# Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?

**Hexiang Tan**[♠♡], **Fei Sun**[♠†], **Wanli Yang**[♠], **Yuanzhuo Wang**[♠], **Qi Cao**[♠], **Xueqi Cheng**[♠♡]
[♠]CAS Key Laboratory of AI Safety,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[♡]University of Chinese Academy of Sciences, Beijing, China
{tanhexiang21s, sunfei, wangyuanzhuo, caoqi, cxq}@ict.ac.cn

## Abstract

While auxiliary information has become a key to enhancing Large Language Models (LLMs), relatively little is known about how LLMs merge these contexts, specifically contexts generated by LLMs and those retrieved from external sources. To investigate this, we formulate a systematic framework to identify whether LLMs' responses are attributed to either generated or retrieved contexts. To easily trace the origin of the response, we construct datasets with conflicting contexts, i.e., each question is paired with both generated and retrieved contexts, yet only one of them contains the correct answer. Our experiments reveal a significant bias in several LLMs (GPT-4/3.5 and Llama2) to favor generated contexts, even when they provide incorrect information. We further identify two key factors contributing to this bias: i) contexts generated by LLMs typically show greater similarity to the questions, increasing their likelihood of being selected; ii) the segmentation process used in retrieved contexts disrupts their completeness, thereby hindering their full utilization in LLMs. Our analysis enhances the understanding of how LLMs merge diverse contexts, offers valuable insights for advancing current LLM augmentation methods, and highlights the risk of generated misinformation for retrieval-augmented LLMs[1].

## 1 Introduction

Recent advancements in augmenting Large Language Models (LLMs) with auxiliary information have significantly revolutionized their efficacy in knowledge-intensive tasks (Chang et al., 2023; Ram et al., 2023). This auxiliary information can originate from contexts generated by LLMs or retrieved from external sources. For the former, Liu et al. (2022); Sun et al. (2023); Yu et al. (2022) instruct LLMs to initially generate a background context tailored to the given question, which then serves as the basis for producing the final answer. In contrast, retrieval-augmented approaches (Lewis et al., 2020; Ram et al., 2023) include relevant passages retrieved from external corpora, such as Wikipedia, as context, thereby notably enhancing LLMs' capability to address challenges like knowledge updates (Jang et al., 2022) and long-tail knowledge (Kandpal et al., 2023).

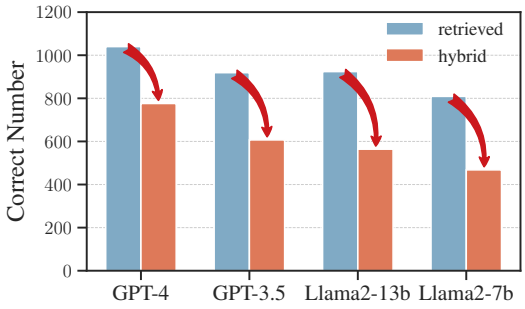



Figure 1: Blue bars show the success number on the NQ test set with only retrieved contexts, while orange bars depict the decline in success for the same questions when generated contexts are additionally incorporated.

Recent hybrid methods have attempted to integrate these two types of contexts to further improve performance in tasks like Question Answering (QA) (Yu et al., 2022; Mallen et al., 2023; Zhang et al., 2023). However, we observe an abnormal phenomenon: in certain cases, models relying solely on retrieval contexts succeeded, whereas hybrid approaches failed, as depicted in Figure 1. This observation implies that LLMs may struggle to effectively integrate diverse types of contexts, overlooking correct information in retrieved contexts. To uncover the reasoning behind this, this study investigates *the underlying mechanisms by which LLMs process the two types of contexts, especially when they contain conflicting information.*

To facilitate the investigation, we proposed a systematic framework to dissect the process by which

[†]Corresponding author.
[1]Code released at https://github.com/Tan-Hexiang/RetrieveOrGenerated.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227

LLMs merge generated and retrieved contexts. We curated tailored **c**ontext-**c**onflicting (CC) datasets in which each question is accompanied by a pair of generated and retrieved contexts. These contexts are deliberately designed to be inconsistent, with only one containing the correct answer to its corresponding question. These datasets provide a solid foundation for determining whether LLMs utilize retrieved or generated context to produce responses in QA tasks.

In this paper, we conducted a series of controlled experiments using our uniquely designed datasets to empirically study this question, focusing on several state-of-the-art closed (GPT-3.5/4) and open (Llama2-7b/13b) LLMs. Surprisingly, our findings reveal a pronounced bias in LLMs to *favor generated contexts* across various LLMs (§4.2), even when the generated contexts offer incorrect information while the retrieval contexts hold the correct answers. Further analysis shows that the bias is prevalent across various retrieval models (§4.3). These findings highlight a critical challenge for existing LLMs in effectively merging contexts from diverse sources, especially in light of the increasing prevalence of LLM-generated content on the internet, which may contain potential misinformation (Pan et al., 2023; Chen and Shu, 2023).

