric preference alignment called **FPA**. The key aspect of our method is to label useful information (i.e., positive documents) and distracting information (i.e., negative documents) from a new perspective, since they are difficult to distinguish in terms of semantic similarity. Inspired by the application of the NLI model in citation quality evaluation (Gao et al., 2023) and conflict detection (Jiayang et al., 2024), we propose a fact-centric positive document mining approach that automatically labels documents based on their entailment relationship with ground truth statements, without relying on manually annotated gold documents. Then, we sample responses according to positive and negative documents and design a citation mismatch data augmentation strategy to construct preference data. By fine-tuning the model for preference alignment, our approach improves LLM's ability to distinguish between useful and distracting information.

The main contributions of this paper are as follows:

- We design a fact-centric positive document mining method to determine the relevance of documents from the perspective of factual mentions, which applies to scenarios where gold documents are hard to obtain.

- We propose a preference alignment training method to improve the noise robustness of RALMs, which finetunes LLM based on the self-generated preference data to enhance LLM's ability to filter distracting information and extract answers from noisy contexts.

- Experimental results show that our method exhibits advanced noise robustness and generalization on both short-form and long-form question answering tasks, and achieves efficient performance improvement with a small scale of training data.

2 Related Work

2.1 Robust Retrieval-Augmented Language Models

RALMs use sparse (Robertson and Zaragoza, 2009) or dense (Karpukhin et al., 2020) retrievers to retrieve documents that are highly relevant to the questions and utilize LLMs to generate answers based on the documents. However, retrieval results are usually imperfect and contain irrelevant content, which may mislead LLMs to generate incorrect responses. Previous studies (Cuconasu et al., 2024; Wu et al., 2024) have observed that distracting documents containing highly semantically related but irrelevant information can cause significant performance degradation, which is attributed to the shortcomings of LLMs in identifying irrelevant information.

Previous works, such as RECOMP (Xu et al., 2023) and LLMLINGUA (Jiang et al., 2023), integrate trainable noise filters to discard irrelevant context. However, these methods may over-filter the context, resulting in the loss of critical information before it is fed into the LLM. There are also methods that finetune the LLM to decide when to retrieve and whether the retrieved documents are relevant (Asai et al., 2023; Yoran et al., 2024). These methods do not focus on improving the model's ability to distinguish distracting documents, and thus may be disturbed when performing retrieval.

2.2 Preference Alignment

Preference alignment methods optimize models by introducing human feedback (preference) to align their outputs with human intentions (Jiang et al., 2024b). RLHF (Ouyang et al., 2022) is one of the major efforts that has achieved breakthrough success in LLM alignment. It collects human-labeled comparison data to train a reward model and optimizes a policy using PPO reinforcement learning algorithm. Despite its effectiveness, the method still faces challenges, including the difficulty in obtaining feedback data and the complexity of the training process. Some works (such as DPO (Rafailov et al., 2023)) are able to bypass the reward modeling and reinforcement learning process, reducing the difficulty of training. In the RAG task, previous work leverages preference optimization to improve the attribution of LLM output (Huang et al., 2024).

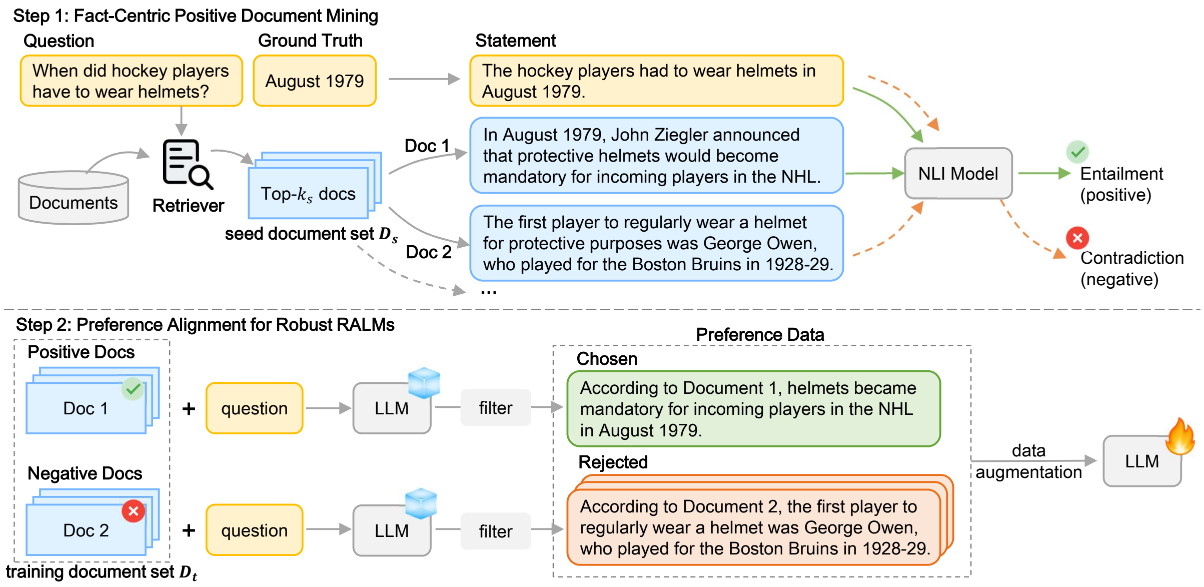

Step 1: Fact-Centric Positive Document Mining
Question
When did hockey players have to wear helmets?
Ground Truth
August 1979
Statement
The hockey players had to wear helmets in August 1979.
Documents
Retriever
Top-k_s docs
seed document set D_s
Doc 1
In August 1979, John Ziegler announced that protective helmets would become mandatory for incoming players in the NHL.
Doc 2
The first player to regularly wear a helmet for protective purposes was George Owen, who played for the Boston Bruins in 1928-29.
...
NLI Model
Entailment (positive)
Contradiction (negative)
Step 2: Preference Alignment for Robust RALMs
Positive Docs
Doc 1
Negative Docs
Doc 2
training document set D_t
+
question
LLM
filter
+
question
LLM
filter
Preference Data
Chosen
According to Document 1, helmets became mandatory for incoming players in the NHL in August 1979.
Rejected
According to Document 2, the first player to regularly wear a helmet was George Owen, who played for the Boston Bruins in 1928-29.
data augmentation
LLM

