Toward Robust RALMs: Revealing the Impact of Imperfect Retrieval on Retrieval-Augmented Language Models

Seong-Il Park
Graduate School of Data Science,
Seoul National University
Seoul, Republic of Korea
athjk3@snu.ac.kr

Jay-Yoon Lee
Graduate School of Data Science,
Seoul National University
Seoul, Republic of Korea
lee.jayyoon@snu.ac.kr

Abstract

Retrieval Augmented Language Models (RALMs) have gained significant attention for their ability to generate accurate answer and improve efficiency. However, RALMs are inherently vulnerable to imperfect information due to their reliance on the imperfect retriever or knowledge source. We identify three common scenarios-unanswerable, adversarial, conflicting-where retrieved document sets can confuse RALM with plausible real-world examples. We present the first comprehensive investigation to assess how well RALMs detect and handle such problematic scenarios. Among these scenarios, to systematically examine adversarial robustness we propose a new adversarial attack method, **Gen**erative model-based **ADV**ersarial attack (**GenADV**) and a novel metric **R**obustness under **A**dditional **D**ocument (**RAD**). Our findings reveal that RALMs often fail to identify the unanswerability or contradiction of a document set, which frequently leads to hallucinations. Moreover, we show the addition of an adversary significantly degrades RALM's performance, with the model becoming even more vulnerable when the two scenarios overlap (adversarial+unanswerable). Our research identifies critical areas for assessing and enhancing the robustness of RALMs, laying the foundation for the development of more robust models.[1]

1 Introduction

Large Language Models (LLMs) are becoming the foundation for various NLP tasks (Brown et al., 2020; Anil et al., 2023; Achiam et al., 2023; Qin et al., 2023). Notably, in open-domain question answering (QA) tasks (Chen et al., 2017) that require substantial knowledge, Retrieval Augmented Language Models (RALMs) have proven to be highly effective (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Izacard et al., 2022; Lin et al., 2023). RALMs generate answers based on external knowledge and shows competitive performance with simple in-context setting without additional training. (Ram et al., 2023)

However, RALMs are known to be sensitive to the quality of external information due to their reliance on it. Common issues such as imperfect retriever or contaminated knowledge sources can affect the robustness of RALMs. (Petroni et al., 2020; Du et al., 2022; Li et al., 2023; Du et al., 2022) For example, Figure 1 shows different types of real-world scenarios, illustrating both incorrect responses by RALM and their ideal responses. If a query "What is the tallest building in the world?" retrieves documents that do not contain the answer, the RALM should classify it as "unanswerable" rather than parroting incorrect answer in the document (*unanswerable scenario*). Furthermore, RALM should ignore documents that do not contain the answer even if it appears to be related to the question (e.g., "The tallest mountain in the world is Mount Everest") and instead extract the answer from documents that do (*adversarial scenario*). In cases where documents provide conflicting answers (e.g., "The tallest building in the world is Burj Khalifa... The Taipei 101 is known for the tallest building in the world"), the model should respond with "conflict" as RALM cannot surely identify which is the correct answer (*conflicting scenario*).

Previous studies have primarily focused on one of these three scenarios, (Chen et al., 2022; Weller et al., 2022; Ren et al., 2023) or on inconsistencies within individual documents (Longpre et al., 2021), addressed methods for mitigating (Asai et al., 2023; Yu et al., 2023; Xu et al., 2023) or examined the relationship between parametric knowledge and documents rather than interactions

[1]The code and data can be found at https://github.com/Atipico1/robust-rag

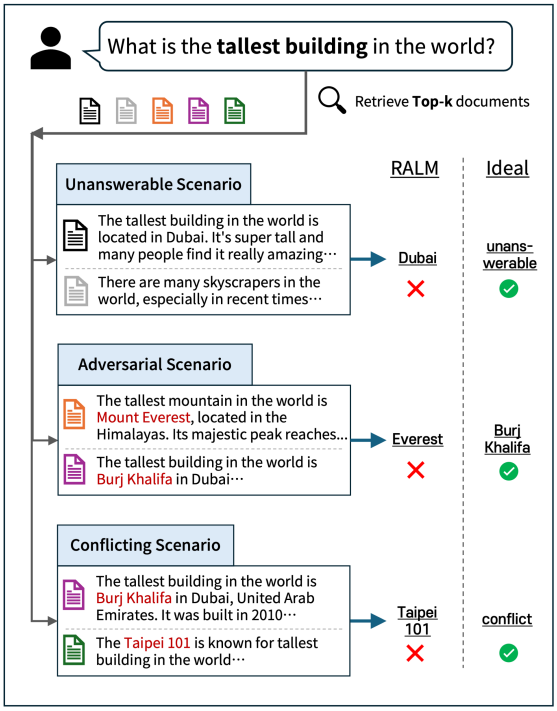

Figure 1: Examples of scenarios with imperfect information. A Robust RALM system can be resilient to imperfections inherent in search engines or knowledge sources.

across multiple retrieved documents (Xie et al., 2023). In contrast, our work systematically analyzes the robustness of RALMs for open-domain QA in scenarios of imperfect information which is a critical factor of RALM's robustness. We define each type of scenario and develop a perturbation method to generate imperfect documents, particularly for simulating adversarial and conflict scenarios in open-domain QA. We also introduce appropriate robustness metrics for each experiment.

For unanswerable scenario, we categorize examples into answerable and unanswerable based on whether the answer strings appear in the retrieved documents, and then measure the accuracy for each subset. Our findings show the challenges RALMs encounter in accurately identifying unanswerability. This often leads to hallucination or parroting incorrect answer in the documents instead of saying unanswerable. For adversarial scenario (Table 1), we introduce a new adversarial attack framework, **Gen**erative model-based **ADV**ersarial attack (**GenADV**) and propose a new metric, **R**obustness under **A**dditional **D**ocument (**RAD**) to measure adversarial robustness of RALMs. Our results indicate that RALMs are vulnerable to adversarial information, particularly GenADV. Also, they show higher vulnerability in adversarial-unanswerable situations where an unanswerable example contain adversarial information in the retrieved documents. Concerning conflicting documents (Table 2), we investigate how well RALMs detect the conflicts and base their responses on conflicting information. This highlight the difficulties RALMs face in detecting conflicts and how easily they can be misled by such misinformation.