Through extensive empirical analyses, we reveal several factors contributing to this bias: (i) *confirmation bias (Xie et al., 2023) is not a key factor* (§5.1): LLMs maintain a significant preference for generated contexts even when they are inconsistent with LLMs' parametric knowledge. (ii) *text similarity is a significant factor* (§5.2): compared to retrieved contexts, generated contexts typically exhibit a higher degree of similarity to the questions, even when they contain incorrect information. The samples with a larger similarity gap between generated and retrieved contexts exhibit a more pronounced bias. (iii) *semantic completeness matters* (§5.3): LLMs tend to favor contexts with semantic integrity. The segmentation process used in retrieved contexts may disrupt their completeness, thereby hindering their full utilization in LLMs. This finding emphasizes the need for optimizing passage segmentation in current retrieval systems.

In summary, we explore the challenges LLMs face when utilizing conflicting contexts and make the following contributions:

- We uncover a critical bias in existing LLMs, where they heavily rely on generated contexts regardless of correctness, indicating an insufficient use of diverse information sources.
- To facilitate controlled experiments, we develop a specialized framework for constructing tailored datasets and excluding confounding factors, e.g. input order, and context length.
- Our extensive analyses have identified two key factors, i.e., text similarity and semantic completeness, in the context utilization of LLMs.

## 2 Background & Study Formulation

In this section, we briefly review three categories of LLMs augmented with auxiliary information for QA tasks: retrieval-augmented, generation-augmented, and hybrid approaches. Additionally, we introduce the framework of our investigation.

### 2.1 Background

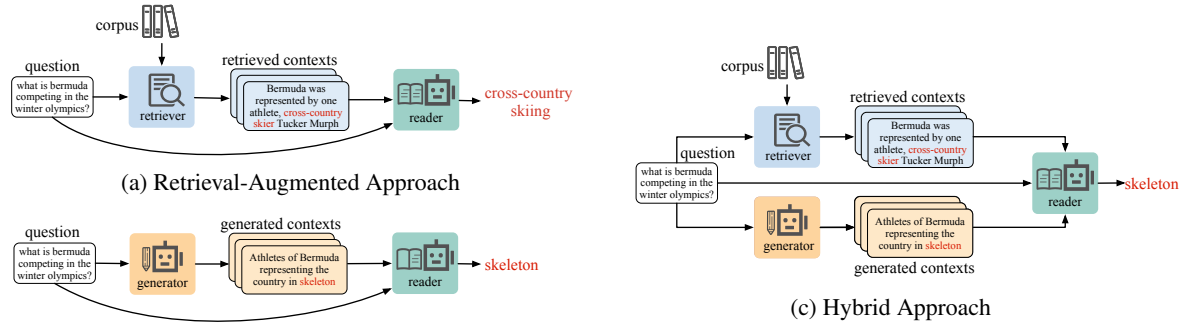
Figure 2 presents high-level abstract frameworks for three typical types of QA systems, each centered around an LLM as the *reader* component, and potentially incorporating additional components like a *retriever*, *generator*, or a blend of both, tailored to the specific methodology.

***Retrieval-Augmented Approach***. As shown in Figure 2a, for a given question $q$ in a set of questions $\mathbb{Q}$, these approaches (Guu et al., 2020; Lewis et al., 2020; Ram et al., 2023; Gao et al., 2023) initially use a retrieval model $\gamma$ to select the top $k$ relevant documents $D_k^\gamma{=}\gamma_k(q,\mathbb{C}){=}\{d_1^\gamma,\ldots,d_k^\gamma\}$ from a corpus $\mathbb{C}{=}\{d_1,\ldots,d_{|\mathbb{C}|}\}$. Then, a reader (often LLM) $\phi$ employs these documents $D_k^\gamma$ to generate an answer $a_\phi^\gamma$, expressed as $a_\phi^\gamma = \phi(q, D_k^\gamma)$.

***Generation-Augmented Approach***. In contrast, as illustrated in Figure 2b, these works (Yu et al., 2022; Sun et al., 2023; Liu et al., 2022) involve an LLM as a generator $\varrho$ to produce $k$ tailored background contexts $D_k^\varrho{=}\varrho_k(q){=}\{d_1^\varrho,\ldots,d_k^\varrho\}$ for a give question $q$, thereby enhancing the utilization of the LLM's internal knowledge. These LLM-generated contexts $D_k^\varrho$ form the input for reader $\phi$ to produce the final answer: $a_\phi^\varrho = \phi(q, D_k^\varrho)$.

***Hybrid Approach***, as depicted in Figure 2c, combines retrieved and generated contexts to enhance performance (Yu et al., 2022; Abdallah and Jatowt, 2023), as $a_\phi = \phi(q, D_k^\gamma, D_k^\varrho)$. These hybrid approaches face a significant challenge: conflicts between diverse sources can impede the effectiveness of information integration (Zhang et al., 2023).