Figure 2: The framework of FPA. FPA first uses the NLI model to estimate the entailment relationship between documents and ground truth statements for positive document mining, then constructs preference data by sampling LLM responses based on positive and negative documents to align the model.

In this work, we use automatically constructed preference data to optimize the ability of LLM to exclude interference from irrelevant information.

3 Methodology

In this section, we introduce the task formulation (§3.1) and our approach FPA, a training method to improve the noise robustness of RALMs using self-generated synthetic data from LLMs, as illustrated in Figure 2. Our method first introduces a new dimension for mining potential positive documents (§3.2), and samples training data for preference alignment based on positive and negative documents. We then use DPO to improve the model's ability to focus on relevant information while excluding distracting context (§3.3) and integrate adaptive retrieval prompting during inference (§3.4).

3.1 Task Formulation

For a question answering dataset $\mathcal{D}$, which contains a question set Q, each question $q \in Q$ corresponds to an answer set A containing one or more answers. In the naive RAG process, for a given question q, the retriever maps the question and documents to embedding vectors using different encoders. The dot product of the two vectors is defined as the relevance score and used to find the most relevant top-k documents:

$$rel(q,d) = Encoder_Q(q)^T Encoder_C(d) \quad (1)$$

The retrieved documents D_r are concatenated with the question q through prompt templates as input to LLM. LLM then generates the answer based on the provided information:

$$y_t = \arg\max_y P(y|D_r, q, y_{1:t-1}) \quad (2)$$

3.2 Fact-Centric Positive Document Mining

The existing retrievers measure relevance based on vector similarity, which will recall documents with high semantic similarity but irrelevant to the question, misleading LLMs to generate incorrect answers. Previous approaches typically define positive documents (i.e., relevant documents) as human-annotated gold documents in the dataset, or documents that contain the ground truth answers. However, gold documents may introduce human bias and are difficult to obtain in some scenarios. As shown in Figure 1, even if a document contains the correct answer, it may still be irrelevant, especially if the answer is a common word (such as year, country, etc.). To this end, we switch to a different perspective: judging the relevance of a document based on its factual mentions.

A question q and its corresponding ground truth answer $\hat{a}$ can form a statement of fact $s = f(q, \hat{a})$. For example, the question *"Where do you cross the Arctic Circle in Norway?"* and its ground truth answer *"Saltfjellet"* form the fact *"You cross the Arctic Circle in Norway at Saltfjellet."* The example in Figure 1 demonstrates that although the

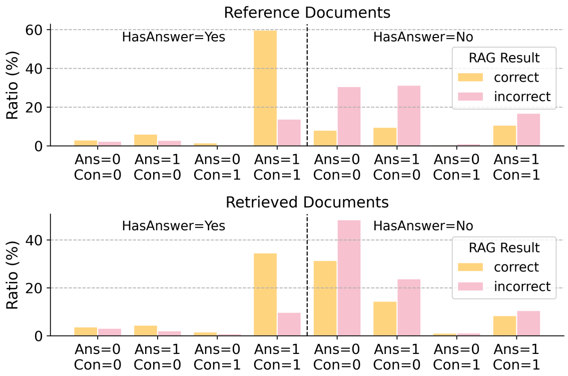


Figure 3: Ratios of documents in context with different answerability and factual consistency when naive RAG answers correctly and incorrectly.

semantic similarity between the distracting and positive documents is so close as to be indistinguishable, the facts they contain are usually inconsistent. Therefore, we mine positive documents by checking whether the facts contained in a given document are consistent with the ground truth facts. The documents that consist with the ground truth facts are classified as positive documents, while the other documents are classified as negative documents.

In preliminary experiments, we prompt LLM to score answerability Ans and factual consistency Con (taking a value of 0 or 1) of the documents, where answerability refers to whether a document contains enough information to answer the question, and factual consistency refers to whether the facts in a document are consistent with the ground truth statement. Based on the scores and whether the document contains ground truth answers ($HasAnswer$), the possible cases can be classified into four categories: (1) **Relevant Documents**: $Ans = 1$, $Con = 1$ and $HasAnswer = Yes$; (2) **Potentially Irrelevant Documents**: $HasAnswer = Yes$ but are considered unanswerable or factual inconsistent; (3) **Irrelevant Documents**: $Ans = 0$, $Con = 0$ and $HasAnswer = No$; (4) **Distracting Documents**: $HasAnswer = No$ but are considered answerable or factual consistent.

We count the ratios of the four types of documents in the context when naive RAG answers correctly or incorrectly, as shown in Figure 3. It can be observed that: (1) for questions answered correctly by RAG, although the ratio of "Relevant Documents" is similar to that of "Irrelevant Documents" in the retrieval results, it is significantly higher in the reference documents, indicating that the ability of LLMs to refer to "Relevant Documents" is very important for the correctness of the RAG's answer; (2) the existence of "Potentially Irrelevant Documents" indicates that some documents, although containing the correct answers, may be considered irrelevant by the LLMs due to incomplete or irrelevant content; (3) for questions answered incorrectly by RAG, "Distracting Documents" account for a higher proportion of both retrieval results and reference documents, indicating that these documents are easily considered relevant by LLMs, which may mislead the model to generate incorrect answers.