In summary, our main contributions are as follows:

- We identify common scenarios involving imperfect information that frequently occur in real-world retrievers, and develop perturbation techniques to simulate these scenarios.
- We design experiments to assess robustness in scenarios of imperfect information and proposed corresponding metrics for a systematic analysis of the results.
- We propose a new adversarial attack method, GenADV, and a metric, RAD, specifically designed to measure adversarial robustness in open-domain QA systems.
- We conduct experiments to evaluate how effectively RALMs detect imperfect information and how often they hallucinate in such situations.

2 Related Works

In-context RALMs Traditionally, Retrieval-Augmented Language Models (RALMs) involved training a generator to generate answers based on the retrieved documents (Lewis et al., 2020; Izacard and Grave, 2021; Izacard et al., 2022). However, recent discoveries show that LLMs can be used as generators for RALM in an in-context setting, without additional training, by simply concatenating retrieved documents to the query (Levine et al., 2022b,a; Kamalloo et al., 2023; Shi et al., 2023b; Ram et al., 2023). Since this method is highly efficient and promising in open-domain QA, we will use the in-context RALMs.

Robustness of LLMs on imperfect information Recent work has demonstrated that LLMs are sensitive to imperfect information, revealing a tendency to adhere to parametric knowledge acquired during pre-training when it conflicts with the given context (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023). In contrast, our research aims to determine whether LLMs can accurately iden-

tify conflicts among multiple documents in a retrieval scenario and ascertain the basis for their responses. Additionally, other previous works have shown that LLMs can produce incorrect answers, a phenomenon often referred to as 'hallucination', especially when the available information is insufficient (Asai and Choi, 2021; Ren et al., 2022; Sulem et al., 2022; Hu et al., 2023; Ren et al., 2023). Building on these findings, our study shifts focus to assess how well LLMs can identify unanswerability in complex scenarios, providing a thorough analysis of situations where detection fails. Moreover, various studies have shown that LLMs can be easily distracted by irrelevant information (Jia and Liang, 2017; Petroni et al., 2020; Creswell et al., 2022; Cao et al., 2022; Shi et al., 2023a; Yoran et al., 2023). To expand on these findings, we demonstrate how to generate distracting information in open-domain QA scenarios to assess the robustness of RALMs.

3 Problem Setup

3.1 Definition of RALM

Our RALM follows in-context RALM framework (Ram et al., 2023), with a particular focus on open-domain QA.

In in-context RALM, for given a input query q and generated response y, we retrieve documents from external knowledge source and use the k highest ranked documents $d = [d_1, d_2, \ldots, d_k]$. We then concatenate q with d for generation. The generation process is represented as:

$$p(y|q) = p(y|d, q)p(d|q) \quad (1)$$

LLMs can directly generate answers for open-domain QA directly through prompting (Levine et al., 2022a,b). Therefore, we utilize a frozen LLM as the generator in our RALM.

3.2 Types of imperfect documents

Our experiments address three scenarios of imperfect information in open-domain QA. Each represents a scenario frequently encountered in retrieval for open-domain QA.

Unanswerable Scenario The first scenario involves unanswerability, in which the set of retrieved documents lacks sufficient information to answer the provided query. In this scenario, there is a high likelihood of hallucination which means parroting the incorrect answer in the documents when RALMs generate responses, thus abstaining is crucial. In our experiments, an unanswerability is identified when all the top-k retrieved documents do not contain the answer string. For detailed experimental settings, refer to section 5.2.

Adversarial Scenario The second scenario, adversarial information refers to situations where the correct answer is not included in the retrieved document, yet the model is misled by distracting information in that document. Table 1 displays real examples of adversarial information present in the open-domain QA dataset, indicating that RALMs can easily be distracted by such adversarial information. Unlike prior studies that primarily focus on adversarial attacks in Machine Reading Comprehension (MRC) systems (Jia and Liang, 2017; Cao et al., 2022) or classification task (Pruthi et al., 2019; Li et al., 2021; Lei et al., 2022), our study addresses adversarial attacks in the context of open-domain QA which utilizes multiple documents for generation.

	TQA
Q	What is the largest city in Ohio?
A	Cleveland (Cincinnati)
Docs	[Doc1] **Cincinnati** is the **third-largest** city in Ohio and 65th in the United States. Its metropolitan area is ... [Doc2] This makes **Dayton** the **fourth-largest** metropolitan area in Ohio and 63rd in the United States...
	NQ
Q	Who got the first nobel prize in physics?
A	Wilhelm Conrad Röntgen (Yuval Katzenelson)
Docs	[Doc1] In 2012, the **first prize winner** was another Israeli teenager, **Yuval Katzenelson** of Kiryat Gat, who presented... [Doc2] ... three names on the list: **Werner Heisenberg**, who received the **Nobel Prize in Physics** in 1932...

Table 1: Actual examples included in the dataset (TriviaQA, Natural Questions) and retrieved documents (Docs) containing adversarial information. Red text indicates the actual output of RALM.

Conflict Scenario The last scenario deals with conflicting information, where there is a contradiction among the information in the retrieved documents. Table 2 shows real example of conflicting information in the dataset. This is a common situation that can occur especially with search engine results, due to the multiple sources involved, or with information that changes over time. Additionally, intentional information poisoning can contaminate knowledge sources (Du et al., 2022; Pan et al., 2023), making it crucial to detect and resolve conflicting information. In our experiments, we follow similar strategies to those described in Xie et al. (2023) for creating conflict-

TQA	
Q	What is a "Scotch Bonnet"?
A	Chili Pepper (Sea snail)
Docs	[Doc1] Scotch bonnet (Semicassis granulata) is a medium-sized to large species of **sea snail**, a marine gastropod... [Doc2] Scotch bonnet, also known as bonney peppers, or Caribbean red peppers, is a variety of **chili pepper**...
NQ	
Q	How many countries are a part of opec?
A	14 (15)
Docs	[Doc1] As of June 2018, OPEC has **15 member countries:** six in the Middle East (Western Asia), seven in Africa, ... [Doc2] As of May 2017, OPEC consists of **14 countries** which earn the majority of their income...

Table 2: Actual examples included in the dataset (TriviaQA, Natural Questions) and retrieved documents (Docs) containing conflicting information. Red text indicates the actual output of RALM.

ing information, in order to test RALM's ability to detect the conflict. For detailed experimental settings, refer to §5.4.1.

4 Experimental Setup

4.1 Task and Datasets

We conducted our experiments focusing on the open-domain QA. We utilized four benchmark datasets: Natural Questions (NQ) (Lee et al., 2019), TriviaQA (TQA) (Joshi et al., 2017), Web Questions (WebQ) (Berant et al., 2013), and PopQA (Mallen et al., 2023). We retrieved the top five documents for each question from Wikipedia[2] based on their cosine similarity to the questions and generated answers using these documents. All our experiments, except for PopQA, were performed on test sets. For the details, please refer to the Table 3.