***Knowledge Conflicts within Contexts***. These studies mainly focus on conflicts within a *single* type of input contexts, either only retrieved (Chen et al.,




(a) Retrieval-Augmented Approach

(b) Generation-Augmented Approach

(c) Hybrid Approach

Figure 2: The frameworks of retrieval-augmented approach, generation-augmented approach, and hybrid approach.

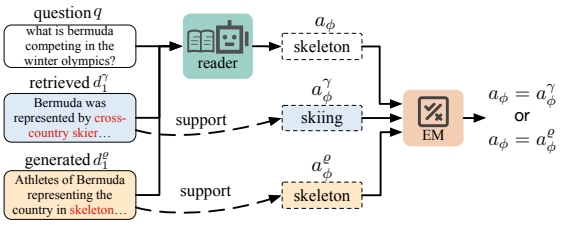



Figure 3: The task to study LLMs' merging mechanisms by tracing the sources of the answers.

2022) or generated (Xie et al., 2023), leaving underexplored how LLMs resolve conflicts between diverse contexts.

### 2.2 Answer Tracing Task

Departing from previous research, our study investigates the mechanisms by which LLMs merge contexts from diverse sources in hybrid approaches. As illustrated in Figure 3, we design a task to ascertain whether an answer $a_\phi$ originates from generated contexts $D_k^\varrho$ or retrieved contexts $D_k^\gamma$. For a more controlled and simpler analysis, we limit the context to a single instance from each source, i.e., $k$=1 and $a_\phi$=$\phi(q, d_1^\gamma, d_1^\varrho)$. Then, by comparing the answer $a_\phi$ with the answers derived from the retrieved context $a_\phi^\gamma$ and the generated context $a_\phi^\varrho$, we can determine its source, thereby analyzing the merging mechanism of LLMs.

We specifically focus on non-tunable LLMs, i.e. in zero-shot settings, to reflect prevalent real-world use cases like ChatGPT. This direction is motivated by the high cost and limited accessibility of fine-tuning, which makes the direct use of non-tunable LLMs popular. Additionally, given the extensive use of LLMs, any bias or issue in their merging mechanisms could lead to serious consequences.

## 3 Experimental Setup

To facilitate our investigation into how LLMs merge generated and retrieved contexts, this section elaborates on the construction of our context-conflicting datasets and the evaluation metric.

### 3.1 Context-Conflicting Datasets

As depicted in Figure 4, in our dataset $\mathcal{D}_{\text{cc}}$, each sample $x$ is a quintet $(q, d_1^\gamma, d_1^\varrho, a_\phi^\gamma, a_\phi^\varrho)$, where $d_1^\gamma$ is the context returned by retriever $\gamma$ for question $q$, $d_1^\varrho$ represents the context generated by LLM $\varrho$, $a_\phi^\gamma$ and $a_\phi^\varrho$ are the candidate answers provided by the reader $\phi$, each based solely on the respective contexts $d_1^\gamma$ and $d_1^\varrho$. To guarantee that our dataset is suitable for controlled experiments aimed at investigating the merging mechanisms of LLMs, it should adhere to specific criteria:

- **Traceability**: $a_\phi^\gamma$ and $a_\phi^\varrho$ should be supported by their corresponding contexts, $d_1^\gamma$ and $d_1^\varrho$.
- **Exclusivity**: Only one of the contexts, $d_1^\gamma$ or $d_1^\varrho$, provides the correct answer, i.e., either $a_\phi^\gamma$ or $a_\phi^\varrho$ matches the gold answer of question $q$.

Such constraints establish a solid basis to identify which context, generated or retrieved, is selected by LLMs to produce answers in hybrid approaches.

We utilize the dev and test sets of two open-domain QA benchmark datasets with golden answers, i.e., NaturalQuestions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (TQA) (Joshi et al., 2017), to assemble our experimental datasets. The overall pipeline for dataset construction is depicted in Figure 4, with detailed steps outlined as follows:

***Context Preparation***. Step 1 in Figure 4 illustrates the process of preparing contexts for each question. For retrieved contexts, it is obtained from the top-1 ranked passage from Wikipedia using Contriever (Izacard et al., 2021), a powerful off-the-shelf retrieval model that is extensively employed in various retrieval-augmented generation systems (Shi et al., 2023; Ram et al., 2023).

For generated context, we follow the GenRead (Yu et al., 2022), instructing the generator, e.g.,

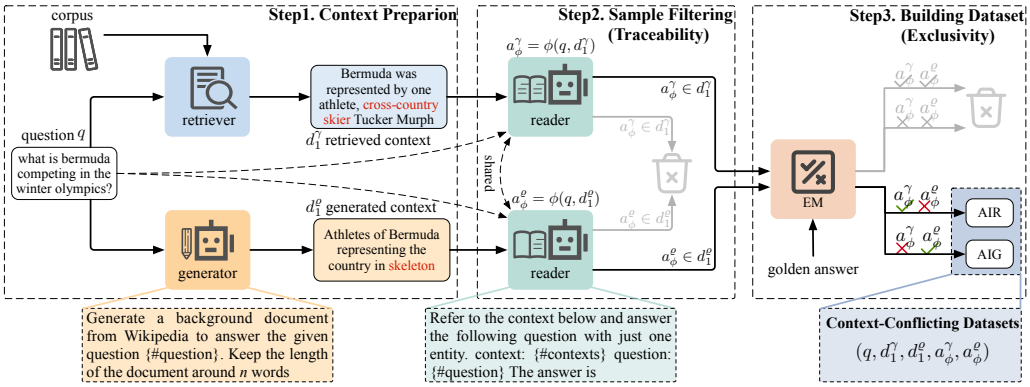



Figure 4: The framework of constructing context-conflicting datasets.

