RbFT: Robust Fine-tuning for Retrieval-Augmented Generation against Retrieval Defects

Yiteng Tu
DCST, Tsinghua University
Beijing, China
yitengtu16@gmail.com

Weihang Su
DCST, Tsinghua University
Beijing, China
swh22@mails.tsinghua.edu.cn

Yujia Zhou
DCST, Tsinghua University
Beijing, China

Yiqun Liu
DCST, Tsinghua University
Beijing, China

Qingyao Ai*
DCST, Tsinghua University
Beijing, China
aiqy@tsinghua.edu.cn

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) by integrating external knowledge retrieved from a knowledge base. However, its effectiveness is fundamentally constrained by the reliability of both the retriever and the knowledge base. In real-world scenarios, imperfections in these components often lead to the retrieval of noisy, irrelevant, or misleading counterfactual information, ultimately undermining the trustworthiness of RAG systems.

To address this challenge, we propose Robust Fine-Tuning (RbFT), a method designed to enhance the resilience of LLMs against retrieval defects through two targeted fine-tuning tasks. Experimental results demonstrate that RbFT significantly improves the robustness of RAG systems across diverse retrieval conditions, surpassing existing methods while maintaining high inference efficiency and compatibility with other robustness techniques.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language generation**.

Keywords

Retrieval-augmented Generation, Fine-tuning, Robust

1 Introduction

Large Language Models (LLMs) have achieved exceptional performance across diverse natural language processing tasks [29, 50], yet they remain constrained by challenges such as hallucinations, outdated or incomplete knowledge, and limited adaptability to specialized domains [19, 24]. Retrieval-augmented generation (RAG) has emerged as a key technique to address these limitations by integrating LLMs with external knowledge sources, enabling enhanced factual accuracy, up-to-date information access, and improved domain-specific performance [17, 26, 44]. Due to its effectiveness, RAG has been widely adopted to provide LLMs with a flexible mechanism for knowledge augmentation, enhancing their performance in varying scenarios [5, 14].

Despite their popularity, existing RAG systems suffer from a critical challenge: the performance of RAG systems heavily relies on the quality of the information provided by the retriever [25, 45, 54]. Since the real-world retriever and its corresponding knowledge base

could be defective and imperfect [7, 36], the retrieved documents provided to the LLM may contain inaccurate, irrelevant, or even malicious and misleading information [4, 48]. Such low-quality or harmful information may hinder the LLM from accessing accurate knowledge, leading to inaccurate or misleading responses [53, 54, 59], significantly degrading the performance and reliability of RAG systems (as discussed in §3.2). Therefore, an imperative challenge arises: *how to enhance the robustness of LLMs and RAG systems against retrieval defects, ensuring their ability to generate accurate responses even with defective retrieval results?*

Although there are several studies working on the robustness of RAG systems against retrieval defects through designing sophisticated inference mechanisms [48, 51–54], they are limited in several perspectives. First, a critical issue is the significant increase in inference costs, which severely limits the RAG pipeline's runtime efficiency. This issue arises because existing studies mostly require generating intermediate results, which the LLM must aggregate or evaluate to produce the final response. Second, more importantly, these methods tend to fail under severe retrieval defects or other challenging conditions since they do not fundamentally address the LLM's dependency on input knowledge. For instance, RobustRAG [53] follows an "isolate-then-aggregate" pipeline, where answers are independently generated for each retrieved document and then aggregated. This approach not only incurs high inference costs but also becomes ineffective when the proportion of negative documents is high. CRAG [54], on the other hand, leverages the large-scale web search to supplement and rely on the vanilla LLM to integrate and refine knowledge from different sources. However, when the vanilla LLM fails to identify the defects in retrieved results, the whole pipeline would be broken and ineffective.

Based on the identified shortcomings of the aforementioned approaches, we believe that building an efficient and robust RAG system requires us to improve the inherent defensive capabilities of LLMs to fundamentally reduce their over-reliance on and blind trust in input information. Specifically, a well-defended LLM should have two characteristics: (1) **Defects Detection**: it should be capable of distinguishing what kind of information facilitates an effective response to the user's query and which documents are irrelevant or even harmful; (2) **Utility Extraction**: it should effectively utilize the limited useful information provided by the retriever while ignoring irrelevant or harmful content, even under adverse retrieval defects. Therefore, we propose two corresponding fine-tuning tasks

aimed at strengthening the LLM's overall defensive capabilities: *Defects Detection* and *Utility Extraction*, collectively referred to as **Ro**bust **F**ine-**T**uning (RbFT). Specifically, we replace the original retrieved documents with defective ones and then train the LLM to (1) determine whether each document contains defects, thereby enhancing its ability to assess inputs critically; (2) generate the correct answers based on the defective inputs, improving its capability to utilize useful information effectively. Experimental results demonstrate that our fine-tuning tasks can deliver superior performance in extremely challenging retrieval conditions, significantly outperforming the state-of-the-art baseline methods.

In summary, this paper makes three key contributions: (1) We conduct a comprehensive analysis of potential retrieval defects in RAG systems from the perspectives of the retrieval system (i.e., the retriever and the corpus) and find that LLMs are highly vulnerable to retrieval defects. (2) We propose RbFT, a set of fine-tuning tasks that improve LLMs' ability to evaluate and utilize retrieved information, enhancing robustness against retrieval defects. (3) We conduct extensive experiments to show that RbFT can achieve superior performance under challenging retrieval conditions, significantly outperforming existing state-of-the-art methods.

2 Related Work

2.1 Retrieval-Augmented Generation

In recent years, Retrieval-Augmented Generation (RAG) has garnered widespread attention in the field of natural language processing (NLP) and shown significant advantages in knowledge-intensive tasks [2, 9, 14, 18, 20, 24, 39, 46, 47]. Traditional RAG typically follows the "Retrieval-then-Read" framework [2, 14, 19, 24], where an external retriever [28, 34, 37, 38, 57] or a complex retrieval system [35, 40] is adopted to search for relevant documents from a large-scale external corpus based on the user's query. The retrieved documents provide external knowledge that supplements the query, allowing the generative model to incorporate relevant information beyond its parametric knowledge when generating a response. To further enhance the retrieval effectiveness, researchers have introduced additional techniques such as query rewriting [8, 27] and re-ranking [1, 13] to refine the quality of retrieved documents before appending them to the generative model.

Building upon the traditional RAG framework, various extensions have been proposed to enhance its effectiveness and efficiency. One such extension, Parametric RAG [43], directly injects the retrieved documents into LLM parameters by offline parameterizing each document into independent plug-in parameters. During the inference process, the retrieved document's parametric representation is merged and integrated into the LLM, enabling knowledge injection without extending the input context. From another angle, GraphRAG [12, 16, 32] leverages pre-constructed knowledge graphs to retrieve graph elements with relational knowledge relevant to a given query. This approach has shown improved performance, especially in tasks that rely on structured and relational information. Another research direction, dynamic RAG [20, 41, 42], dynamically triggers the retrieval module during the generation process when the LLM exhibits high uncertainty during the generation process.

In contrast to existing works that focus on improving retrieval quality, refining retrieval pipelines (e.g., through query rewriting or re-ranking), or reorganizing knowledge representation, we propose a fundamentally different perspective: directly enabling the LLM itself to handle imperfect or even malicious retrieval results. Instead of assuming perfectly relevant or pre-filtered documents, we train the model to detect flaws in retrieved texts and extract only useful evidence, mitigating the impact of noisy, irrelevant, or incomplete information. This shift not only enhances accuracy but also fosters a more resilient and trustworthy RAG framework.