4.2 Metrics

We use accuracy (Mallen et al., 2023) as a primary metric. Unlike Exact match score, we consider a prediction correct if any substring of the prediction exactly matches any of the answers. This emphasis aligns with our goal to test the robustness of models rather than their extractive QA capabilities from LLMs. Specific metrics defined to test robustness in each experiment will be discussed in the respective experimental sections.

[2]We used preprocessed data following (Karpukhin et al., 2020)

Datasets	Size	Recall@5	Unans
NQ	3610	0.68	0.32
TQA	11313	0.76	0.24
WebQ	2032	0.65	0.35
PopQA	14267	0.67	0.33

Table 3: Dataset statistics and Recall for Top-5 retrieval. Recall means top-k retrieval accuracy as used in (Karpukhin et al., 2020). Unans denotes the proportion of examples for which none of the top-5 documents contain the answer string.

4.3 Models

We use ColBERTv2 (Santhanam et al., 2022) as a dense retriever. We experimented following DPR style passage retrieval (Karpukhin et al., 2020). The LLMs used for generating answers were publicly available instruction following models capable of RALM while being in a frozen state. Models included Llama2 chat (Touvron et al., 2023), Mistral Instruct-v2 (Jiang et al., 2023), Orca2 (Mitra et al., 2023), Qwen 1.5 chat (Bai et al., 2023), and Gemma instruct (Team et al., 2024). Our experiments were conducted using 7B size models, with additional analysis on larger sizes within the same family. Additionally, we used OpenAI's `gpt-4o-mini-2024-07-18` API [3] (abbreviated as **GPT4o-mini**) as a closed-source model for further comparison. To minimize randomness in the generative model, greedy decoding was used during generation, and all random seeds were fixed. For faster inference, we used vLLM (Kwon et al., 2023) in all experiments.

4.4 Prompting

We crafted instructions to assess how well LLMs can detect unanswerability and conflicts in a zero-shot RALM setting. The primary focus is on enhancing a standard retrieval-augmented QA system by integrating capabilities to recognize unanswerability and identify conflicts within the provided documents. The types of prompts are as follows:

Normal prompt This is our basic instruction for retrieval-augmented QA, enabling the LLM to utilize external information retrieved in response to questions.

Unans prompt This instruction incorporates unanswerability detection into the Normal prompt,

[3]For detailed information on the model, refer to https://platform.openai.com/docs/models/gpt-4o-mini.

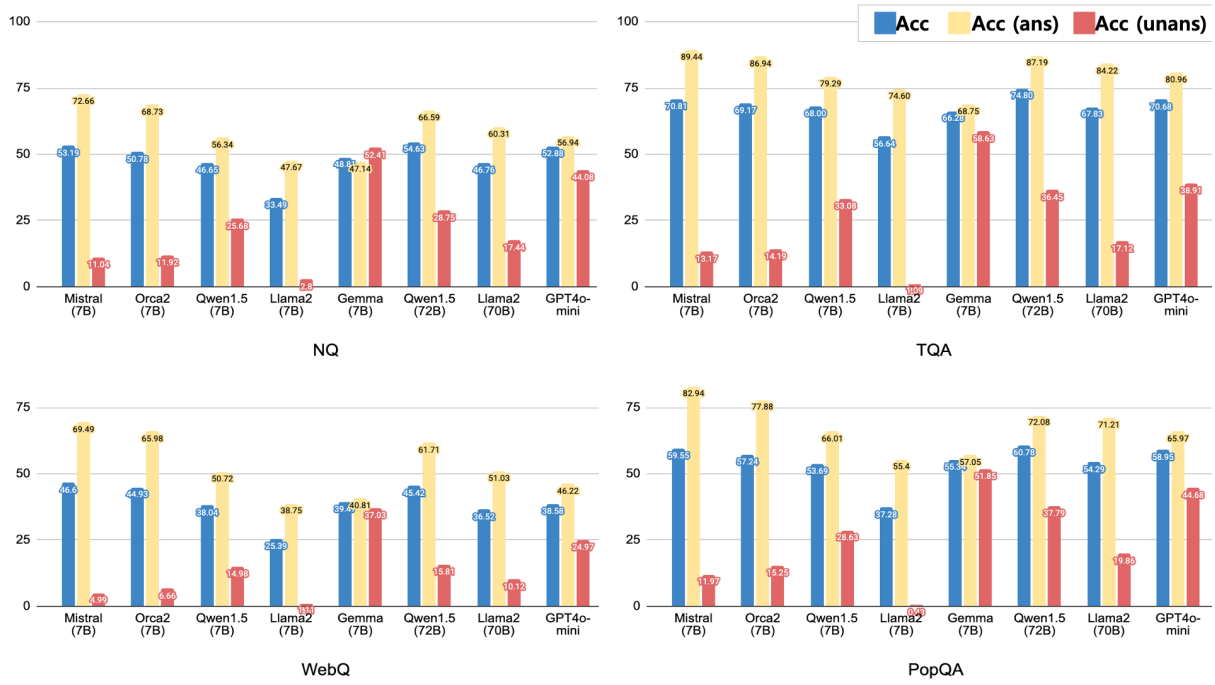

Figure 2: Experimental results on identifying unanswerable examples. The x axis represents the models (size). Acc means accuracy for all examples, Acc (ans) means accuracy for answerable examples and Acc (unans) means accuracy for unanswerable examples. The two models on the far right represent the results of experiments conducted on the largest models within their respective families.

requiring the LLM to not only search for answers but also assess whether the question can be answered based on the provided documents.

Conflict prompt This instruction introduces conflict detection to the Normal prompt, compelling the LLM to meticulously examine the retrieved data for any inconsistencies or contradictory information.

For the details of prompts, refer to Appendix A

5 Experiments

5.1 Baseline QA Performance as a reference

As a reference, we report standard QA performances on RALM and closed-book settings, without our curated prompts for the analysis, on the datasets we utilize in the main experiment.