LLM like GPT-4, to generate a background document based on the question. All LLMs in this paper, unless otherwise noted, have a temperature setting of zero to ensure result reproducibility. However, this method often yields contexts much longer (>250 words) than the retrieved contexts (typically truncated to ∼100 words (Karpukhin et al., 2020; Izacard et al., 2021)). The discrepancy in length could potentially affect the merging mechanisms of LLMs (Xie et al., 2023). To exclude this disturbance, we regulate the length of the generated context by incorporating length constraint in the prompt, resulting in an average length discrepancy below 3%. All subsequent experiments, unless otherwise specified, employ this method to eliminate the impact of length variations. More details can be found in Appendix A.1.

***Sample Filtering for Traceability***. With each question paired with a *single* context (either generated or retrieved) established in the initial stage, the reader generates the corresponding candidate answer, as shown in Step 2 of Figure 4. To unravel the mechanisms of LLMs in context merging, it is essential to ensure the *traceability*, i.e., the output answer is derived from the input context, rather than the intrinsic parametric knowledge of the LLMs. To achieve this, we only keep samples in which both the generated and retrieved contexts exactly include their respective generated answers, exemplified by $a_\phi^\gamma \in d_1^\gamma$, where $\in$ denotes $d_1^\gamma$ contains the answer string $a_\phi^\gamma$. This practice is grounded in the findings of Chen et al. (2022); Xie et al. (2023), which demonstrate that in the presence of external context, LLMs mostly rely on external context rather than their intrinsic parametric knowledge. Despite these efforts, the complete elimination of the influence from parametric knowledge remains challenging. Therefore, we separately examine its impact in §5.1, where we demonstrate that it has a negligible effect on our conclusions.

| Generator &Reader | NQ (12367) | | TQA (20150) | |
|---|---|---|---|---|
| | NQ-AIG | NQ-AIR | TQA-AIG | TQA-AIR |
| GPT-4 | 1120 | 763 | 1712 | 681 |
| GPT-3.5 | 1337 | 857 | 2389 | 1042 |
| Llama2-13b | 1441 | 1336 | 2982 | 2091 |
| Llama2-7b | 1423 | 1381 | 3064 | 2604 |
| Avg. Prop. | 10.8% | 8.8% | 12.6% | 8.0% |

Table 1: Dataset size across LLMs. "Avg. Prop." shows average proportions of subsets to original datasets.

***Building Context-Conflicting Dataset***. Having obtained answers for each type of context, we are now positioned to construct our **c**ontext-**c**onflicting (CC) datasets, as depicted in Step 3 of Figure 4. Initially, We employ the exact match metric (Yu et al., 2022) to evaluate the correctness of candidate answers derived from contexts, considering an answer correct if its normalized form matches any of the golden answers.

Subsequently, the context-conflicting datasets are composed of samples for which only one of the two types of contexts, either generated or retrieved, yields the correct answer, thereby ensuring the *exclusivity*. Notably, each dataset comprises two distinct subsets: **AIG**, consisting of samples with correct **a**nswers only **i**n the **g**enerated context; and **AIR** comprising samples with correct **a**nswers only **i**n the **r**etrieved context.

### 3.2 Statistics of Datasets

For each reader-generator pair, we respectively construct context-conflicting datasets from test and dev sets of NQ and TQA: NQ-CC (NQ-AIG + NQ-AIR) and TQA-CC (TQA-AIG + TQA-AIR).

We initially adopt a typical and simple setting in which an LLM serves as both the generator and reader. Table 1 provides statistics for the constructed subsets corresponding to various LLMs, including GPT-4 (gpt-4-0613), GPT-3.5 (gpt-3.5-turbo-0613), Llama2-7b/13b (Llama2-7b/13b-chat

(Touvron et al., 2023)). The statistics show that the context-conflicting subsets form a substantial part of the datasets, underscoring the need to investigate how LLMs integrate these distinct contexts. Notably, GPT-4 has fewer conflicting instances than other LLMs, because of its higher efficacy in answering questions using either solely retrieved or generated contexts.

Section 4.2 also explores a more complex scenario in which the generator and reader are distinct LLMs, with the statistics shown in Appendix A.2.