2.2 Robustness in RAG

The robustness of RAG systems refers to the ability of LLMs to consistently extract and apply relevant knowledge, even when exposed to varying or defective retrieval inputs [60]. Existing works have found that misinformation and corruption retrieval inputs pose significant challenges to the robustness of RAG systems. Adversarial Addition and Modification [10] demonstrates the vulnerability of automated fact-checking systems when confronted with synthetic adversarial evidence. Pan et al. [30] and Pan et al. [31] explore the threat posed by misinformation (whether manually crafted or generated by LLMs) to open-domain question-answering (ODQA) systems, highlighting the vulnerability of these systems when exposed to misinformation corruption. By injecting malicious texts into the knowledge base, PoisonedRAG [61], GARAG [6] and Phantom [3] can manipulate LLMs into generating specific incorrect or harmful responses. To address these vulnerabilities, researchers have proposed strategies focusing on input optimization and knowledge integration. Weller et al. [52] conducts query augmentation and introduces a novel confidence method based on answer redundancy. RobustRAG [53] employs an isolate-then-aggregate strategy to ensure the robustness of LLM responses against retrieval corruption attacks. By generating self-synthesized rationales, InstructRAG [51] explicitly denoises the retrieved content, thereby enhancing the robustness of RAG systems. CRAG [54] and AstuteRAG [48] turn to refine and integrate knowledge derived from different sources to improve knowledge utilization and enhance the robustness of the generated answer. However, despite the progress achieved by these methods, as discussed in §1, they can only partially control the retrieved content that the LLM accesses in RAG systems and fail to address the core issue of LLMs' excessive reliance on the retrieval inputs. Unlike these works, we focus on enhancing the inherent defensive capabilities of LLMs to mitigate the impact of retrieval defects. By reducing dependency on external retrieval, our approach fundamentally improves RAG system robustness, offering a more resilient and trustworthy framework for real-world applications.

3 Task Formulation

In this section, we first formalize the workflow of the RAG system and then discuss three types of potential retrieval defects that may occur in the RAG process.

3.1 Workflow of RAG

Following previous works [24, 54], a vanilla RAG system typically consists of a retrieval component $\mathcal{R}$, a generation component (i.e., the LLM) $\mathcal{G}$, and a corresponding corpus $C = \{d\}$ containing a large collection of knowledge documents. Whenever the system receives a user query q, the retrieval component $\mathcal{R}$ first retrieves
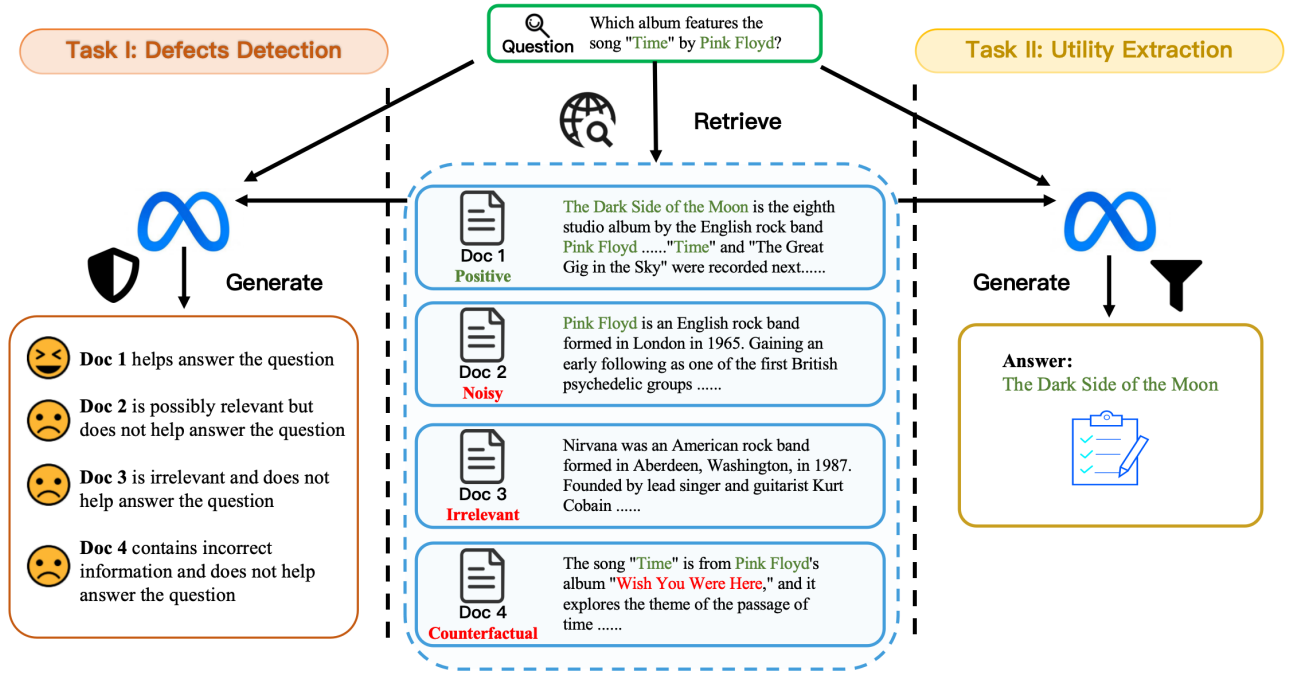
Figure 1: Overview of our RbFT. Specifically, RbFT consists of two sub-tasks: Defects Detection and Utility Extraction, which aim to identify the types of retrieval defects and generate the final answer with limited useful information, respectively. In the figure, green text indicates relevant information, while red text represents incorrect counterfactual information.

the top-k most relevant documents $\mathcal{D}^q = \{d_1^q, d_2^q, ..., d_k^q\} \subset C$ from the corpus C:

$$\mathcal{D}^q = \mathcal{R}(q, C) \tag{1}$$

Then, the LLM $\mathcal{G}$ generates a response r based on the query q and relevant documents $\mathcal{D}^q$, where the expected output r should ideally match the ground-truth answer a. Thus, the entire workflow can be formalized as:

$$r = \mathcal{G}(q, \mathcal{D}^q) = \mathcal{G}(q, \mathcal{R}(q, C)) \tag{2}$$

It is evident that, aside from the understanding and generation capabilities of the LLM itself, the quality of the generated response r is highly dependent on the capability of the retriever $\mathcal{R}$ along with the quality of the corpus C. Either an underperforming retriever or a low-quality corpus can significantly degrade the response quality. Since these retrieval-side issues are unavoidable in real-world scenarios, our work focuses on enhancing the capability of the generation component $\mathcal{G}$ to minimize their negative impact.

3.2 Retrieval Defects

As mentioned above, due to the limitations of the retriever's performance and the quality of the corpus, the retrieval component often cannot guarantee that all returned documents fully meet the user query's information needs, resulting in various types of retrieval defects. These defects can be broadly categorized into three types:

3.2.1 Noisy Documents. Noisy documents refer to content that is relevant to the query topic but does not directly answer the query. For example, given the query in Figure 1: "*Which album features*

the song 'Time' by Pink Floyd?", the retrieval system might return a general overview of the band *Pink Floyd* (Doc 2 in Figure 1). Although such a document is related to the band and its music, it does not explicitly mention the album containing the song "*Time*", thereby failing to address the core question.