5.2 Identifying Unanswerability

In this experiment, we aim to test the zero-shot capability of the RALM system to detect when retrieved documents do not contain the answer to a question, known as selective prediction. We will also test how much the models hallucinate in such situations. Since all our datasets are based on extractive QA, we determined the unanswerability of a question by checking if none of the top-5 retrieved documents contained the answer string (we refer to these as *unanswerable examples*). Unlike (Ren et al., 2023), which studied the RALM's ability to determine unanswerability through an additional verification step, we directly test whether the model can identify unanswerability by changing the original answer to *unanswerable*. This is because, in the real world, directly identifying unanswerability allows us to choose other options, such as using a closed book method instead of RALM, or attempting retrieval again. Therefore, if no document included the answer string, we change the original answer to *unanswerable*[4] and if the model responds with *"unanswerable"*, it is considered accurate. Additionally, we instructed the LLM with the *Unans prompt* to indicate unanswerability when it cannot find an answer in the given documents.

5.2.1 Results and Analysis

Answerable vs. Unanswerable We assessed the zero-shot capability of RALMs by dividing test

[4]We use the specific term *unanswerable* instead of a more general expression (e.g., "I don't know") because LLMs showed better identification performance with *unanswerable*.

examples into answerable and unanswerable examples and calculating accuracy for each subset. Figure 2 displays the results, which show significantly lower accuracy for unanswerable examples across most models and datasets. These results indicate that RALMs generally struggle to identify unanswerable scenarios, even in large models and commercial model with strong reasoning capabilities. There were variations in performance among models; for instance, the Llama2 exhibited near-zero accuracy, whereas the Gemma demonstrated higher unanswerable accuracy on the NQ dataset. However, high unanswerable accuracy isn't always reliable. We examined how often models incorrectly responded *"unanswerable"* to answerable examples on the NQ dataset. Gemma, despite high unanswerable accuracy, did this for 28.59% of answerable examples, while Llama2 only 0.7%. This suggests high unanswerable accuracy could simply result from a high propensity of answering *"unanswerable"*, rather than truly identifying unanswerabilities.

	NQ			TQA		
Models	**Acc.**	**Hallu.**	**Cor.**	**Acc.**	**Hallu.**	**Cor.**
Llama2	2.8	95.79	1.4	1.09	94.06	4.85
Mistral	11.04	84.31	4.65	13.17	76.37	10.46
Orca2	11.92	82.03	6.05	14.19	74.3	11.51
Qwen1.5	25.68	73.09	1.23	33.08	63.34	3.58
Gemma	52.41	47.5	0.09	58.63	39.96	1.41
Qwen1.5*	28.75	67.74	3.51	36.45	53.38	10.17
Llama2*	17.44	79.67	2.89	17.12	73.43	9.45
GPT4o-mini	44.08	53.90	2.02	38.91	52.95	8.14

Table 4: Detailed experimental results for unanswerable examples in the NQ and TriviaQA. Acc. indicates percentage of examples where the model accurately identified a question as *unanswerable* (same as Acc (unans) in Figure 2). Cor. means examples where the model provided the true answer to the question. Hallu. represents examples that are neither Acc. nor Cor. * indicates largest models within the family (70B for Llama2 and 72B for Qwen1.5, respectively); all others are 7B models.

Not Responding "Unanswerable" Does Not Imply Correctness Table 4 categorizes the results for the unanswerable examples three groups: those that accurately identified the question as *unanswerable* (abbreviated as **Acc.**), those that provided the original answer to the question (abbreviated as **Cor.**), and those that produced a hallucinated response (abbreviated as **Hallu.**). In all cases, the hallucination ratio significantly outweighed the corrects. Notably, in the NQ dataset, Llama2 hallucinated 95.79% of the time, and 94.06% in the TriviaQA dataset. The large-sized models exhibited similar trends. The Qwen1.5 (with 72B parameters) provide the original answer in only 3.51% of the examples and Llama2 (with 70B parameters) did so in just 2.89% of cases. This demonstrates that failing to correctly respond "unanswerable" does not mean the model has provided the original correct answer; rather, it indicates that the models are mostly **hallucinating**.

Model Size and Robustness Figure 2 also show the results of the larger models. Across all four datasets, larger models exhibited higher accuracy for both answerable and unanswerable examples than their smaller model. Specifically, the Llama2 model, except on the Web Questions, showed a greater performance gain for unanswerable than for answerable examples. This suggests that more complex models possess superior reasoning abilities in more challenging scenarios. However, despite this, the relative accuracy for unanswerable examples remains very low in models larger than 70B, indicating that relying solely on LLM responses to identify unanswerability could be potentially risky.

5.3 Robustness on Adversarial Information

In this experiment, our objective is to test the RALM's robustness in generating correct answers when adversarial documents designed to distract the model are included in the retrieved documents. For the test, we developed a new adversarial attack method for open-domain QA.

5.3.1 Crafting adversarial information

Traditional adversarial information generation in QA systems typically uses word substitution at the entity level, suitable for MRC tasks that rely on a given gold context. (Jia and Liang, 2017; Jin et al., 2020; Cao et al., 2022). However, this approach is less suitable for open-domain QA, which requires multiple passages and does not provide a gold context in advance. Additionally, such adversarially crafted sentences are often grammatically or contextually inconsistent with other documents. To address these issues, we generated adversarial passages using a **Gen**erative model based **ADV**ersarial attack (**GenADV**). GenADV is a hybrid approach. By replacing entities within sentences with similar entities, it retains semantic similarity to the original sentence, while also leveraging LLM to enhance consistency and nat-

uralness in adversaries. However, unlike previous approaches, which depended on gold context (Cao et al., 2022) or human annotation (Jia and Liang, 2017), it relies solely on question-answer (Q-A) pairs. The following describes the process of generating adversarial information using GenADV (Table 5).

1. **Creating an answer sentence and detecting entities:** Initially, we generated an answer sentence using a Q-A pair. After detecting all named entities in the created answer sentence, sentences containing fewer than two entities were excluded. The reason is that when only one entity is detected, there is a high likelihood of conflict if substitution is performed.
2. **Generating Adversarial sentence and filtering:** Using an LLM, we substituted entities in the answer sentence with similar ones to create an adversarial sentence that retained similar meanings but differed in information from the original answer sentence. We excluded adversarial sentences that contained information about the correct answers. Specifically, we removed any sentence that included the answer string or exhibited a cosine similarity of 0.8 or higher with the answer sentence.[5]
3. **Generating Adversarial passage and filtering:** We used the LLM to create supporting passages for the sentences generated in Step 2. Similar to step 2, any adversarial passage containing the answer string was excluded.

Throughout this process, we employed the OpenAI's `gpt-3.5-turbo-0125` model with default generation parameters as our LLM, and used SpaCy NER model[6]for named entity recognition. All the prompts we used can be found in Table 10 in Appendix A.