### 3.3 Evaluation Metric

Besides datasets, we also develop metrics to study how LLMs merge generated and retrieved contexts in hybrid approaches. Specifically, the selection of LLMs towards either generated or retrieved context can be measured by the proportion of answers that exactly match the answer produced solely by the corresponding context, denoted as

$$\rho_{\text{gen}}=\text{avg}(\texttt{em}(a_\phi, a_\phi^o)), \quad \rho_{\text{ret}} = \text{avg}(\texttt{em}(a_\phi, a_\phi^\gamma))$$

where $\texttt{em}(a, b)$ returns 1 if $a$ exactly match $b$, and 0 otherwise. The proportion of instances where $a_\phi$ does not match either $a_\phi^o$ or $a_\phi^\gamma$ is negligible to the conclusion in this work, as demonstrated in Table 17. To facilitate a simple and efficient experiment, we define a synthesized metric as follows:

$$\text{DiffGR} = \frac{\rho_{\text{gen}} - \rho_{\text{ret}}}{\rho_{\text{gen}} + \rho_{\text{ret}}} \quad (1)$$

The metric DiffGR, ranging from $[-1, 1]$, quantifies the extent of LLMs' tendency to rely on generated contexts over retrieved contexts. Using AIR as an example, where all correct answers come from retrieved contexts, an ideal DiffGR is $-1$, i.e., LLMs should always rely on retrieved contexts.

## 4 How LLMs Merge Contexts?

This section conducts experiments on the constructed datasets to investigate the merging mechanism of the LLMs in hybrid approaches. We first consider a typical setting where the generator and reader share a single LLM, to explore how LLMs merge retrieved and *self-generated* contexts (§4.1). Then, we extend our experiments to include more flexible combinations of generator and reader (§4.2), and various retrieval models (§4.3), to investigate their effects.

### 4.1 LLMs Prefer Self-Generated Contexts

Our preliminary experiments, in which a single LLM serves both as generator and reader, are designed to explore how LLMs integrate information from retrieved and *self-generated* contexts. The LLMs under evaluation are tasked with answering questions using both types of contexts on their corresponding CC datasets. In all experiments, we employ a *randomized* input sequence of contexts to mitigate the influence of order, which is further discussed in Appendix B.3.

| Generator &Reader | NQ-CC | | TQA-CC | |
|---|---|---|---|---|
| | NQ-AIG | NQ-AIR | TQA-AIG | TQA-AIR |
| GPT-4 | 91.34 | 17.69 | 94.57 | 19.09 |
| GPT-3.5 | 91.85 | 14.94 | 94.14 | 18.52 |
| Llama2-13b | 90.22 | 18.64 | 92.12 | 20.28 |
| Llama2-7b | 70.77 | 21.51 | 81.17 | 22.16 |

Table 2: The Exact Match (EM) scores (%) of hybrid approaches on corresponding context-conflicting datasets.

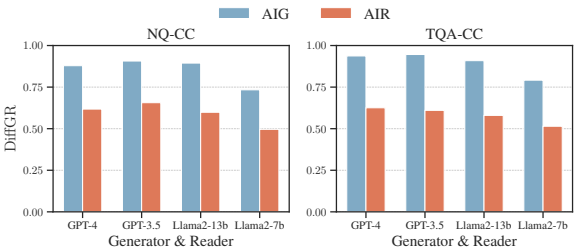



Figure 5: The DiffGR of LLMs on their corresponding context-conflicting datasets.

We begin our analysis by examining LLMs' QA performance on CC datasets to reveal how well can LLMs utilize both types of contexts. Table 2 presents the Exact Match scores (Yu et al., 2022) across various LLMs. Surprisingly, LLMs demonstrate significantly low performance ($\leq 22.16\%$) on AIR subsets, despite the fact that the retrieved context alone consistently yields the correct answer on these subsets. In contrast, LLMs exhibit strong performance on AIG subsets (most near $90\%$). Overall, all LLMs exhibit a significant performance gap between AIR and AIG datasets, with a pronounced decline in performance when the correct answers come from retrieved contexts.

To further reveal LLMs' behavior underlying the QA performance, we trace the source contexts of LLMs' answers using the proposed DiffGR metric. An ideal LLM should always rely on retrieved contexts on AIR subsets (DiffGR $= -1$), and always rely on generated contexts on AIG subsets (DiffGR $= 1$). Contrary to expectations, Figure 5 illustrates that LLMs fail to identify the correct information and consistently tend to rely on generated contexts on both AIG and AIR subsets. This result indicates a pronounced bias in LLMs to **favor generated contexts, even when they provide**

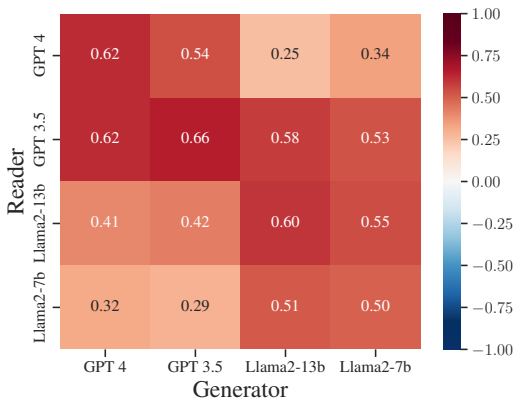



Figure 6: DiffGR with different (reader, generator) pairs on their corresponding NQ-AIR datasets.