3.2.2 Irrelevant Documents. Irrelevant documents are those that bear no connection to the query topic. Such documents are typically retrieved due to inaccuracies in the retrieval model's judgment. For instance, in response to the same query, the system might retrieve a document introducing another band like *Nirvana* (Doc 3 in Figure 1). While about music, it has no relevance to *Pink Floyd* or its albums, making it a clear example of an irrelevant document.

3.2.3 Counterfactual Documents. Counterfactual documents consist of false or misleading information, often resulting from inaccuracies or malicious manipulation within online content. They fail to answer the question and may even lead to misconceptions. For example, while the correct answer to the query is "*The Dark Side of the Moon*", a counterfactual document might falsely claim that the song "*Time*" appears on another album "*Wish You Were Here*" (Doc 4 in Figure 1). Such incorrect information undermines the reliability of the retrieval process and can mislead both users and LLMs.

To simulate different levels of retrieval defects, we use varying probabilities τ to randomly replace the original retrieved documents with defective documents of different or identical types. Let the
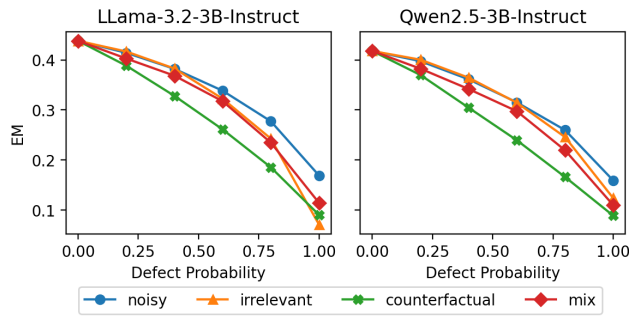
Figure 2: Emperical study: the impact of different types of retrieval defects on Vanilla RAG. The average EM metric on NQ, HQA, and TQA datasets is reported.

modified retrieval result be denoted as $\mathcal{D}_\tau^q$, our goal is for the LLM to generate the correct answer a even when provided with $\mathcal{D}_\tau^q$.

To validate the negative impact of these retrieval defects on RAG systems, we conduct preliminary experiments on three datasets: Natural Questions [23], HotpotQA [56], and TriviaQA [21], and report the average results. We select e5-base-v2 [49] as the retriever, returning the top-5 retrieved results, while LLama-3.2-3B-Instruct [11] and Qwen2.5-3B-Instruct [55] are adopted to generate the final answers, which are measured using the exact match (EM) metric. The tests are conducted using the aforementioned three types of retrieval defects and their mixture (i.e., randomly selecting one type of defective document for replacement). The defect replacement probability τ is set to {0, 0.2, 0.4, 0.6, 0.8, 1.0}. As observed in Figure 2, various types of retrieval defects pose a significant threat to the reliability of the vanilla RAG system. Specifically, misleading counterfactual documents have the greatest impact on RAG systems, while even noisy documents that are relatively less harmful can be quite disruptive. When all input documents are noisy, accuracy drops to below 40% of the original. Therefore, the severe impact of retrieval defects on the reliability of RAG systems underscores the urgent need to enhance their robustness.

4 Robust Fine-Tuning (RbFT)

As mentioned in the §1, since the performance of an RAG system heavily depends on the quality of the retrieved documents and the model's ability to effectively utilize them, we aim to enhance the robustness of the overall RAG system by fine-tuning the LLM to strengthen its intrinsic defensive capabilities. Here, the robustness is reflected not only in the model's ability to remain stable when confronted with retrieval defects or low-quality inputs but, more importantly, in its capacity to extract and utilize useful information effectively. Specifically, we believe that a well-defended LLM should possess the following two key characteristics:

- **Ability to assess the quality of input content.** A robust LLM should effectively distinguish the quality of documents, determining which documents are genuinely helpful in addressing the question and which are useless. On the one hand, distinguishing the input documents helps LLMs develop critical thinking and reduces their dependency on the retrieved results. On the other

hand, accurate identification of useful information is also critical to prevent retrieval defects from affecting the final output.
- **Ability to fully utilize useful information in the overall context.** Once the model has developed an initial ability to assess input quality, it should be able to extract and exploit key information from high-quality and useful content while filtering out irrelevant or misleading content. Besides, the model should not only be capable of information filtering but also be able to synthesize multi-source information during answer generation to ensure the accuracy and completeness of the output.

Therefore, the core of our strategy lies in equipping LLMs with stronger self-detection and extraction capabilities, enabling them to maintain efficient and accurate outputs in complex real-world scenarios. To achieve this goal, we design two specialized training tasks, namely *Defect Detection* and *Utility Extraction*, corresponding to input content assessment and effective information filtering, respectively (as shown in Figure 1). The joint training of these tasks enables the LLM to improve its resistance to interference in complex input environments, thereby enhancing the overall robustness of the RAG system.

4.1 Task I: Defects Detection

The Defect Detection task aims to train the LLM to identify whether each retrieved document contributes to answering the user's query. If a document is useless, the LLM must also classify it into one of three defect types, i.e., noisy, irrelevant, or counterfactual document. We treat the original retrieved documents as positive examples and randomly replace them with different types of defective documents at a probability of τ. To improve the efficiency, we adopt a listwise input format, where the LLM evaluates the entire list of retrieved documents at once. The prompt for this task is as follows:

Input:
Determine whether the following documents help answer the given question. The assessment includes:
Assessment 1: The document helps answer the question.
Assessment 2: The document is possibly relevant but does not help answer the question.
Assessment 3: The document is irrelevant and does not help answer the question.
Assessment 4: The document contains incorrect information and does not help answer the question.
Only give me your assessment for each document and do not output any other words.
Documents:
Doc 1: { document 1 }
Doc 2: { document 2 }
......
Question: { question }

Output:
Doc 1 helps answer the question. / Doc 1 is possibly relevant but does not help answer the question. / Doc 1 is irrelevant and does not help answer the question. / Doc 1 contains incorrect information and does not help answer the question.
Doc 2 ......

4.2 Task II: Utility Extraction

In the Utility Extraction task, we aim to train the LLM to extract as much useful information as possible from the defective retrieval result. The LLM can either directly utilize the extracted relevant information or leverage the relevant context to activate its internal parametric knowledge to generate the correct answer. Meanwhile, Utility Extraction training also enables the LLM to directly and efficiently handle low-quality or contaminated contexts without prior cleanup. Similarly, the original documents are replaced with defective documents at a probability τ, and the LLM is required to answer based on these defective retrieval results while producing correct outputs. The prompt used for this task is as follows:

Input:
Answer the question based on the given document. Only give me the answer and do not output any other words.
The following are given documents.
Doc 1: { document 1 }
Doc 2: { document 2 }
......
Question: { question }

Output:
{ answer }

4.3 Training Objective

RbFT aims to directly fine-tune the LLM using the two afore-mentioned tasks, thereby enhancing its organic defense capability. To achieve efficient training while preserving the LLM's general-purpose capabilities, we adopt the Low-Rank Adaptation (LoRA) [15] technique for fine-tuning. Specifically, given the input text x, our goal is to maximize the probability of producing the correct output text y:

$$\max_{\Theta} \sum_{(x,y)} \sum_{i=1}^{|y|} \log \left(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_i | y_{<i}, x) \right) \tag{3}$$

where Φ_0 and $\Delta\Phi(\Theta)$ denote the LLM's original parameters and the learned parameter adjustments during fine-tuning, respectively.