Afterward, a single adversarial document is randomly inserted among the top-5 retrieved documents. This adversarial addition is semantically very similar to the question but is unrelated to the answer, thus acting as a distraction for the RALM.

[5]We used the Sentence Transformers (Reimers and Gurevych, 2019) library, specifically employing the `all-MiniLM-L6-v2` model for sentence embedding.

[6]We used the `en_core_web_trf` model. The link is as follows. https://spacy.io/models/en

Question	Who got the first nobel prize in physics?
Answer	Wilhelm Conrad Röntgen
Answer Sentence	Wilhelm Conrad Röntgen was awarded the first Nobel Prize in Physics.
Adversarial Sentence	**Jesse Douglas** was the first recipient of the **Fields Medal**
Adversarial Passage	**Jesse Douglas**, an American mathematician, was awarded **the first Fields Medal** in 1936 during the International Congress of Mathematicians in Oslo. He was recognized for his work on the Plateau problem, an important problem...

Table 5: An example of adversarial document generated by GenADV.

5.3.2 RAD score

Our objective is to observe how the performance of the RALM changes when adversarial documents are added. Therefore, a mere decrease in the Exact Match (EM) score may not suffice for precise analysis. To systematically study this, first, we determine the **A**ccurate set of **R**etrieval **A**ugmentation(**ARA**), which consists of instances where the model provides correct answers with retrieved documents. We then define the **ARA-Add** as the instances in the ARA that maintained correctness even when an extra document was added. Consequently, the **R**obustness under **A**dditional **D**ocument (**RAD**) score is calculated as follows:

$$\textbf{RAD} = \frac{\text{\# of ARA-Add}}{\text{\# of ARA}} \times 100 \qquad (2)$$

Using the RAD score, we can precisely analyze the impact that the addition of documents has on the results of the RALM.

5.3.3 Results and Analysis

To assess the effectiveness of GenADV, we compared the RAD scores based on the type of additional document used: a random document (selected randomly from the retrieved documents of other questions) and a top-k document (in the top-5 setting, this refers to the 6th highest-ranked document).

RALMs are Not Robust to Adversaries Figure 3 shows the results of our experiments. In this experiment, we used the *Normal prompt* to obtain the ARA and ARA-Add. Contrary to expectations, adding the top-k documents (referred to as Top-k in the figure) did not result in an RAD close to 100. In fact, in some cases, the RAD was lower than when random documents were added. Particularly in the NQ, RAD was lower in five out of seven models. This suggests that the retrieved documents can contain adversarial information that distracts the model, and that ignoring such documents

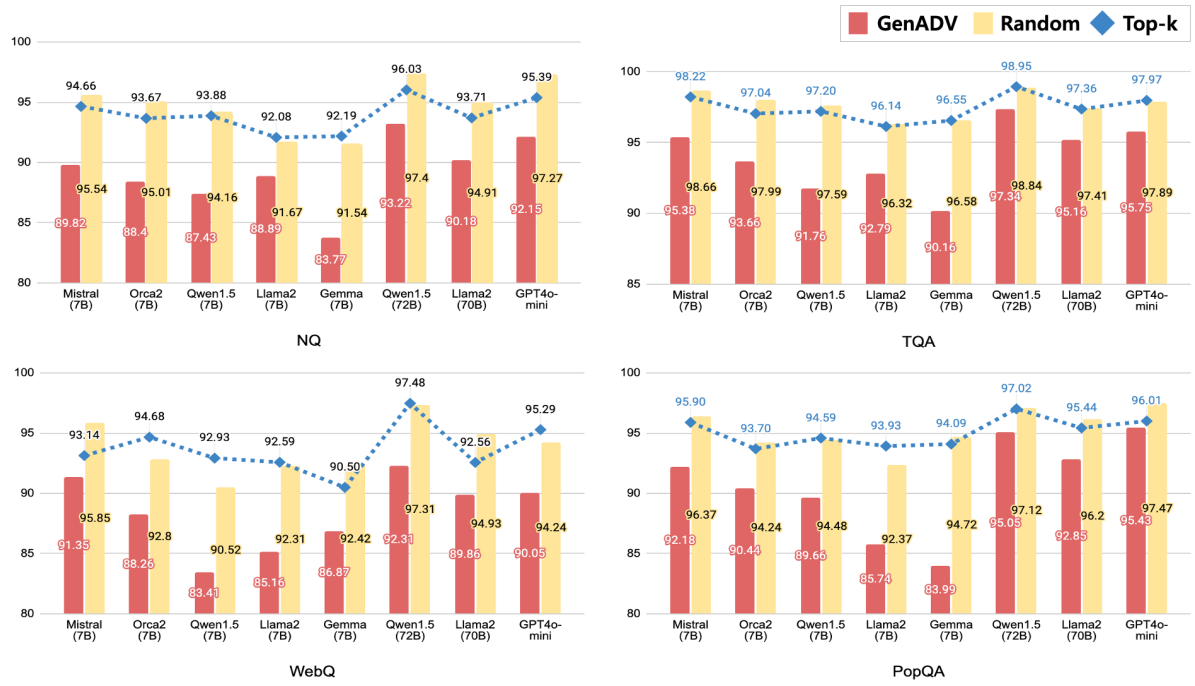

Figure 3: Experimental results on the effects of adding documents. The y-axis represents RAD score. GenADV refers to adding an adversarial document, Random to adding a randomly selected document, and Top-k to adding the sixth highest-ranked document.

can be more challenging for the RALM than disregarding completely unrelated documents. Moreover, the GenADV approach consistently resulted in the lowest RAD across all datasets and models, indicating that our method has the most significant distracting effect on the models.

For example, with Gemma, in the NQ and PopQA datasets, RAD scores of 83.77 and 83.99 were reported, respectively. This indicates that adding just one adversarial passage can induce hallucinations in about 17 out of 100 answers. These results suggest that merely increasing the number of retrieved documents may have limited effects on enhancing the performance of open-domain QA, contrary to previous studies (Ram et al., 2023) and that high-ranked documents also can distract RALMs.

Challenges with Adversarial-Unanswerable Scenario We also created a scenario called *adversarial unanswerable* scenario. This scenario involves a situation where the retrieved documents contain no correct answers (*unanswerable*) while also including adversarial information (*adversarial*). Consequently, RALMs should be able to identify unanswerability without being distracted by adversarial information.