**incorrect information**. This bias leads to the insufficient utilization of retrieved contexts mentioned above and highlights a critical challenge for existing LLMs in effectively merging generated and retrieved contexts. As the bias on AIR subsets has a more direct impact on the performance, the following experiments and analysis will focus on the biases on these subsets to conserve space. Results on the AIG subsets can be found in Appendix B.1.

### 4.2 LLMs Broadly Prefer Generated Contexts

The above experiments reveal the bias in LLMs to favor the *self-generated* context. A question emerges: *Do LLMs also prefer contexts generated by other LLMs?* To investigate this question, this section extends the experiments to more flexible combinations of generators and readers. This setting is also of practical significance, as recent works have explored the decoupling of generators and readers to achieve modularization of knowledge (Luo et al., 2023; Feng et al., 2023).

We construct context-conflicting datasets for each (generator, reader) pair respectively. The statistics of these datasets are shown in Appendix A.2. Based on these datasets, we then compute DiffGR metric to examine biases across various (generator, reader) pairs, as shown in Figure 6, and observe two notable insights: (i) **LLMs are also biased towards contexts generated by other LLMs.** This suggests that such bias in LLMs is widespread and not limited to self-generated contexts. (ii) **LLMs usually exhibit a stronger bias to contexts generated by themselves.** The sole exception is Llama2-7b, which shows the strongest bias when paired with Llama2-13b as its generator. This phenomenon likely results from their highly similar model structures and training processes (Touvron et al., 2023).

| Retriever | Generator & Reader | DiffGR |
|---|---|---|
| BM25 | Llama2-7b | 0.5070 |
| Contriever | Llama2-7b | 0.5016 |
| Gold | Llama2-7b | 0.4656 |

Table 3: DiffGR of different retrievers on respectively AIR datasets constructed from NQ dev set.

| Reader | Generator | NQ (12367) | | TQA (20150) | |
|---|---|---|---|---|---|
| | | NQ-AIG | NQ-AIR | TQA-AIG | TQA-AIR |
| GPT 3.5 | GPT 3.5 | 500 | 457 | 524 | 322 |
| GPT 3.5 | Llama2-13b | 271 | 665 | 359 | 574 |
| Llama2-13b | GPT 3.5 | 1318 | 553 | 1889 | 467 |
| Llama2-13b | Llama2-13b | 633 | 841 | 928 | 1020 |

Table 4: The number of data in selected subsets where $a_\phi^{\text{llm}} \neq a_\phi^{\gamma} \neq a_\phi^{o}$. More results are shown in Table 10.

### 4.3 Consistent Preferences Across Retrievers

The above experiments have demonstrated the bias in LLMs against contexts retrieved by **Contriever**, a dense retrieval model. This section further explores whether such bias remains consistent across different retrieval models. To this end, we incorporate **BM25** as a representative of sparse retrieval techniques, along with human-labeled golden passages, to mimic an ideal retrieval model (**Gold**). For each retrieval model, we respectively construct AIR context-conflicting datasets based on the NQ dev sets from Karpukhin et al. (2020), which include the golden passages. As illustrated in Table 3, the results indicate that LLMs consistently display a pronounced bias in favor of generated contexts regardless of the retrieval model used.

## 5 Why LLMs Prefer Generated Contexts

In this section, we investigate the causes of the observed bias from several perspectives: the effect of parametric knowledge in §5.1, context similarity to the question in §5.2, and context completeness in §5.3. This section primarily presents the result for Llama2-13b and GPT-3.5, while the results for the other LLMs are provided in Appendix B.

### 5.1 Effect of Parametric Knowledge

Recently, Xie et al. (2023) demonstrated that LLMs exhibit a bias towards contexts consistent with their parametric knowledge (or memory), a phenomenon termed *confirmation bias*. This section explores whether the observed bias arises from the potential consistency between generated contexts and parametric knowledge. To investigate this, we select subsets where the answers from "retrieved context", "generated context", and "parametric knowledge" were all different from one another.