5 Experimental Settings

5.1 Datasets and Evaluation Metrics

We conduct experiments on three widely used Question Answering (QA) datasets: Natural Questions (NQ) [23], HotpotQA (HQA) [56], and TriviaQA (TQA) [21], which cover both factoid QA and multi-hop QA tasks. Each data instance consists of a query and its corresponding ground truth answer. To create the training and validation sets, we randomly sample a total of 20,000 instances from the training splits of these three datasets, where 10% of the training data is reserved for validation. For evaluation, to ensure efficient experiments, following [48, 53], we sample 1,000 queries from the test sets of NQ and TQA, as well as the HQA validation set (since HQA does not provide the test set), respectively (i.e., a total of 3,000 test queries). The e5-base-v2 [49] retriever is adopted to retrieve the top 100 most relevant documents for each query from the Wikipedia

corpus [1]. To assess the performance of different RAG systems under varying levels of retrieval defects, we employ the standard QA evaluation metrics: exact match (EM) and token-level F1 score (F1), which measure the precision of the generated answers.

5.2 Baselines

Our RbFT is primarily compared with No RAG, Vanilla RAG, as well as four state-of-the-art robustness approaches for the RAG system: RobustRAG [53], CRAG [54], InstructRAG [51] and AstuteRAG [48]. RobustRAG leverages an "isolate-then-aggregate" strategy, where the LLM independently generates responses for each retrieved passage and then aggregates these individual responses to produce the final output. CRAG introduces a lightweight retrieval evaluator that triggers different knowledge retrieval actions based on evaluation results and enables knowledge refinement. To ensure a fair comparison of different RAG systems, we disable the module in CRAG that is responsible for large-scale web searches to acquire additional knowledge. InstructRAG instructs LLMs to explicitly denoise retrieved content by generating self-synthesized explanatory rationales, which explain how the answer is derived from the retrieved documents. AstuteRAG, on the other hand, focuses on resolving conflicts between the internal knowledge of the LLM and the external knowledge provided by the retriever. It achieves this goal through an "iterative source-aware knowledge consolidation" process that integrates the two kinds of knowledge and handles knowledge conflicts.

5.3 Data Generation

We simulate three types of defective documents using different approaches. For noisy and irrelevant documents, inspired by the negative sampling methods commonly used in training dense retrieval models [22, 33, 58], these two types of defective documents can be analogized to hard negatives and random negatives, respectively. Accordingly, noisy documents can be obtained by randomly sampling from lower-ranked retrieval results (e.g., documents ranked after 50 in the retrieval results), while irrelevant documents can be randomly sampled from the entire corpus. For counterfactual documents, we adopt a two-step generation strategy: first, given the query, the correct answer, and the original retrieval results, we use Llama-3.2-3B-Instruct [11] to generate a misleading incorrect answer. Then, we call the LLM again to rewrite all original documents by replacing all information related to the correct answer with the misleading incorrect answer:

Step 1 Input:
Based on a given question and its correct answer, generate a misleading wrong answer. You can refer to some relevant documents for inspiration. The wrong answer should belong to the same entity type as the correct answer (e.g., person, time, place, organization, data, etc.) to enhance its confusion. If the answer does not contain an entity, replace a key entity in the question and treat it as the wrong answer. Only give me the wrong answer and do not output any other words.
The following are given documents.
Doc 1: { document 1 }
Doc 2: { document 2 }
......
Question: { question }
Correct Answer: { answer }

Step 2 Input:
You are a writing AI. Rewrite the passage by replacing all content and information related to { correct answer } with { wrong answer }. Ensure that the rewritten passage is fluent and concise, maintaining a language style similar to the original. Only give me the rewritten passage and do not output any other words.
Original Document: { document }

5.4 Implementation Details

We fine-tune two LLMs on the RbFT task, Llama-3.2-3B-Instruct [11] and Qwen2.5-3B-Instruct [55], to enhance their robustness against retrieval defects through the LLaMA-Factory toolkit [2]. In the following text, we refer to these two LLMs as Llama and Qwen for convenience. The fine-tuning is conducted for 2 epochs, with a learning rate of 1e-5, and a per-device batch size of 16, setting $lora\_rank = 16$ and $lora\_alpha = 64$. LLMs are fine-tuned on four types of defective data: *Noisy*, *Irrelevant*, *Counterfactual*, and *Mix* (randomly selected from the first three types), with the probability of replacing original retrieval results with defective documents (τ) selected from {0.2, 0.4, 0.6, 0.8, 1.0}. During the evaluation phase, the same τ values are used, with particular attention given to $\tau = 0.4$ and $\tau = 1.0$, referred to as the *Normal* and *Hard* settings, respectively, representing moderate and severe retrieval defects. Additionally, to compare the original performance, we also report their results with the original retrieval results ($\tau = 0$), referred to as the *Clean* setting. The retrieval list size k is set to 5. Our code and data are available at the URL [3].

6 Results and Analysis

6.1 Main Results

Table 1 shows the performance of all methods under different levels of retrieval defects. We observe that:

(1) **In the *Clean* setting, RbFT is the only method that surpasses Vanilla RAG.** Unlike other approaches that experience performance degradation in defect-free environments, RbFT can

---

[2]https://github.com/hiyouga/LLaMA-Factory
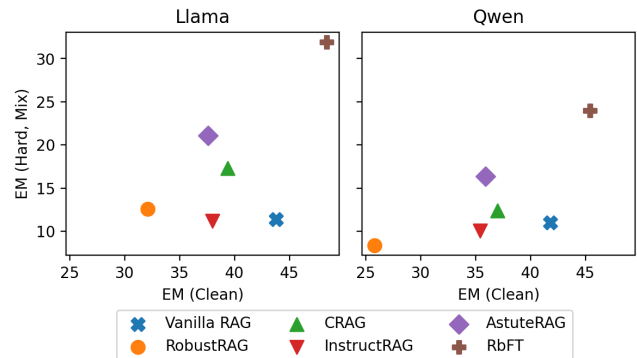[3]https://github.com/StibiumT16/Robust-Fine-tuning

Figure 3: The effectiveness-robustness trade-off scatter diagram. The x-axis represents effectiveness measured by the EM scores of each model in the Clean setting, and the y-axis represents robustness measured by the EM scores of each model in the Hard + Mix setting.

enhance robustness while maintaining or even improving performance in such scenarios. For example, when using Qwen as the base model, RbFT achieves a 12.0% improvement in F1 score compared to Vanilla RAG. This advantage may be attributed to the fact that the top-ranked results returned by the retriever are rarely flawless and may contain a certain proportion of defective documents (i.e., false positives). RbFT actively identifies and adapts to these potentially defective documents, thereby preventing performance degradation.

(2) **In the *Normal* setting, RbFT consistently achieves the best performance across all retrieval defect scenarios and is still the only method that significantly outperforms Vanilla RAG.** Notably, RbFT shows the most substantial improvement in the counterfactual defect scenario: when using the Llama model, the EM metric improves by 37.3% compared to the second-best method, while using the Qwen model also results in a 32.9% improvement, far exceeding other approaches. In other defect scenarios, RbFT also demonstrates strong superiority, with improvements in both metrics mainly ranging between 15% and 20%.

(3) **In the *Hard* setting, RbFT continues to outperform all other methods and further widens the gap with the second-best approach**, highlighting its exceptional performance and adaptability in extremely adverse retrieval environments. Traditional methods often exhibit instability when handling these complex issues in such high-difficulty scenarios, where the retrieval results consist entirely of defective documents. However, RbFT consistently maintains high performance, particularly in the Counterfactual scenario, where using the Llama model results in an EM metric improvement of over 70%. This significant advantage further validates RbFT's robustness and reliability in dealing with various complex retrieval scenarios.