Typically, when retriever fails to fetch accurate information, there is a high likelihood of encountering only adversarial information that is related to the correct answer. Therefore, this scenario is both realistic and crucial. To test this scenario, we used the *Unans Prompt* to obtain

Dataset	Models	GenADV (Ans.)	Random (Ans.)	GenADV (Unans.)	Random (Unans.)
NQ	Mistral	**89.26**	94.84	**49.25**	74.63
	Orca2	**89.15**	94.57	**44.44**	69.44
	Qwen1.5	**86.35**	92.73	**64.86**	89.86
	Gemma	**78.54**	90.83	**83.15**	94.14
	Qwen1.5*	**91.45**	96.76	**69.41**	88.24
	Llama2*	**91.24**	94.88	**68.42**	85.09
	GPT4o-mini	**90.94**	97.34	**84.58**	92.89
TQA	Mistral	**95.38**	98.73	**48.61**	70.14
	Orca2	**93.09**	97.45	**45.24**	77.38
	Qwen1.5	**90.65**	97.17	**64.94**	93.77
	Gemma	**86.49**	95.18	**78.46**	89.37
	Qwen1.5*	**96.44**	98.84	**77.22**	95.44
	Llama2*	**95.46**	98.22	**54.84**	83.87
	GPT4o-mini	**93.52**	97.98	**81.78**	94.13

Table 6: Experimental results on the adversarial unanswerable scenario. Llama2 (7B) was excluded from the results because it correctly identified fewer than 30 unanswerable examples. * indicates largest models within the family (70B for Llama2 and 72B for Qwen1.5, respectively); all others are 7B models.

the ARA[7], then categorized the ARA into answerable and unanswerable examples. Subsequently, we identified the ARA-Add for each category and calculated the RAD score respectively. In this experiment, answerable examples refer to those that are not unanswerable. Through this experiment, we were able to assess the impact of document addition on the performance of RALMs for both answerable and unanswerable examples. Table 6 shows the experimental results for the NQ and TQA datasets under adversarial-unanswerable scenarios, with random method added for comparison. The numbers in the table represent the RAD scores. Across all models, the impact of the adversary was more pronounced on unanswerable examples (Unans.) compared to answerable ones (Ans.). While Gemma showed relative robustness to unanswerable scenarios in NQ, it displayed the lowest RAD scores for answerable examples. These findings indicate that LLMs may struggle more with identifying examples as unanswerable when they fail to retrieve correct answers, particularly in adversarial settings. This result shows that it is more challenging for the RALM to confirm the absence of an answer where none exists than to find the correct answer where one is present. Thus, the close relationship between unanswerable examples and adversaries in real-world contexts implies that merely supplying gold documents with the correct answers is inadequate to address all challenges.

5.4 Identifying Conflicting Information

In this experiment, we evaluate the robustness of LLMs based on two criteria. First, we evaluate the model's ability to detect conflicts in the retrieved documents in a zero-shot setting (conflict detection). Second, we assess whether the model can generate accurate responses when presented with conflicting documents (stubbornness).

Conflict Detection For this evaluation, we used the *Conflict prompt* specifically designed for conflict detection and measured how accurately the model identified the presence of a *conflict*. This experiment was conducted on the answerable examples, with accuracy as the metric.

Stubbornness To evaluate this, we used a *Normal prompt* and measured how well the model generated original answers to the questions despite conflicting information within the documents. This experiment, evaluating stubbornness in retrieval augmentation results, differs from previous studies focused on stubbornness in parametric knowledge (Mallen et al., 2023; Xie et al., 2023). We conducted this experiment on the ARA, using accuracy as our metric.

5.4.1 Crafting conflicting documents

We followed Xie et al. (2023)'s method to create conflicting documents. Specifically, similar to the process of creating adversarial documents, we first generated an answer sentence. Unlike (Xie et al., 2023), we do not perform random substitutions within the same type entity; instead, replaced answer entities in the answer sentence with similar entities of the same type to create a conflicting sentence. This is because we assume a retrieval scenario, and therefore, the conflicting information must also contain information similar to the original answer. We utilized the SpaCy[8] token embedding model to calculate the cosine similarity between entities, and to exclude aliases, we substituted entities with a cosine similarity score of 0.8 or lower.[9] Then, using an LLM, we generated a supporting conflicting passage and a single conflict passage was randomly inserted among the top-5 documents, similar to our approach with adversarial documents.

5.4.2 Results and Analysis

LLMs as Poor Conflict Detectors Table 7 shows the experimental results for identifying conflicting information. Across all datasets and models, the accuracy for conflicting examples was notably low. Although Mistral and Gemma showed relatively high accuracy, even the largest models of Qwen1.5 and Llama2 performed worse than these models. Despite the fact that we carefully crafted the conflicting information to be highly intuitive, the results indicate that LLMs significantly struggle to detect conflicts within documents. This underscores the challenge LLMs face in recognizing conflicts, particularly when they are involved in retrieving and generating content from multiple sources.

RALMs are Vulnerable to Misinformation Table 8 illustrates the results of experiments test-

[7]Also we replaced the original answers with *unanswerable* for unanswerable examples.

[8]We used the `en_core_web_lg` model.

[9]The entity pool for substitutions was created by extracting entities from all texts in the `Wikitext-103-raw-v1` dataset

Models	NQ			TQA			WebQ			PopQA		
	Acc	Acc (C)	Acc (NC)	Acc	Acc (C)	Acc (NC)	Acc	Acc (C)	Acc (NC)	Acc	Acc (C)	Acc (NC)
Mistral	48.68	35.88	66.63	65.58	30.09	88.96	43.17	32.34	62.18	59.77	52.42	80.84
Orca2	25.84	0.35	61.58	50.95	0.18	84.4	21.21	0.48	57.56	19.88	0.14	76.48
Qwen1.5	23.45	9.65	42.8	46.56	5.39	73.69	24.64	17.13	37.82	23.61	11.56	58.12
Llama2	21.39	5.27	43.97	46.33	0.24	76.7	23.04	13.29	40.13	16.62	0.51	62.81
Gemma	40.5	34	49.61	57.16	32.01	73.73	34.71	31.38	40.55	44.85	38.5	63.06
Qwen1.5*	27.99	6.66	57.88	53.17	3.53	85.88	25.48	7.31	57.35	26.67	12.66	66.81
Llama2*	34.67	20.68	54.28	52.44	6.3	82.85	29.14	20.6	44.12	39.59	27.13	75.3
GPT4o-mini	26.65	11.52	47.86	51.56	7.95	80.29	21.36	10.54	40.34	30.65	21.15	57.88

Table 7: Experimental results on identifying conflicts. Acc (C) means the accuracy for conflicting examples, Acc (NC) means the accuracy for non-conflicting examples, and Acc refers to the accuracy for all answerable examples. * indicates largest models within the family (70B for Llama2 and 72B for Qwen1.5, respectively); all others are 7B models.

ing the stubbornness of RALMs. We analyzed the proportion of examples in the ARA that either maintained correctness (A→A), sourced answers from conflicting documents (A→C), or did neither (A→U) when the conflicting document was added. The results highlight the susceptibility of LLMs to misinformation. Gemma retained only 57.88% accuracy in the NQ dataset, while even the highest-performing Llama2 managed only 75.92%. This suggests that in the presence of deliberate misinformation, LLMs are prone to abandoning correct answers in favor of the erroneous information. Particularly in the real world, if contaminated information such as fake news from the internet is retrieved, it implies that RALMs can potentially provide incorrect responses based on such sources.