In practice, we transform the task of checking the factual consistency between a document and a ground truth statement into a recognizing textual entailment (RTE) task. Consider the document as *premise* and the ground truth statement as *hypothesis*; if the two texts are entailment, they are deemed to be factual consistent. We use a NLI model to predict whether two texts are entailment, which is a three-classification task with categories Contradiction (C), Neutral (N) and Entailment (E). A document d is considered factual consistent with a ground truth statement s if the model predicts the highest probability for the Entailment category:

$$con(d, s) = \mathbb{I}(P(E) > max(P(C), P(N))) \quad (3)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. This ensures that irrelevant documents are considered as negative documents, since they are typically classified as Neutral or Contradiction in the NLI-based annotation.

3.3 Preference Alignment for Robust RALMs

In this section, we use the obtained positive and negative documents to construct the training data by automatically sampling the responses, and improve the model's ability to distinguish between relevant and distracting documents through preference alignment.

Construct the training dataset. For a subset of the training set of QA dataset, the top-k_s retrieval results (seed document set D_s) for each question are first obtained, with varying percentages of positive documents. The positive and negative labels of the documents are obtained using the method outlined in Section 3.2. k_t documents are selected from D_s to form the training document set D_t, including no more than $k_t - 1$ positive documents with the highest relevance scores and randomly selected negative documents. We randomly shuffle the order of selected k_t documents (instead of sorting by relevance) to increase the difficulty, and obtain the positive and negative document indexes corresponding to the current order. When implemented, k_s is set slightly larger than k_t to ensure more diverse seed documents. We take $k_s = 8$ and $k_t = 5$ in the experiment.

By adding a trigger with document indexes (e.g., *"According to documents 1 and 5"*) at the end of the prompt, it is possible to control which documents are referenced when sampling the response. We expect LLMs to reference as many positive documents as possible, thus we add the index of all positive documents in D_t into the trigger to sample the preferred (chosen) response. To ensure response quality, only the correct response is retained. In addition, we expect LLMs to avoid referencing any negative documents, thus the index of negative document in D_t is added into the trigger sequentially (one at a time) for sampling the dispreferred (rejected) responses and the incorrect responses are retained. The preference alignment training dataset comprises input prompt and preferred-dispreferred response pairs. Given that the number of negative documents and the order of documents varies between different questions, our method constructs preference data with varying noise ratios and positive document positions.

To avoid LLM's hallucination of references (i.e., document indexes are out of range), we design a **citation mismatch data augmentation** strategy, which modifies the reference document index in the chosen response to the unreferenced document index as the rejected response.

Training objective. The training objective of the generator (LLM) is to avoid referring to negative documents while referring to positive documents as much as possible when answering questions, and to extract correct answers based on positive documents. In order to make full use of the negative document samples, we use Direct Preference Optimization (DPO) to fine-tune the LLMs:

$$\mathcal{L}_{DPO} = -\mathbb{E}[\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))] \quad (4)$$

where

$$r_\theta(x, y) = \beta log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \quad (5)$$

where the input $x = \mathcal{I} \bigoplus D_t \bigoplus q$ consists of the instruction $\mathcal{I}$, question q, and training document set D_t. y_w is the denoised response that refers to positive documents and contains the correct answer, while y_l is the noisy response that refers to negative documents and is answered incorrectly.

Dataset	Train	Test	Retriever
NQ	79168	3610	DPR
PopQA	12868	1399	Contriever
TriviaQA	78785	11313	Contriever
ASQA	4353	948	GTR

Table 1: Dataset statistics.

3.4 Inference

Over-reliance on context can lead to incorrect responses when there is no relevant information in the retrieval results. In contrast to Self-RAG (Asai et al., 2023) and InstructRAG (Wei et al., 2024), which finetune LLM to learn when to retrieve, we incorporate adaptive retrieval prompting into the instruction. It instructs LLM to use internal knowledge to make responses when the retrieval results do not contain enough useful information. The prompt templates are available in Appendix C.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

We choose three short-form QA datasets, including Natural Questions (NQ) (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023) and TriviaQA (Joshi et al., 2017), and one long-form QA dataset ASQA (Stelmakh et al., 2022) to evaluate our approach. Following previous works (Asai et al., 2023; Wei et al., 2024), we use Wikipedia as an external corpus and leverage different retrievers (including DPR (Karpukhin et al., 2020), Contriever-MS MARCO (Izacard et al., 2021), and GTR (Ni et al., 2022)) to retrieve top-5 documents for each question. The dataset statistics are shown in Table 1. For NQ, PopQA and TriviaQA, we evaluate the accuracy based on whether the output contains ground truth answers (EM). For ASQA, we use the official metrics in ALCE (Gao et al., 2023), including correctness (str-em), citation precision (pre) and recall (rec).

4.2 Baselines

We compare our method with other noise robustness improvement methods without and with finetuning LLM, respectively.

For *w/o Retrieval* method, we utilize LLM to directly answer questions with parametric knowledge.