(4) **RbFT demonstrates significant advantages in balancing effectiveness and robustness.** We refer to the EM score under the *Clean* setting as a method's organic capability in QA tasks (i.e., effectiveness), and refer to the EM score under the *Hard+ Mix* defective setting as the method's ability to handle complex retrieval defects (i.e., robustness). Based on these two metrics, we plot the

Table 1: The average evaluation results of each model on the three datasets under the Clean ($\tau = 0$), Normal ($\tau = 0.4$), and Hard ($\tau = 1.0$) settings. "*" refers to a significant improvement compared to the Vanilla RAG baseline at $p < 0.05$ level using the two-tailed pairwise t-test. The best and second-best methods are marked in bold and underlined, respectively. The improvement ratio of the best model over the second-best model is also reported.

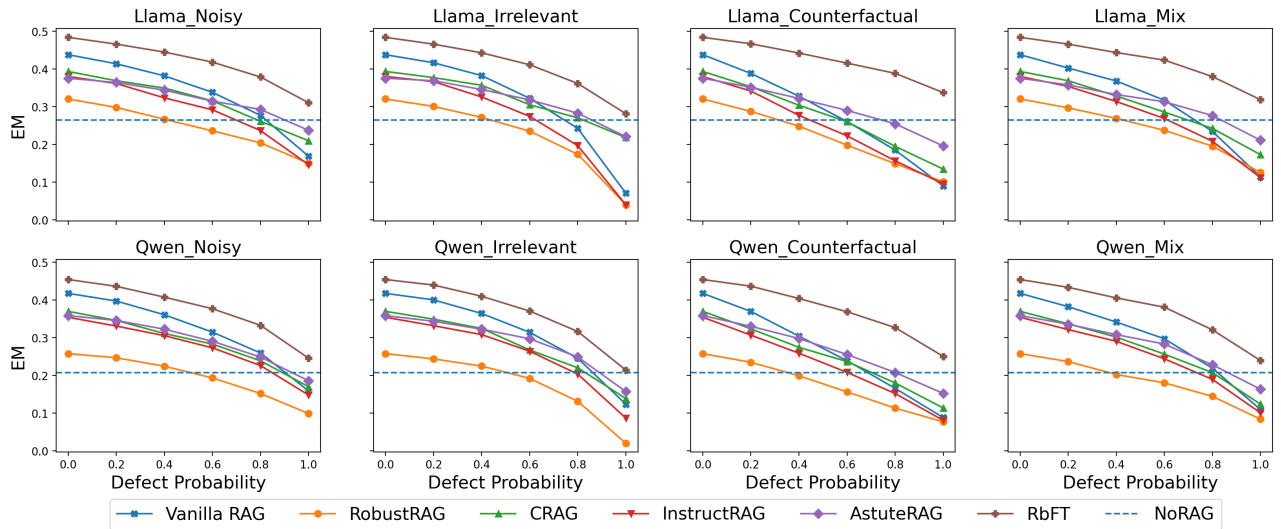| LLM | Method | Clean ($\tau = 0$) | | Normal ($\tau = 0.4$) | | | | | | | | Hard ($\tau = 1.0$) | | | | | | | |
---	---	---	---	Noisy		Irrelevant		Counterfactual		Mix		Noisy		Irrelevant		Counterfactual		Mix	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Llama	No RAG	26.5	34.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Vanilla RAG	<u>43.8</u>	<u>52.7</u>	<u>38.2</u>	<u>46.2</u>	<u>38.3</u>	<u>46.3</u>	<u>32.8</u>	40.9	<u>36.8</u>	<u>45.2</u>	16.9	22.0	7.1	9.5	9.0	15.4	11.4	17.1
	RobustRAG	32.1	43.5	26.7	37.0	27.2	37.6	24.9	35.2	26.9	37.5	15.0	22.2	4.0	5.8	10.1	18.4*	12.6	19.7*
	CRAG	39.4	48.5	35.0	43.1	35.7	43.6	30.4	39.3	32.9	41.6	21.0*	27.4*	21.9*	27.6*	13.5*	20.7*	17.3*	23.8*
	InstructRAG	38.0	47.6	32.4	41.3	32.6	41.3	27.7	36.6	31.4	40.2	14.7	21.3	4.0	8.3	9.6	16.6*	11.2	17.2
	AstuteRAG	37.6	47.5	34.4	43.8	34.6	43.7	32.2	<u>41.5</u>	33.2	42.7	<u>23.8*</u>	<u>32.1*</u>	<u>22.1*</u>	<u>30.2*</u>	<u>19.6*</u>	<u>28.4*</u>	<u>21.1*</u>	<u>29.6*</u>
	RbFT (Ours)	**48.4***	**58.5***	**44.5***	**53.9***	**44.3***	**53.7***	**44.2***	**54.2***	**44.4***	**54.0***	**31.1***	**39.1***	**28.2***	**36.4***	**33.8***	**43.1***	**31.9***	**40.9***
	Improvement	10.5%	11.0%	16.5%	16.7%	15.7%	16.0%	37.3%	29.4%	20.7%	19.5%	30.7%	21.8%	27.6%	20.5%	72.4%	51.8%	51.2%	38.2%
Qwen	No RAG	20.8	27.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Vanilla RAG	<u>41.8</u>	<u>50.7</u>	<u>36.1</u>	<u>44.3</u>	<u>36.4</u>	<u>44.6</u>	<u>30.4</u>	38.6	<u>34.2</u>	<u>42.6</u>	15.9	21.6	12.4	15.5	8.9	15.4	11.0	16.9
	RobustRAG	25.8	37.6	22.5	33.0	22.5	33.2	19.9	29.9	20.3	31.3	9.9	15.7	2.0	3.3	7.7	14.9	8.4	14.7
	CRAG	37.0	45.5	31.1	39.0	32.5	40.0	27.5	35.3	30.2	38.4	17.0*	22.6	13.9*	17.2*	11.4*	17.6*	12.4*	18.0
	InstructRAG	35.4	46.6	30.6	40.5	30.8	40.6	25.9	35.9	29.1	38.9	14.8	20.9	8.7	13.1	8.1	15.5	10.1	16.6
	AstuteRAG	35.9	46.3	32.3	41.9	32.3	41.6	29.8	<u>39.2</u>	30.8	40.4	<u>18.6*</u>	<u>26.0*</u>	<u>15.7*</u>	<u>21.8*</u>	<u>15.3*</u>	<u>23.4*</u>	<u>16.4*</u>	<u>24.1*</u>
	RbFT (Ours)	**45.4***	**56.8***	**40.8***	**51.7***	**41.0***	**51.6***	**40.4***	**51.5***	**40.6***	**51.5***	**24.6***	**33.3***	**21.4***	**29.9***	**25.1***	**34.7***	**24.0***	**33.2***
	Improvement	8.6%	12.0%	13.0%	16.7%	12.6%	15.7%	32.9%	31.4%	18.7%	20.9%	32.3%	28.1%	36.3%	37.2%	64.1%	48.3%	46.3%	37.8%

Figure 4: The EM performance of all methods under 4 types of defective data with $\tau = \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

effectiveness-robustness scatter diagram. Results in Figure 3 indicate that RbFT outperforms all other methods in both dimensions, achieving the best overall performance. Specifically, in terms of effectiveness, RbFT surpasses the second-best method, Vanilla RAG; while in robustness, it significantly outperforms the most competitive method, AstuteRAG. This indicates that RbFT not only maintains high answering accuracy but also effectively defends against various complex retrieval defects, striking a better balance between effectiveness and robustness.