Models	NQ			TQA		
	A→A	A→C	A→U	A→A	A→C	A→U
Mistral	68.34	25.61	6.05	79.74	16.84	3.42
Orca2	69.36	23.63	7.01	76.49	18.13	5.38
Qwen1.5	66.45	24.36	9.19	68.83	23.34	7.83
Llama2	75.92	13.23	10.88	80.66	11.07	8.27
Gemma	57.88	33.45	8.67	63.11	28.55	8.35
Qwen1.5*	82.64	12.37	4.99	90.5	7.07	2.43
Llama2*	74.02	18.56	7.42	82.23	12.52	5.25
GPT4o-mini	74.84	16.14	9.02	82.07	11.04	6.89

Table 8: Experimental results on the changes in answers when conflicting documents are added. A→A (Answer to Answer) indicates cases where the answer remained the same after the addition of a conflicting document, A→C (Answer to Conflict) indicates cases where the answer was based on information in the conflicting document, and A→U (Answer to Uncertain) refers to all other cases.

6 Conclusion

In this study, we conducted a comprehensive evaluation of the robustness of RALMs under various imperfect retrieval conditions. Our findings revealed significant challenges faced by these models in handling unanswerablity, adversarial and conflicting information.

Through extensive experiments, we demonstrated that RALMs struggle to identify unanswerable scenarios, often hallucinating responses even when the retrieved documents do not contain the answer. Additionally, we introduced a new method, GenADV, for generating adversarial information, which proved highly effective in distracting the models and causing them to abandon correct answers. Furthermore, our results highlighted the vulnerability of RALMs to conflicting information, as they exhibited poor performance in both detecting conflicts within the retrieved documents and generating accurate responses in the presence of such conflicts. Our study provides a foundation for evaluating the robustness of RALMs, crucial for their safe use. Based on this foundation, further exploration into developing robust models is necessary.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2023-00280883, RS-2023-00222663), by the National Super computing Center with super computing resources including technical support (KSC-2023-CRE-0176), and partially supported by New Faculty Startup Fund from Seoul National University

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking qa: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

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

Yu Cao, Dianqi Li, Meng Fang, Tianyi Zhou, Jun Gao, Yibing Zhan, and Dacheng Tao. 2022. TASA: Deceiving question answering models by twin answer sentences attack. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11975–11992, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

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*, volume 36, pages 10581–10589.

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

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. *arXiv preprint arXiv:2307.02394*.

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

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Mandar Joshi, Eunsol Choi, Daniel S 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.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

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*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Yibin Lei, Yu Cao, Dianqi Li, Tianyi Zhou, Meng Fang, and Mykola Pechenizkiy. 2022. Phrase-level textual adversarial attack with label preservation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1095–1112.

Yoav Levine, Itay Dalmedigos, Ori Ram, Yoel Zeldes, Daniel Jannai, Dor Muhlgay, Yoni Osin, Opher Lieber, Barak Lenz, Shai Shalev-Shwartz, et al. 2022a. Standing on the shoulders of giant frozen language models. *arXiv preprint arXiv:2204.10019*.

Yoav Levine, Ori Ram, Daniel Jannai, Barak Lenz, Shai Shalev-Shwartz, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022b. Huge frozen language models as readers for open-domain question answering. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

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*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Alex Troy 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 *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke

Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, no or idk: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1075–1085.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2022. Defending against poisoning attacks in open-domain question answering. *arXiv preprint arXiv:2212.10002*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

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

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

A Prompts

To facilitate the reproducibility of our experiments, we are releasing all the prompts used in our study. Table 9 and 10 show the instructions. The curly brackets denote placeholders where actual values will be inserted.

Name	Instruction
Normal	Documents:{retrieved documents} Use the above documents to answer the subsequent question. Please provide the answer as a single word or term, without forming a complete sentence. Question: {question} Answer:
Unans	Documents:{retrieved documents} Use the above documents to answer the subsequent question. Please provide the answer as a single word or term, without forming a complete sentence. If the answer cannot be found, write 'unanswerable' Question: {question} Answer:
Conflict	Documents:{retrieved documents} Use the above documents to answer the subsequent question. Please provide the answer as a single word or term, without forming a complete sentence. If multiple documents present different answers, please respond with 'conflict' to indicate the presence of conflicting information. Question: {question} Answer:

Table 9: The prompts used in the experiment.

Step	Instruction
Answer Sentence Generation	Please write a single sentence using the following question and answer. The sentence should include the answer and be as realistic as possible.: Question: {question} Answer: {answer} Sentence:
Adversarial Sentence Generation	Rewrite the sentence by replacing the specified words with others, ensuring that the new sentence retains a meaning as close as possible to the original while not being identical. The words to replace are named entities, which should be substituted with entities of the same type. The revised sentence must also remain factually accurate. Original sentence: {answer sentence} Words to replace: {named entities} Revised sentence:
Adversarial Passage Generation	Given a claim, write a concise, factual passage using 50 to 100 words to support it. Please write the passage in the style of Wikipedia: Claim: {adversarial sentence} Passage:

Table 10: The prompts used for crafting adversarial information.

B Baseline performance

We evaluated the performance of models on closed QA as well as retrieval augmented QA. These experiments followed the settings outlined in §4. Closed QA refers to the task where no retrieved documents are provided, allowing us to gauge how much knowledge the model possesses about the dataset, known as *parametric knowledge*. For closed QA, we used the following prompt: *Answer the following question. Please provide the answer as a single word or term, without forming a complete sentence. Q: {question} A:*

To comprehensively assess QA performance, we additionally calculated the exact match score (EM) and F1 score (F1), following the (Izacard et al., 2022). The results of these experiments are presented in Table 11.