The *w/o Finetuning LLM* methods include (1) Vanilla RAG, which provide top-5 documents retrieved by the retriever from Wikipedia as exter-

	NQ	PopQA	TriviaQA	ASQA		
Methods	(acc)	(acc)	(acc)	(str-em)	(pre)	(rec)
w/o Retrieval	48.2	28.0	70.3	32.1	-	-
w/o Finetuning LLM						
Vanilla RAG	56.5	62.7	71.7	42.5	50.9	76.8
RAG w/ rerank (Zhuang et al., 2022)	58.0	**69.5**	74.9	43.4	28.7	46.7
RAG w/ compress (Jiang et al., 2024a)	41.3	44.9	64.8	27.7	-	-
w/ Finetuning LLM						
Self-RAG (Asai et al., 2023)	42.8	55.8	71.4	36.9	**69.7**	69.7
RetRobust (Yoran et al., 2024)	54.2	56.5	71.5	40.5	-	-
InstructRAG-FT (Wei et al., 2024)	65.7	66.2	78.5	47.6	65.7	70.5
InstructRAG-FT †(Wei et al., 2024)	65.7	65.2	76.5	45.3	40.0	63.8
FPA (Ours) †	**66.9**	67.1	**78.7**	**48.8**	58.3	**77.9**

Table 2: Experimental results of our FPA and baselines on four QA benchmarks. The best results are in **bold** and second best results are in underlined. †indicates the model is trained on NQ only, for which we report the in-domain performance for NQ and out-of-domain performance for the other benchmarks. Part of the baseline results are referenced from (Wei et al., 2024). "-" indicates the results are not reported in the original paper or not applicable.

nal knowledge; (2) RAG w/ rerank, which use RankT5-large (Zhuang et al., 2022) to re-score and rerank retrieval results for more than 5 documents, and choose top-5 documents with the highest reranking scores; (3) RAG w/ compress, which use LongLLMLingua (Jiang et al., 2024a) to compress the top-5 retrieved documents.

The *w/ Finetuning LLM* methods include (1) Self-RAG (Asai et al., 2023), which finetunes LLM to judge the need for retrieval and evaluate the quality of generated responses by predicting special reflection tokens; (2) RetRobust (Yoran et al., 2024), which adaptively refers to retrieved documents and finetunes LLM to ignore irrelevant contexts; (3) InstructRAG-FT (Wei et al., 2024), which uses LLM self-generated explicit denoising rationales as demonstrations for in-context learning, or supervised data for LLM fine-tuning.

4.3 Implementation Details

Our experiments employ Llama-3-8B-Instruction (Dubey et al., 2024) to construct 5k preference data from a subset of the NQ training set and perform preference alignment, and BART-Large-MNLI (Lewis et al., 2020a) for positive document mining. We use LoRA to fine-tune LLM on two NVIDIA RTX 4090 GPUs for 5 epochs, which takes 3.5 hours. We apply greedy decoding to sample the LLM response. More experimental details are presented in Appendix A.

5 Experimental Results

5.1 Main Results

Table 2 shows the overall experimental results of our method and the baselines on four QA benchmarks.

RAG without fine-tuning LLM. The experimental results indicate that our method generally outperforms the noise filtering methods of reranking and compression. Compared to *Vanilla RAG*, reranking or compressing the retrieval results does not significantly improve the performance of RAG, and even severely decreases the accuracy. One exception is that *RAG w/ rerank* achieves competitive performance on PopQA. Although reranker's more precise relevance evaluation helps to improve the recall of relevant documents, it is not effective for all tasks. For *RAG w/ compress*, we implement 3× compression on retrieved documents with a length of around 700 tokens, and there may be over-compression leading to loss of key information. Although LongLLMLingua performs well in processing long contexts with gold documents, the actual quality of the retrieval results can be worse, which may be the reason for the significant performance degradation due to compression. Our approach uses LLM to process the noisy text directly, avoiding the information loss caused by post-retrieval processing.

RAG with fine-tuning LLM. The results indicate that our method outperforms the baseline methods or is comparable to the state-of-the-art InstructRAG-FT for both in-domain performance

	NQ	PopQA	ASQA			Avg. Acc
Ablation	(acc)	(acc)	(str-em)	(pre)	(rec)	
base model	56.5	62.7	42.5	50.9	76.8	53.9
+ preference alignment	62.9	64.9	44.9	48.9	73.4	57.6
+ adaptive retrieval	65.1	65.3	47.6	49.7	73.6	59.3
+ preference alignment & adaptive retrieval	**67.1**	66.3	48.5	49.6	74.9	60.6
FPA (Ours)	66.9	**67.1**	**48.8**	**58.3**	**77.9**	**60.9**

Table 3: Ablation study of preference alignment, adaptive retrieval prompting and citation mismatch data augmentation in FPA. **Avg. Acc** calculates the average of acc and str-em on the three benchmarks.

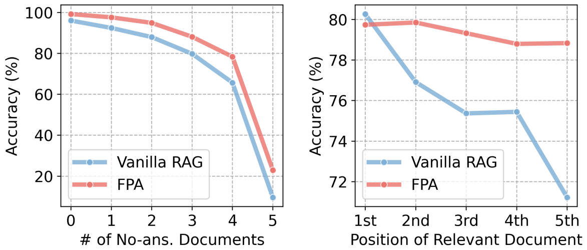


Figure 4: Accuracy of FPA and baselines at different noise ratios (left) and position of relevant document (right) on NQ benchmark.

on NQ and out-of-domain performance on the other datasets. Our method does not achieve optimal citation precision because the preference data encourages referencing as much useful information as possible, rather than citing a minimum sufficient subset of the documents. For a fair comparison, we test the generalization performance of the InstructRAG-FT model trained only on NQ. The results show that our method also has superior generalization performance over InstructRAG-FT. Although our method fine-tunes LLM using the retrieval results from DPR, there are also significant improvements when using the Contriever and GTR, indicating that our method can be effectively combined with different retrievers. Meanwhile, FPA achieves optimal str-em on ASQA, revealing that our method generalizes well from short-form QA tasks to long-form QA tasks.