To achieve the subset, we first establish the

| Reader | Generator | $\rho_{\text{gen}}$ | $\rho_{\text{ret}}$ | $\rho_{\text{llm}}$ | DiffGR | DiffGR (ori) |
|---|---|---|---|---|---|---|
| **NQ-AIR** | | | | | | |
| GPT 3.5 | GPT 3.5 | 67.83 | 12.91 | 0.88 | 0.68 | 0.66 |
| GPT 3.5 | Llama2-13b | 67.22 | 15.34 | 1.20 | 0.63 | 0.58 |
| Llama2-13b | GPT 3.5 | 62.39 | 26.76 | 1.08 | 0.40 | 0.42 |
| Llama2-13b | Llama2-13b | 69.92 | 18.67 | 1.43 | 0.58 | 0.60 |
| **TQA-AIR** | | | | | | |
| GPT 3.5 | GPT 3.5 | 72.05 | 16.15 | 1.55 | 0.63 | 0.61 |
| GPT 3.5 | Llama2-13b | 70.21 | 16.38 | 2.61 | 0.62 | 0.62 |
| Llama2-13b | GPT 3.5 | 61.88 | 26.34 | 1.71 | 0.40 | 0.35 |
| Llama2-13b | Llama2-13b | 72.55 | 17.75 | 1.96 | 0.61 | 0.58 |

Table 5: The columns $\rho_{\text{gen}}$, $\rho_{\text{ret}}$, and $\rho_{\text{llm}}$ represent the proportion (%) of responses matching answers from generated contexts, retrieved contexts, and parametric knowledge, respectively. These values are computed on the filtered subset where $a_\phi^{\text{llm}} \neq a_\phi^\gamma \neq a_\phi^o$, except DiffGR (ori), which is computed on the original dataset. Results for other LLMs are shown in Table 11.

LLM's parametric knowledge about a question using a closed-book QA task, following Xie et al. (2023), which can be expressed as $a_\phi^{\text{llm}} = \phi(q)$. Then, we select the cases that satisfy $a_\phi^{\text{llm}} \neq a_\phi^\gamma \neq a_\phi^o$ from the AIR datasets constructed in §3. Table 4 shows the number of samples after filtering.

The filtered datasets not only exclude the effect of confirmation bias but also help us identify whether LLMs use parametric knowledge or contexts to answer the question. Table 5 illustrates the proportion of LLMs choosing the answer provided by generated contexts, retrieved contexts, or parametric knowledge. We observe two key insights: (i) The proportion of choosing LLMs' parametric knowledge is very small. This result is consistent with the findings of several previous works (Xie et al., 2023; Chen et al., 2022), which found that LLMs mostly rely on the context even when it conflicts with the parametric knowledge. (ii) LLMs still show a significant preference for generated contexts when excluding the influence of parametric knowledge. This suggests that the confirmation bias is not the primary cause of the observed bias.

To ensure rigorous analysis, we use the filtered data in Table 4 for subsequent experiments to eliminate any potential distractions.

## 5.2 Effect of Text Similarity

The **text similarity** between a context and a question can reflect the degree of their relevance. To investigate the potential effect of the similarity, we employ Jaccard similarity and BERTScore (Zhang et al., 2020) to analyze the contexts on the constructed context-conflicting datasets with the reader and generator sharing a single LLM. Figure 7 shows that generated contexts exhibit a significantly higher similarity to the question on AIR

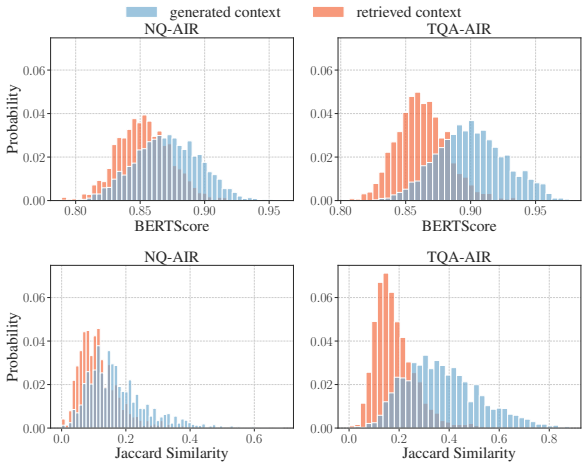



Figure 7: Context-question similarity distribution of generated and retrieved contexts on the union of AIR subsets for different LLMs. More results are shown in Appendix B.5.

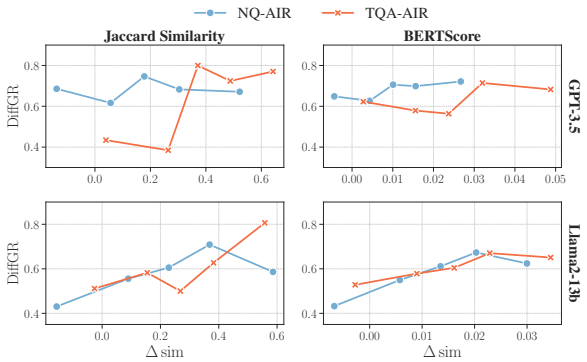



Figure 8: The DiffGR in slices with different average $\Delta$ sim. Results for other LLMs are in Figure 15.

subsets, despite the fact that generated contexts are incorrect on these subsets. This similarity discrepancy between generated and retrieved contexts persists whether assessed by term-based overlap (average 0.37 vs. 0.18 on TQA-AIR) or semantic similarity (0.90 vs. 0.86).