Next, for performance in retrieval augmentation, we provided the top-5 retrieved documents and used a *Normal prompt* for inference. We also computed additional metrics. The results are shown in Table 12.

Finally, to further investigate the model's parametric knowledge, we calculated the accuracy rate of correct answers for examples where the top-5 retrieved documents did not contain the answer (same as "unanswerable" in §5.2). If the model correctly answered the question without any relevant information provided, it likely relied on its parametric knowledge. Therefore, a higher rate suggests greater parametric knowledge. We define this metric as the Parametric Answer Rate (PAR). The results for PAR are presented in Table 13.

	Baselines without retrieval											
	NQ			**TQA**			**WebQ**			**PopQA**		
Models	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1
Mistral	33.55	3.38	12.92	60.44	25.29	37.63	43.90	4.63	21.35	24.52	6.09	13.27
Orca2	32.63	4.02	16.93	55.74	21.77	38.17	43.95	6.99	25.73	24.84	5.04	15.52
Qwen1.5	16.29	12.30	18.86	34.88	30.06	37.08	24.90	13.39	28.44	15.36	13.02	17.23
Llama2	14.96	12.02	21.58	38.50	37.34	47.26	19.09	15.75	32.03	16.44	16.14	21.85
Gemma	14.96	3.99	11.49	37.64	14.55	27.22	23.47	5.02	19.19	15.06	3.40	9.31
Qwen1.5*	35.41	23.80	34.95	64.08	56.02	64.93	43.21	19.34	37.69	31.73	22.89	28.95
Llama2*	24.82	20.36	31.16	55.26	54.12	64.39	23.62	20.28	36.23	25.53	25.35	29.69
GPT4o-mini	29.73	29.70	41.27	58.94	59.44	69.84	24.26	22.74	38.69	27.58	27.38	32.90

Table 11: Experimental Results for Closed QA. * indicates largest models within the family (70B for Llama2 and 72B for Qwen1.5, respectively)

	Baselines with retrieval											
	NQ			**TQA**			**WebQ**			**PopQA**		
Models	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1
Mistral	51.61	12.19	26.95	70.67	46.73	59.12	47.05	9.30	25.60	57.12	21.55	33.21
Orca2	49.34	5.57	22.15	69.11	17.05	37.11	44.39	4.08	22.87	53.14	2.34	23.40
Qwen1.5	39.36	33.63	43.57	63.40	59.91	67.43	33.86	19.73	35.60	47.23	42.15	47.47
Llama2	34.90	32.58	42.54	60.05	59.11	66.95	27.61	18.31	33.12	40.30	39.69	44.42
Gemma	39.81	26.32	35.94	60.69	48.59	58.35	31.89	15.50	29.74	45.57	32.92	39.27
Qwen1.5*	48.89	39.39	50.36	71.49	65.31	73.78	42.27	22.93	39.58	51.78	43.89	50.42
Llama2*	40.00	35.68	47.14	63.80	63.17	71.52	32.04	20.18	35.36	48.76	45.22	49.94
GPT4o-mini	41.16	39.97	51.11	65.53	65.05	74.73	30.81	25.10	40.22	48.85	48.05	52.67

Table 12: Experimental results for retrieval augmentation QA. * indicates largest models within the family (70B for Llama2 and 72B for Qwen1.5, respectively)

Parametric Answer Rate				
Models	NQ	TQA	WebQ	PopQA
Mistral	4.82	11.62	5.96	2.83
Orca2	6.13	12.21	6.13	2.79
Qwen1.5	1.49	5.25	1.82	1.62
Llama2	1.41	5.75	1.39	1.32
Gemma	0.26	2.97	1.66	1.38
Qwen1.5*	5.61	15.67	4.58	2.55
Llama2*	3.24	10.46	3.05	2.02
GPT4o-mini	4.47	14.32	2.64	1.94

Table 13: Results of parametric answer rate. We used the Normal prompt for inference. A higher rate indicates that the model correctly answered more questions even without the relevant document being provided, suggesting a greater amount of parametric knowledge about the dataset.

C Human evaluation

C.1 GenADV

Score	Description
Consistency	
1	The passage is very awkward, with poor sentence flow.
2	The passage is somewhat natural but has minor issues in expression or flow.
3	The passage is very natural and flows smoothly.
Similarity	
1	The passage is completely unrelated to the question's topic.
2	Some content or words in the passage are related to the question's topic.
3	Most of the passage content is closely related to the question's topic.
Relevance	
1	It is impossible to infer the correct answer from the passage.
2	The passage content is related to the correct answer but does not directly provide it.
3	The passage content allows for direct or indirect inference of the correct answer.

Table 14: Guidelines provided to evaluators for assessing passages generated by GenADV

To validate the effectiveness of our GenADV, we conducted a human evaluation. We randomly sampled 25 questions, answers, and generated adversarial passages from NQ, TQA, WebQ, and PopQA datasets. These samples were then evaluated on three criteria: *consistency, similarity*, and *relevance*, each scored on a scale from 1 to 3. Consistency assesses the fluency and coherence of the passage. Similarity measures how closely the passage's topic aligns with the question's topic. Relevance evaluates how well the passage allows one to infer the answers. According to these criteria, a good adversarial passage should score high in consistency and similarity, but low in relevance. The guidelines provided to the evaluators can be found in Table 14. To ensure objectivity, we did not disclose the purpose of the evaluation to the evaluators, nor did we inform them that the passages were generated by AI.

Criteria	Score
Consistency	2.83
Similarity	2.24
Relevance	1.14

Table 15: Evaluation results for each criterion of passages generated by GenADV

We selected 10 non-native English speakers with high proficiency in English as evaluators. Each of the 100 samples was reviewed by two evaluators, and the final score was the average of their individual scores. The average scores from the evaluation are shown in Table 15. These results indicate that while GenADV effectively generates passages related to the question's topic, the relevance to the actual answers remains low.

C.2 Conflict

We also manually verify the presence of conflicting information in the conflict passages we created. Similar to C.1, we sampled 25 instances from each dataset and evaluated them. If the passage contained information that contradicted the original answer, we labeled it as a "conflict". If there was no contradiction, it was labeled as "non-conflict". Instances where the conflict was uncertain, such as when the passage was not relevant to the question or did not provide a clear answer, were labeled as "uncertain". In our evaluation, 83 passages contained conflicting information, while only one passage had none, meaning that we successfully generated conflicting passages in over 80% of the instances.