Figure 4 further illustrates the EM performance of each model under different retrieval defect types and various τ values. It can be observed that RbFT consistently achieves the best performance across all defect levels and retrieval defect scenarios (i.e., Noisy, Irrelevant, Counterfactual, and Mix), demonstrating outstanding robustness and broad applicability. Whether in a noise-free standard environment or extreme conditions with highly noisy, irrelevant, or misleading information, RbFT significantly outperforms

Table 2: Ablation study on the impact of two fine-tuning tasks, Defects Detection (referred to as DD in the table) and Utility Extraction (referred to as UE in the table). The EM metrics on the test set and the change ratios of EM between single-task fine-tuning and RbFT are reported. "*" denotes the result is significantly worse than RbFT with $p < 0.05$ level.

LLM		Llama				Qwen			
Method		Vanilla	Vanilla + DD	Vanilla + UE	RbFT	Vanilla	Vanilla + DD	Vanilla + UE	RbFT
Clean ($\tau = 0$)		43.8*	49.7(↑ 2.7%)	42.7* (↓ 11.7%)	48.4	41.8*	45.5(↑ 0.0%)	41.5* (↓ 8.6%)	45.4
Normal ($\tau = 0.4$)	Noisy	38.2*	45.2(↑ 1.6%)	39.7* (↓ 10.8%)	44.5	36.1*	40.7 (↓ 0.0%)	37.9* (↓ 7.1%)	40.8
	Irrelevant	38.3*	43.8(↓ 1.1%)	39.5* (↓ 10.8%)	44.3	36.4*	40.2(↓ 2.0%)	37.7* (↓ 8.0%)	41.0
	Counterfactual	32.8*	41.3* (↓ 6.6%)	39.9* (↓ 9.7%)	44.2	30.4*	37.8*(↓ 6.4%)	37.1* (↓ 8.2%)	40.4
	Mix	36.8*	43.7 (↓ 1.6%)	39.2 (↓ 11.7%)	44.4	34.2*	39.5(↓ 2.7%)	37.1* (↓ 8.6%)	40.6
Hard ($\tau = 1.0$)	Noisy	16.9*	28.4*(↓ 8.7%)	28.9* (↓ 7.1%)	31.1	15.9*	22.9* (↓ 6.9%)	23.1* (↓ 6.1%)	24.6
	Irrelevant	7.1*	24.3*(↓ 13.8%)	25.4* (↓ 9.9%)	28.2	12.4*	18.3*(↓ 14.5%)	18.6* (↓ 13.1%)	21.4
	Counterfactual	9.0*	22.7*(↓ 32.8%)	30.3* (↓ 10.4%)	33.8	8.9*	16.4*(↓ 34.7%)	21.6* (↓ 13.9%)	25.1
	Mix	11.4*	25.7*(↓ 19.4%)	29.5* (↓ 7.5%)	31.9	11.0*	19.5* (↓ 18.8%)	21.2* (↓ 11.7%)	24.0

Vanilla RAG and other enhancement methods. Moreover, in high-difficulty scenarios, RbFT further expands its lead over competing approaches, indicating its superior capability in handling harsh retrieval environments. This consistent and substantial performance improvement indicates that RbFT is not only highly effective in addressing complex retrieval defects but also better suited for real-world applications, where potential retrieval defects are common. As a result, since fluctuations in the quality of retrieved documents are unavoidable in real-world scenarios, our RbFT method, with its capability to provide stable and reliable retrieval-generation responses and maintain strong and consistent robustness, is better suited to meet practical requirements, making it a valuable solution for practical application.

6.2 Ablation Study

To further verify the effectiveness and interrelation of the two tasks in RbFT (i.e., Defects Detection and Utility Extraction), we conduct ablation experiments by fine-tuning LLMs using each task individually (referred to as *Vanilla + DD* and *Vanilla + UE*, respectively) to explore their respective roles. Specifically, we adopt the same training steps, learning rate, and LoRA parameters as those used in RbFT. It is worth noting that, since the instructions and output format of the Defects Detection task differ from those of the original QA tasks, fine-tuning using only Defects Detection data inevitably results in degraded performance on QA tasks. Therefore, we supplement the training data for *Vanilla + DD* with QA training data in the *Clean* setting (i.e., the original training data). It can be viewed as the version of RbFT with all defective documents in the Utility Extraction task replaced with their original retrieved versions.

The results of the ablation study, as shown in Table 2, indicate that both training tasks contribute to improving the robustness of LLMs and RAG systems to some extent, though their improvements are still weaker than RbFT. Specifically, the model fine-tuned solely with the Utility Extraction task exhibits a performance drop of approximately 10% compared to RbFT across all three settings. In contrast, the model fine-tuned with Defects Detection demonstrates different features. Under settings with weaker retrieval defects (i.e., *Clean* and *Normal*), *Vanilla + DD* achieves performance comparable

to RbFT. However, in the more challenging retrieval environment of the *Hard* setting, *Vanilla + DD* falls short of *Vanilla + UE* in terms of robustness, especially on the counterfactual data. Therefore, the Defects Detection and Utility Extraction training tasks are mutually complementary, working in tandem to reinforce each other's effectiveness. Only by combining both can we maximize effectiveness in low-defect scenarios while simultaneously enhancing robustness in high-defect environments.

6.3 Case Study

In Figure 5, we attempt to analyze further how RbFT enhances the defense capability of LLMs by examining the attention distribution over tokens of the input document. Specifically, we select one case for each of the three types of retrieval defects (i.e., noisy, irrelevant, and counterfactual) and apply retrieval augmentation to the Llama model using two corresponding defective documents. For noisy documents, as shown in Figure 5a, the model fine-tuned with RbFT distributes its attention more evenly across a broader range of contextually relevant information. In contrast, the Vanilla model tends to concentrate its attention on distracting and misleading entities, for example, "Fiormonda" in Figure 5a. Similarly, in the case of counterfactual documents (as shown in Figure 5c), the RbFT-enhanced model focuses less on the incorrect answer "Toronto" and more broadly on multiple relevant pieces of contextual information, thereby mitigating the impact of erroneous and misleading content. For irrelevant documents (Figure 5b), the Vanilla model also over-focuses on certain specific tokens, whereas the RbFT model distributes its attention more broadly across the context. In summary, the attention distribution of LLMs fine-tuned with RbFT becomes smoother compared to the Vanilla LLMs when processing defective input documents. This smoother attention distribution helps in two ways. First, it increases the model's resistance against incorrect or irrelevant information by reducing excessive attention to and reliance on such content. Second, even when the input document does not directly contain the ground-truth answer, attending to more relevant information in the overall context may better activate the internal parametric knowledge and memory of the LLM, thereby facilitating more accurate responses.

Question: Which character does this protagonist, who secretly loves and marries a member of the rival house, of William Shakespeare's tragedy that has a fictional character Benvolio slay?