5.2 Ablation Study

We add single or multiple components separately to analyze the role of each component in our method. The experimental results are shown in Table 3.

Effect of preference alignment. Preference alignment is used to improve LLM's ability to extract useful information from noisy text. In this setting, LLM is fine-tuned only on preference data constructed from positive and negative documents without citation mismatch data augmentation and inference without adaptive retrieval prompting. The results demonstrate that preference alignment using constructed preference data can effectively improve the performance of LLM in processing noisy contexts.

Effect of adaptive retrieval prompting. Adaptive retrieval prompting instructs LLMs to respond directly when retrieval results do not contain useful information through zero-shot prompting. The results indicate that adaptive retrieval prompting significantly improves the answer accuracy of the base model, especially in the case of poor retrieval quality. When adaptive retrieval prompting is combined with aligned LLM, the performance of RALM exceeds that of either alone.

Effect of citation mismatch data augmentation. Citation mismatch data augmentation is applied in the construction of preference data to avoid LLM's hallucination of references. The results reveal that citation mismatch data augmentation can improve the citation quality in LLM responses. However, it has no significant effect on the accuracy of QA.

5.3 Analysis

Robustness to noise ratios. We simplify the noise ratio as the ratio of documents that do not contain ground truth answers (No-ans.) in the top-5 retrieval result. Figure 4 shows the accuracy at different noise ratios on NQ. The results demonstrate that as the noise ratio increases, the accuracy of RALM decreases due to the lower density of useful information. Compared to vanilla RAG, our method improves the accuracy of RALM at different noise ratios by preference alignment. Even when all retrieved documents are noisy, our method still has an accuracy of 22.88%. We attribute this to adaptive retrieval prompting, which encourages LLM to answer based on internal knowledge when no useful information is available.

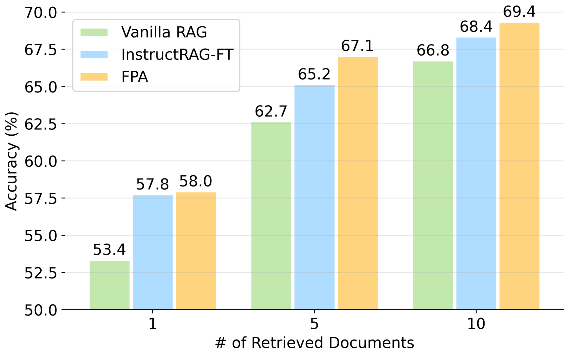


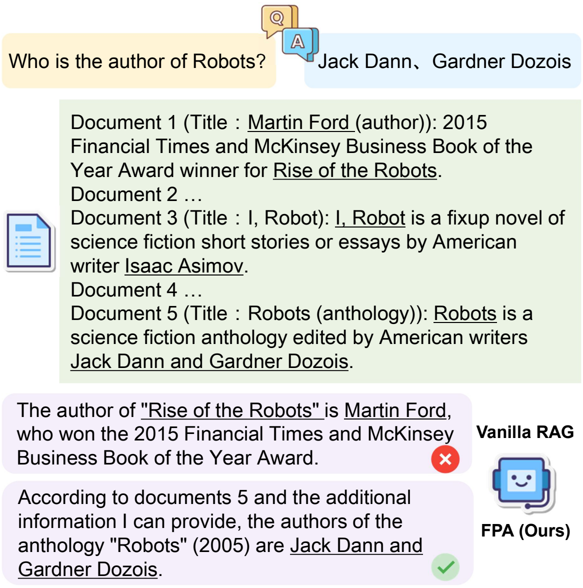
Figure 5: Accuracy of FPA and baselines when top-k retrieved documents are provided on PopQA benchmark, where k is set to 1, 5, 10.

Settings	NQ (acc)	PopQA (acc)	TriviaQA (acc)	ASQA (acc)
w/o evidence	49.9	29.6	71.1	34.1
w/ evidence	66.9	67.1	78.7	48.8

Table 4: Accuracy of finetuned FPA model with and without retrieved evidence.

Robustness to relevant document positions. Similar to the setting of Liu et al. (2024), we investigate the scenario where the provided documents contain 1 gold document (relevant document) and 4 top-ranked No-ans. documents (distracting documents). The accuracy of our method when relevant documents are in different positions is illustrated in Figure 4. For vanilla RAG, the accuracy decreases as the relevant document approaches the end, up to 11.26%. This decline can be attributed to the model's preference to focus on information at the beginning of the context. In our approach, the order of documents in the preference data is shuffled to make the distribution of relevant document positions more balanced, which enables the fine-tuned LLM to focus on relevant information at different positions.

Generalisability to the number of retrieved documents. We evaluate the accuracy of our FPA and baselines when different numbers of retrieved documents are provided in the context, as shown in Figure 5. As the number of retrieved documents increases, the accuracy of all three methods increases, on account of more useful information in the retrieval results. Our method significantly improves the performance of vanilla RAG and outperforms the state-of-the-art InstructRAG-FT, showing excellent generalization to varying numbers of retrieved documents.


Who is the author of Robots?

Jack Dann、Gardner Dozois

Document 1 (Title：Martin Ford (author)): 2015 Financial Times and McKinsey Business Book of the Year Award winner for Rise of the Robots.
Document 2 …
Document 3 (Title：I, Robot): I, Robot is a fixup novel of science fiction short stories or essays by American writer Isaac Asimov.
Document 4 …
Document 5 (Title：Robots (anthology)): Robots is a science fiction anthology edited by American writers Jack Dann and Gardner Dozois.

The author of "Rise of the Robots" is Martin Ford, who won the 2015 Financial Times and McKinsey Business Book of the Year Award.