Answer: Tybalt

Vanilla Output: Fiormonda

Doc 1 (Title : " Love 's Sacr ifice ") "" Ford 's most typical play ."" The play is , in one view , "" the most puzz ling of Ford 's works ," " and in another , "" a bot ched mess ."" Phill ippo Car aff a , Duke of P avia , has accidentally caught sight of a beautiful young woman named Bian ca , the daughter of a Milan ese gentleman , while he was hunting . Car aff a falls in love with her , and mar ries her . Yet the Duke 's close friend Fernando also falls in love with the new d uch ess ; she rejects him at first , but eventually acknowledges similar feelings for him . Fi orm onda , the wid owed sister of Car aff a , has long suffered an un requ ited passion for Fernando ; she perce ives the Doc 2 (Title : " Sh akespeare in Love ") would go on to win seven Academy Awards , including Best Picture , Best Actress (G wyn eth P alt row), Best Supporting Actress (J udi Den ch), and Best Original Screen play . In 159 3 London , William Shakespeare is a sometime player in the Lord Chamber lain 's Men and poor playwright for Philip H ens low e , owner of The Rose Theatre . Shakespeare is working on a new comedy , "" R ome o and Eth el , the Pirate 's Daughter "". S uffer ing from writer 's block , he has barely begun the play , and is further distracted by attempts both to sed uce Ros al ine , the mistress of Richard Burb age , owner of the rival Curtain Theatre and to convince Burb age to buy the

RbFT Output: Tybalt

Doc 1 (Title : " Love 's Sacr ifice ") "" Ford 's most typical play ."" The play is , in one view , "" the most puzz ling of Ford 's works ," " and in another , "" a bot ched mess ."" Phill ippo Car aff a , Duke of P avia , has accidentally caught sight of a beautiful young woman named Bian ca , the daughter of a Milan ese gentleman , while he was hunting . Car aff a falls in love with her , and mar ries her . Yet the Duke 's close friend Fernando also falls in love with the new d uch ess ; she rejects him at first , but eventually acknowledges similar feelings for him . Fi orm onda , the wid owed sister of Car aff a , has long suffered an un requ ited passion for Fernando ; she perce ives the Doc 2 (Title : " Sh akespeare in Love ") would go on to win seven Academy Awards , including Best Picture , Best Actress (G wyn eth P alt row), Best Supporting Actress (J udi Den ch), and Best Original Screen play . In 159 3 London , William Shakespeare is a sometime player in the Lord Chamber lain 's Men and poor playwright for Philip H ens low e , owner of The Rose Theatre . Shakespeare is working on a new comedy , "" R ome o and Eth el , the Pirate 's Daughter "". S uffer ing from writer 's block , he has barely begun the play , and is further distracted by attempts both to sed uce Ros al ine , the mistress of Richard Burb age , owner of the rival Curtain Theatre and to convince Burb age to buy the

(a) The attention distribution over two input noisy documents of Vanilla RAG and RbFT.

Question: Which airport Pago Pago International Airport or Chattanooga Metropolitan Airport is closer to the central business district of their local community ?

Answer: Chattanooga Metropolitan Airport

Vanilla Output: Neither

Doc 1 (Title : " Brian Bark ") Brian Bark Brian Stuart Bark (born August 26 , 196 8), also known as Sammy B , is a former relief pitcher in Major League Baseball who played for the Boston Red Sox in the 199 5 season . Bark b atted and threw left –handed . He attended both Randall stown High School and North Carolina State University , where he played college baseball for the Wolf pack . He is Jewish . He was a contestant on an episode of the CBS show "" The Brief case ," " but producers decided not to air the episode due to limited excitement . Bark was first drafted by the Baltimore Orioles in the 28 th round of the Doc 2 (Title : " A gricult ure in Par aguay ") the 198 0 s . In fact , Par aguay 's per capita consumption of fuel wood was the highest in all of Latin America and the Caribbean and nearly three times the level of other South American countries . The def orestation question was complicated by the distribution of forest lands and population . Southeast Par aguay was being def ore sted the most rapidly . From the mid – 197 0 s to the mid – 198 0 s , that region 's forest land decreased from just under 45 percent of all land to 30 percent . The Ch aco maintained a large number of forest lands and shr ubs , but they could not be economically exploited . Government policy was slow to respond to def orestation because of

RbFT Output: Chattanooga Metropolitan Airport

Doc 1 (Title : " Brian Bark ") Brian Bark Brian Stuart Bark (born August 26 , 196 8), also known as Sammy B , is a former relief pitcher in Major League Baseball who played for the Boston Red Sox in the 199 5 season . Bark b atted and threw left –handed . He attended both Randall stown High School and North Carolina State University , where he played college baseball for the Wolf pack . He is Jewish . He was a contestant on an episode of the CBS show "" The Brief case ," " but producers decided not to air the episode due to limited excitement . Bark was first drafted by the Baltimore Orioles in the 28 th round of the Doc 2 (Title : " A gricult ure in Par aguay ") the 198 0 s . In fact , Par aguay 's per capita consumption of fuel wood was the highest in all of Latin America and the Caribbean and nearly three times the level of other South American countries . The def orestation question was complicated by the distribution of forest lands and population . Southeast Par aguay was being def ore sted the most rapidly . From the mid – 197 0 s to the mid – 198 0 s , that region 's forest land decreased from just under 45 percent of all land to 30 percent . The Ch aco maintained a large number of forest lands and shr ubs , but they could not be economically exploited . Government policy was slow to respond to def orestation because of

(b) The attention distribution over two input irrelevant documents of Vanilla RAG and RbFT.

Question: Roden Brothers were taken over in 1953 by a group headquartered in which Canadian city?

Answer: Montreal

Vanilla Output: Toronto

Doc 1 (Title : Rod en Brothers tea service for Princess Elizabeth 's wedding in 194 7 , and they were attributed to helping with the production of the Stanley and Grey Cup . Rod en Brothers Rod en Brothers was a Canadian table ware design and manufacturing company . It was founded June 1 , 189 1 in Toronto by Thomas and Frank Rod en . In the 191 0 s the firm became known as Rod en Bros . Ltd . and were later taken over by Henry Bir ks and Sons in 195 3 . Rod en Bros Ltd .' s silver was supplied by the province of British Columbia and with it they produced a wide range of silver hollow are and flat ware in traditional English styles .) Doc 2 (Title : Rod en Brothers Rod en Brothers was a Canadian table ware design and manufacturing company . It was founded June 1 , 189 1 in Toronto by Thomas and Frank Rod en . In the 191 0 s the firm became known as Rod en Bros . Ltd . and were later taken over by Henry Bir ks and Sons in 195 3 . Rod en Bros Ltd .' s silver was supplied by the province of British Columbia and with it they produced a wide range of silver hollow are and flat ware in traditional English styles . The company offered a variety of flat ware patterns that included Strat ford , Queens , and Louis XV . Gold smith s Stock Company were their exclusive selling agents .)

RbFT Output: Montreal

Doc 1 (Title : Rod en Brothers tea service for Princess Elizabeth 's wedding in 194 7 , and they were attributed to helping with the production of the Stanley and Grey Cup . Rod en Brothers Rod en Brothers was a Canadian table ware design and manufacturing company . It was founded June 1 , 189 1 in Toronto by Thomas and Frank Rod en . In the 191 0 s the firm became known as Rod en Bros . Ltd . and were later taken over by Henry Bir ks and Sons in 195 3 . Rod en Bros Ltd .' s silver was supplied by the province of British Columbia and with it they produced a wide range of silver hollow are and flat ware in traditional English styles .) Doc 2 (Title : Rod en Brothers Rod en Brothers was a Canadian table ware design and manufacturing company . It was founded June 1 , 189 1 in Toronto by Thomas and Frank Rod en . In the 191 0 s the firm became known as Rod en Bros . Ltd . and were later taken over by Henry Bir ks and Sons in 195 3 . Rod en Bros Ltd .' s silver was supplied by the province of British Columbia and with it they produced a wide range of silver hollow are and flat ware in traditional English styles . The company offered a variety of flat ware patterns that included Strat ford , Queens , and Louis XV . Gold smith s Stock Company were their exclusive selling agents .)