According to documents 5 and the additional information I can provide, the authors of the anthology "Robots" (2005) are Jack Dann and Gardner Dozois.

Figure 6: Case study.

Methods	Training Data	FT Models	Require GPT-4
Self-RAG	150k	cri.+gen.	✓
RetRobust	1k	gen.	✗
InstRAG	70k	gen.	✗
FPA	5k	gen.	✗

Table 5: Efficiency Analysis of FPA and baselines. FT stands for Finetuned. cri. denotes critical model and gen. denotes generator.

Better memory or better analysis? We conduct additional experiments to verify that our FPA model tends to select appropriate evidence rather than simply showing internal knowledge. Specifically, we use the finetuned FPA model to answer questions directly without providing evidence to simulate a scenario where the model completely ignores retrieved content. As shown in Table 4, the accuracy is significantly lower in the *w/o evidence* setting than in the *w/ evidence* setting, indicating that the FPA model does not rely solely on internal knowledge in the RAG process. This confirms that when retrieved documents are provided, the model does not neglect to analyze the evidence.

Efficiency. We compare the efficiency of our method with baselines, as illustrated in Table 5. Compared to Self-RAG and InstructRAG-FT, our method constructs less training data without access to GPT-4 and fine-tunes the generator only without training additional critical models, which reduces the cost of data construction and training. Similar to RetRobust, our approach works on improving the noise robustness of the model with a small number of samples and achieves better performance.

Case study. Figure 6 shows the responses generated by vanilla RAG and our FPA. The question pertains to the author of the book “Robots”. However, the vanilla RAG is misled by information about the author of “Rise of the Robots” in Document 1, resulting in the incorrect answer. In contrast, our FPA is not disturbed by the noisy information about other works in Documents 1 and 3, and correctly outputs the authors of “Robots” based on the evidence in Document 5. Further case studies are presented in Appendix B.

6 Conclusion

In this work, we propose FPA, a new fact-centric preference alignment training method for improving the noise robustness of RALMs. FPA first utilizes the entailment relationship between documents and ground truth statements for positive document mining, and constructs preference data by sampling LLM responses according to positive and negative documents. The preference data are used to align LLM for enhancing its capacity to distinguish distracting documents and generate correct answers. Experimental results on four QA benchmarks indicate that our approach achieves higher accuracy than existing methods with less training data, and exhibits strong generalization from short-form QA to long-form QA.

Limitations

We focus on improving the robustness of RALM when faced with noisy retrieval results and achieve significant performance gains. However, in the case of extremely poor retrieval quality, the improvement of our method will be considerably limited. Therefore, improving retriever performance remains a key aspect for the overall performance of RALMs. In addition, our experiments only involve single-hop ODQA and do not evaluate domain-specific tasks or multi-hop QA. Future work directions will include extending our approach to more complex knowledge-intensive tasks and optimizing the retriever.

Acknowledgments

We appreciate the anonymous reviewers for their valuable feedback and suggestions. This work was supported by Beijing Association of Higher Education project (MS2023151) and Education and Teaching Reform Project of Beijing University of Posts and Telecommunications (2024ZD18).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jiaqi Bai, Hongcheng Guo, Jiaheng Liu, Jian Yang, Xinnian Liang, Zhao Yan, and Zhoujun Li. 2023. Griprank: Bridging the gap between retrieval and generation via the generative knowledge improved passage ranking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 36–46.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. 2023. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15492–15509, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

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*, page 719–729.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. 2024. The llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. Advancing large language model attribution through self-improving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3822–3836, Miami, Florida, USA. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024a. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024b. A survey on human preference learning for large language models. *Computing Research Repository*, arXiv:2406.11191.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. ECON: On the detection and resolution of evidence conflicts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7816–7844, Miami, Florida, USA. Association for Computational Linguistics.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10438–10451, Bangkok, Thailand. Association for Computational Linguistics.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

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.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*.

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

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *Computing Research Repository*, arXiv:2310.04408.

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

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng

Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1069, Bangkok, Thailand. Association for Computational Linguistics.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2022. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A Experimental Details

A.1 Datasets

For PopQA, we use the English Wikipedia dump from December 2020 as the retrieval source; for other benchmarks, we use Wikipedia from December 2018. Under the default setting, we use top-5 retrieval results returned by the retriever. Under the RAG w/ rerank setting, we use the top-25 retrieval results[2] provided by Asai et al. (2023) for PopQA, and the top-100 retrieval results for NQ[3], TriviaQA and ASQA[4].

We count the positive document position distribution in our constructed preference dataset, as shown in Figure 7. Due to the shuffled document order, the distribution of positive document positions in our preference dataset is relatively balanced, which is beneficial to improve the model's robustness to the positions of relevant documents. The distribution of the number of positive documents in the context is shown in Figure 8. This distribution is close to the actual retrieval results. The samples with fewer positive documents account for the majority of the samples, which helps the model learn to focus on useful information under a high noise ratio.

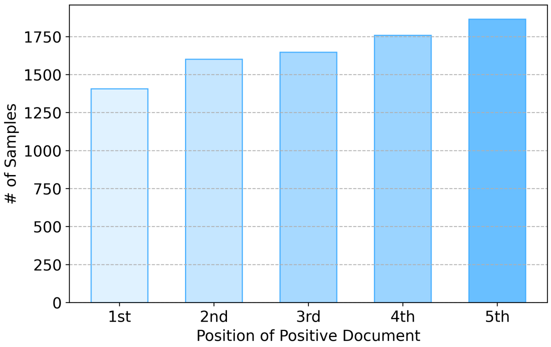


Figure 7: Distribution of positive document positions in our preference dataset.

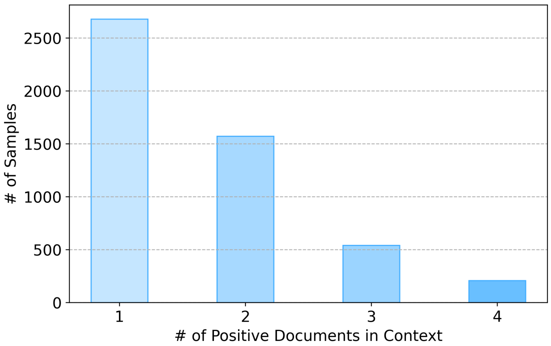


Figure 8: Distribution of the number of positive document in our preference dataset.

A.2 Baselines

w/o Retrieval. LLM directly answers questions with parametric knowledge.

Vanilla RAG. LLM answers questions based on top-5 documents retrieved by the retriever from Wikipedia.

RAG w/ rerank. RankT5-large[5] (Zhuang et al., 2022) is used to re-score and rerank retrieval results for more than 5 documents. We provide top-5 documents with the highest reranking scores to LLM as the external knowledge.

RAG w/ compress. We use LongLLMLingua (Jiang et al., 2024a) to compress the top-5 retrieved documents. LongLLMLingua uses a small language model to filter out noisy text by capturing key information related to the question and discarding unimportant tokens in the prompt. We employ the default hyperparameter settings of 10× compression example[6] to get compressed text.

Self-RAG (Asai et al., 2023). LLM is trained to predict special reflection tokens to judge the need for retrieval and evaluate the quality of retrieval results and generated responses. The method trains a small critical model to add reflection tokens to the input-output pair for generator LLM training.

RetRobust (Yoran et al., 2024). The method determines whether to refer to retrieved document based on its relevance and trains LLM to ignore irrelevant contexts. A document is considered rel-

[2] https://github.com/AkariAsai/self-rag

[3] https://github.com/facebookresearch/DPR

[4] https://github.com/princeton-nlp/ALCE

[5] https://huggingface.co/Soyoung97/RankT5-large

[6] https://github.com/microsoft/LLMLingua/blob/main/examples/RAG.ipynb

evant if it entails the LLM zero-shot answer. The training data for fine-tuning LLM use top-1 document returned by the retriever as positive document, and low-ranked or random documents as negative documents.

InstructRAG-FT (Wei et al., 2024). The method prompts LLM to generate explicit denoising rationales for deriving ground truth answers from retrieved documents, and then uses these rationales as demonstrations for in-context learning, or supervised data for LLM fine-tuning.

A.3 Implementation details

We set the learning rate to 1e-5, with a warm-up ratio of 0.1 and a cosine scheduler. The dropout probability is set to 0.1. Both fine-tuning and inference are implemented on the basis of the Llama Factory[7] framework. Consistent with Wei et al. (2024), the maximum token length is set to 4096. All the results we report are for a single run.

In the ASQA evaluation, we employ the metrics and methods provided by ALCE[8] (Gao et al., 2023), which uses the TRUE[9] (Honovich et al., 2022) model to evaluate the precision and recall of citations.

B Case Study

Table 6 shows an example where none of the top-5 retrieved documents contain useful information. Since the documents in the Wikipedia dump corpus are truncated, incomplete statements may cause the model to misinterpret the context. For example, the information of Mad Hatter player in Document 1 is truncated and the document does not actually contain the correct answer. However, vanilla RAG and InstructRAG-FT identify Document 1 as relevant to the question, but mistakenly assume that "David Warner", which is adjacent to the keyword "Mad Hatter", is the answer. Our FPA correctly identifies the absence of useful information in Document 1 and gives an answer based on parameter knowledge.

Table 7 shows an example where the top-5 documents contain useful information. Vanilla RAG fails to recognize that Document 4 contains relevant information and incorrectly treats "Emile Berliner" in Document 1 as the correct answer. InstructRAG-FT similarly ignores the key information in Document 4 and mistakes "Alexander Graham Bell" as the inventor. In contrast, FPA focuses on the information "Thomas Edison invented the phonograph" in Document 3 and combines it with internal knowledge to give the correct conclusion.

C Prompt Templates

In the *w/o Retrieval* setting, no retrieval documents are provided in the context and LLM answers directly based on internal knowledge. The prompt we use is shown in Table 8.

In the *RAG* setting, the top-5 retrieved documents is provided to LLM. The initial prompt are the non-italic, non-bold parts of Table 9. In our FPA, we add adaptive retrieval prompting (highlighted in **bold**) to the initial prompt for inference. To construct preference data, we sample LLM responses based on the initial prompt for RAG. To evaluate citation quality on ASQA, since neither our method nor the base model is trained to give citations in the response at any time, we add additional instructions (highlighted in *italics*) for LLM to output accurate citations.

In the preliminary experiments in Section 3.2, we use LLM to score the answerability and factual consistency of retrieved documents using the prompts shown in Tables 10 and 11, respectively.