(c) The attention distribution over two input counterfactual documents of Vanilla RAG and RbFT.

Figure 5: Case studies on the attention distribution over input documents of Vanilla RAG and RbFT under different retrieval defects. The greener a document token, the higher the attention it receives during the answer generation process.

6.4 Efficiency Analysis

In Table 3, we assess the time efficiency of different methods during inference, reporting the average time required by each RAG system to process a single user query. It can be observed that RbFT, by only fine-tuning the LLMs, maintains an inference speed comparable to Vanilla RAG. In contrast, other robustness-oriented methods, except for InstructRAG, adopt more complex inference mechanisms that require multiple generation steps or rounds, leading to significantly higher time costs than Vanilla RAG and RbFT. This demonstrates that RbFT not only excels in performance but also offers a notable advantage in efficiency over other baseline models. On the other hand, RbFT is vertical to these methods and can be integrated with them to further enhance system robustness.

Table 3: The inference efficiency of each method.

Method	Inference Efficiency (s / query)	
	Llama	Qwen
Vanilla RAG	0.193	0.198
RobustRAG	1.207	1.300
CRAG	0.401	0.401
InstructRAG	0.198	0.224
AstuteRAG	3.417	3.369
RbFT	0.196	0.196

7 Conclusion

In this work, we introduce Robust Fine-Tuning (RbFT), a novel fine-tuning approach to enhance the robustness of RAG systems against retrieval defects. By addressing the critical vulnerabilities in RAG systems, specifically their susceptibility to defective retrieval results, RbFT equips LLMs with improved defensive capabilities. Our dual-task fine-tuning strategy mitigates the impact of defective retrieval inputs and ensures effective knowledge utilization even under adverse retrieval conditions. Extensive experimental evaluations demonstrate that RbFT significantly outperforms existing state-of-the-art methods in terms of robustness and inference efficiency. Notably, RbFT maintains high effectiveness even in clean environments while offering reliable responses in high-defect settings, making it a robust and practical solution for real-world RAG applications. In future works, we plan to extend RbFT beyond QA tasks to a broader range of applications. Additionally, since RbFT is theoretically vertical to other baselines focusing on inference strategies and mechanisms, we intend to explore their integration to develop more efficient and robust RAG systems further.

References

[1] Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal* 7 (2024), 100065.

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR, 2206–2240.

[3] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. 2024. Phantom: General Trigger Attacks on Retrieval Augmented Language Generation. *arXiv preprint arXiv:2405.20485* (2024).

[4] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.

[5] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928* (2022).

[6] Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C Park. 2024. Typos that Broke the RAG's Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations. *arXiv preprint arXiv:2404.13948* (2024).

[7] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Unifying Bias and Unfairness in Information Retrieval: A Survey of Challenges and Opportunities with Large Language Models. *arXiv preprint arXiv:2404.11457* (2024).

[8] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755* (2022).

[9] Qian Dong, Qingyao Ai, Hongning Wang, Yiding Liu, Haitao Li, Weihang Su, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2025. Decoupling Knowledge and

Context: An Efficient and Effective Retrieval Augmented Generation Framework via Cross Attention. In *Proceedings of the ACM on Web Conference 2025*.

[10] Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10581–10589.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[12] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[13] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300* (2022).

[14] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[16] Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. GRAG: Graph Retrieval-Augmented Generation. *arXiv preprint arXiv:2405.16506* (2024).

[17] Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 11–20.

[18] Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282* (2020).

[19] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.

[20] Zhengbao Jiang, Luyu Gao, Jun Araki, Haibo Ding, Zhiruo Wang, Jamie Callan, and Graham Neubig. 2022. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. *arXiv preprint arXiv:2212.02027* (2022).

[21] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[22] 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* (2020).

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

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

[25] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110* (2022).

[26] Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024. Citation-Enhanced Generation for LLM-based Chatbot. *arXiv preprint arXiv:2402.16063* (2024).

[27] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283* (2023).

[28] Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. CaseEncoder: A Knowledge-enhanced Pre-trained Model for Legal Case Encoding. *arXiv preprint arXiv:2305.05393* (2023).

[29] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. GPTEval: A survey on assessments of ChatGPT and GPT-4. *arXiv preprint arXiv:2308.12488* (2023).

[30] Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Attacking open-domain question answering by injecting misinformation. *arXiv preprint arXiv:2110.07803* (2021).

[31] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661* (2023).

[32] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921* (2024).

[33] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191* (2020).

[34] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[35] Alireza Salemi and Hamed Zamani. 2024. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 741–751.

[36] Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore. *arXiv preprint arXiv:2407.12854* (2024).

[37] Weihang Su, Qingyao Ai, Xiangsheng Li, Jia Chen, Yiqun Liu, Xiaolong Wu, and Shengluan Hou. 2023. Wikiformer: Pre-training with Structured Information of Wikipedia for Ad-hoc Retrieval. *arXiv preprint arXiv:2312.10661* (2023).

[38] Weihang Su, Qingyao Ai, Yueyue Wu, Yixiao Ma, Haitao Li, and Yiqun Liu. 2023. Caseformer: Pre-training for Legal Case Retrieval. *arXiv preprint arXiv:2311.00333* (2023).

[39] Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Quezi Bing, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. STARD: A Chinese Statute Retrieval Dataset Derived from Real-life Queries by Non-professionals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10658–10671. https://doi.org/10.18653/v1/2024.findings-emnlp.625

[40] Weihang Su, Xiangsheng Li, Yiqun Liu, Min Zhang, and Shaoping Ma. 2023. Thuir2 at ntcir-16 session search (ss) task. *arXiv preprint arXiv:2307.00250* (2023).

[41] Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region.* 23–31.

[42] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081* (2024).

[43] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025. Parametric Retrieval Augmented Generation. *arXiv preprint arXiv:2501.15915* (2025).

[44] Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448* (2024).

[45] Chao-Hong Tan, Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. TegTok: Augmenting text generation via task-specific and open-world knowledge. *arXiv preprint arXiv:2203.08517* (2022).

[46] Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2024. Knowledge Editing through Chain-of-Thought. *arXiv preprint arXiv:2412.17727* (2024).

[47] Changyue Wang, Weihang Su, Hu Yiran, Qingyao Ai, Yueyue Wu, Cheng Luo, Yiqun Liu, Min Zhang, and Shaoping Ma. 2024. LeKUBE: A Legal Knowledge Update BEnchmark. *arXiv preprint arXiv:2407.14192* (2024).

[48] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176* (2024).

[49] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).

[50] Shuo Wang, Shotaro Kinoshita, and Hiromi M Yokoyama. 2022. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* (2022), 10–1227.

[51] Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. InstructRAG: Instructing Retrieval Augmented Generation via Self-Synthesized Rationales. *arXiv preprint arXiv:2406.13629* (2024).

[52] Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2022. Defending Against Misinformation Attacks in Open-Domain Question Answering. *arXiv preprint arXiv:2212.10002* (2022).

[53] Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably Robust RAG against Retrieval Corruption. *arXiv preprint arXiv:2405.15556* (2024).

[54] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884* (2024).

[55] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[56] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).

[57] ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis lectures on human language technologies* 1, 1 (2008), 1–141.

[58] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1503–1512.

[59] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

[60] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102* (2024).

[61] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867* (2024).