[7] https://github.com/hiyouga/LLaMA-Factory

[8] https://github.com/princeton-nlp/ALCE

[9] https://huggingface.co/google/t5_xxl_true_nli_mixture

Question Who played the Mad Hatter in the Batman TV show? **Answers** Roddy McDowall, David Wayne, Benedict Samuel **Context** Document 1 (Title: Batman: The Animated Series): as the Mad Hatter, David Warner as Ra's al Ghul, Michael York as Count Vertigo, Kate Mulgrew as Red Claw, George Murdock as and George Dzundza as the Ventriloquist. One of the series' best-known inventions is the Joker's assistant, Harley Quinn, who became so popular that DC Comics later added her to mainstream Batman comic book continuity. ... Document 2 (Title: Batman (TV series)): ... Document 3 (Title: Batman (TV series)): ... Document 4 (Title: Batman (TV series)): ... Document 5 (Title: Batman (TV series)): ...
Vanilla RAG According to Document 1, the Mad Hatter was played by *David Warner* in the Batman: The Animated Series. (✗)
InstructRAG-FT After analyzing the provided documents, I found that Document 1 is the most relevant to answer the question "who played the Mad Hatter in the Batman TV show?". Document 1 mentions the voice actors for various characters in the Batman: The Animated Series, including the Mad Hatter, played by David Warner. The other documents do not provide information about the Mad Hatter or the Batman TV show, as they focus on different aspects of the Batman franchise, such as the 1960s TV series, the 1990s animated series, and the character's comic book history. Therefore, based on Document 1, the answer to the question is: *David Warner played the Mad Hatter in the Batman: The Animated Series.* Note that the other documents do not provide any information about the Mad Hatter or the Batman TV show, and therefore are not relevant to answering the question. (✗)
FPA (Ours) **The provided documents do not directly answer who played the Mad Hatter in the Batman TV show.** However, according to my knowledge, the Mad Hatter was played by Malkovich in the 1990s Batman: The Animated Series (not mentioned in the provided documents). If we were to look at other adaptations, the Mad Hatter has been portrayed by various actors, including **David Wayne** in the 1966 Batman TV series (not mentioned in the provided documents) and Anthony Newley in the 1966 film "Batman: The Movie" (also not mentioned in the provided documents). It's worth noting that the provided documents do mention other characters and actors from various Batman adaptations, but not the Mad Hatter specifically. (✓)

Table 6: An example where none of the top-5 retrieved documents contain useful information. We highlight wrong answers in *italics* and correct answers in **bold**.

Question Who made possible the first reproduction of the human voice on the phonograph?
Answers Thomas Edison, Thomas Alva Edison
Context
Document 1 (Title: History of sound recording): this idea, and it was not until 1887 that yet another inventor, Emile Berliner, actually photoengraved a phonautograph recording into metal and played it back.
...
Document 2 (Title: Gianni Bettini): ...
Document 3 (Title: Phonograph): Alexander Graham Bell and his two associates took Edison's tinfoil phonograph and modified it considerably to make it reproduce sound from wax instead of tinfoil.
Document 4 (Title: Phonograph record): Along with a tuning fork tone and unintelligible snippets recorded as early as 1857, these are the earliest known recordings of sound. In 1877, Thomas Edison invented the phonograph. Unlike the phonautograph, it could both record and reproduce sound.
Document 5 (Title: Sound recording and reproduction): ...

Vanilla RAG According to the provided information, it was *Emile Berliner* who made possible the first reproduction of the human voice on the phonograph in 1887. He photoengraved a phonautograph recording into metal and played it back, making it possible to reproduce sound. (✗)

InstructRAG-FT The documents that are useful to answer the question "who made possible the first reproduction of the human voice on the phonograph?" are: * Document 1: History of sound recording * Document 4: Phonograph record * Document 5: Sound recording and reproduction. ... However, Document 3: Phonograph provides the crucial information needed to answer the question. It states that Alexander Graham Bell and his associates modified Edison's tinfoil phonograph to make it reproduce sound from wax instead of tinfoil, and that they began their work at Bell's Volta Laboratory in Washington, D.C. in 1879. This information suggests that Bell and his associates were involved in the development of the phonograph and may have made possible the first reproduction of the human voice on the phonograph. Based on this information, it can be concluded that *Alexander Graham Bell* and his associates made possible the first reproduction of the human voice on the phonograph. (✗)

FPA (Ours) Based on the provided information, it is not explicitly stated who made possible the first reproduction of the human voice on the phonograph. **However, according to documents 4 and 3, Thomas Edison invented the phonograph in 1877**, and Alexander Graham Bell and his associates modified Edison's tinfoil phonograph to record in wax in 1879. According to my knowledge, Thomas Edison is credited with inventing the first device that could both record and reproduce sound, which is known as the phonograph. He demonstrated the first recording and playback of sound on July 18, 1877, using a cylinder covered in tinfoil. Therefore, based on my knowledge and the provided information, I would answer that **Thomas Edison** made possible the first reproduction of the human voice on the phonograph. (✓)

Table 7: An example where the top-5 documents contain useful information. We highlight wrong answers in *italics* and correct answers in **bold**.

Based on your knowledge, answer the question:
{question}

Table 8: Prompt templates for generation without retrieval.

Answer the given question and cite the index of all references used in your answer. Ensure both the answer and citations are accurate.

Document 1 (Title: ...): {contents}
...
Document 5 (Title: ...): {contents}

Based on your knowledge and the provided information, answer the question:
{question} **If the provided documents are irrelevant or do not contain complete information to answer the question, supplement and answer with your knowledge.**

Table 9: Prompt templates for RAG. *Italic* text is used for ASQA evaluation and **bold** text is adaptive retrieval prompting.

Read the following document:

Document (Title: ...): {contents}

For the given question: '{question}', can we derive the answer to the question from the above document? If your answer is yes, output 1, otherwise output 0. Your output should only include 0 or 1.

Table 10: Prompt templates for document answerability scoring by LLM.

Read the following document:

Document (Title: ...): {contents}

Based on the above document, can we conclude that '{ground truth statement}'? If your answer is yes, output 1, otherwise output 0. Your output should only include 0 or 1.

Table 11: Prompt templates for document factual consistency scoring by LLM.