To further clarify the influence of the observed similarity discrepancies, we rank the samples according to the similarity gap $\Delta$ sim between generated and retrieved contexts.

$$\Delta \text{sim} = \frac{\text{sim}(q, d^o) - \text{sim}(q, d^\gamma)}{\text{sim}(q, d^o) + \text{sim}(q, d^\gamma)}$$

Here, $\text{sim}(q, d^o)$ is the similarity between generated context and question, and $\text{sim}(q, d^\gamma)$ is for retrieved context. Then, we divide the dataset into $n$ ($n = 5$ here[2]) slices with an equal number of samples. Ensuring that each slice contains an equal number of samples helps to avoid fluctuations caused by sample size variations.

[2]Similar results and observations are found with other $n$.

| Context | Completeness | | Similarity | | Length |
|---|---|---|---|---|---|
| | Sentence | Semantic | Jaccard | BERTScore | |
| Retrieved | ✗ | ✗ | 0.1162 | 0.8552 | 107.1 |
| Nature | ✔ | ✔ | 0.1748 | 0.8730 | 105.7 |
| S-Trunc. | ✔ | ✗ | 0.1733 | 0.8730 | 105.8 |
| Trunc. | ✗ | ✗ | 0.1769 | 0.8736 | 107.1 |

Table 6: Average length and similarity of contexts with different completeness on NQ-AIR dataset for GPT-3.5 (more details in Table 14).

| Context Pair | NQ-AIR | | TQA-AIR | |
|---|---|---|---|---|
| | GPT-3.5 | Llama2-13b | GPT-3.5 | Llama2-13b |
| Nature vs. Ret. | **0.7519** | **0.6082** | **0.6353** | **0.6207** |
| S-Trunc. vs. Ret. | 0.4864 | 0.2551 | 0.5802 | 0.2779 |
| Trunc. vs. Ret. | 0.4792 | 0.2198 | 0.5663 | 0.2787 |

Table 7: DiffGR with different completeness in generated context. "Nature", "Trunc." and "S-Trunc." represent three types of generated contexts with different completeness. "Ret" means retrieved contexts.

Figure 8 illustrates the relationship between the average $\Delta\,\mathrm{sim}$ within each slice and the corresponding DiffGR[3]. From the results, we observe a general trend that **LLMs exhibit an increased bias to generated contexts on slices with a larger average similarity gap**, which indicates that text similarity is a significant factor in the preference for generated contexts. These findings suggest that generated contexts should be applied with greater caution to mitigate the influence of highly relevant but misleading information.

To facilitate understanding why the similarity affects LLMs' preference, we include some examples in Table 16. From these cases, we observe that contexts with higher similarity often support candidate answers more straightforwardly, for instance, by mirroring the phrasing used in the questions. Conversely, the contexts with low similarity introduce more challenges, often necessitating an understanding of synonyms and even some inferences.

### 5.3 Effect of Context Completeness

In all the above experiments, there is a key difference between generated and retrieved contexts that may affect the context preference: **semantic and sentence completeness**. Concretely, current retrieval systems typically employ fixed-length truncation to divide a complete article into multiple passages, which serve as the fundamental units for retrieval tasks (Karpukhin et al., 2020; Wang et al., 2019; Zhu et al., 2021). This truncation often results in retrieved contexts with incomplete semantic meaning, as well as sentences that are cut off at beginnings or endings. In contrast, generated contexts in the above experiments are naturally produced by LLMs (**Nature**), resulting in enhanced semantic and sentence completeness.

To investigate the potential effects of completeness on the observed bias, we conduct controlled experiments that vary the semantic and sentence completeness of generated contexts[4] using the following methods: (a) **Truncation (Trunc.)** eliminates the length constraints from the generation prompt of §3, allowing LLMs to generate extended contexts. These generated contexts are then truncated to match the length of retrieved contexts, thereby simulating both semantic and sentence incompleteness of retrieved contexts. (b) **Sentence Truncation (S-Trunc.)**: Based on the method (a), we truncate generated contexts only at the end of a sentence to preserve the sentence completeness, while simulating the semantic incompleteness.

To eliminate the interference of similarity factors, we select questions whose three types of generated contexts have nearly equivalent similarity, with BERTScore differences less than 0.05[5]. Table 6 demonstrates that three types of generated contexts also have similar average lengths after filtering. This means that both the influences of similarity and length are mitigated, thereby highlighting the principal disparities in semantic and sentence completeness.

We evaluate LLMs' preference between generated versus retrieved context, varying the completeness of generated context, following the same pipeline in §4.1. Table 7 presents the DiffGR with different semantic and sentence completeness in generated contexts. A comparison between "Trunc." and "S-Trunc." reveals that sentence completeness has a very slight impact on LLMs' preference for generated contexts. In contrast, comparing "Nature" and "S-Trunc.", we find a significant increase in bias towards generated contexts that are semantically more complete. These findings indicate that **LLMs tend to favor contexts with enhanced semantic completeness**, underscoring the necessity to investigate improved passage segmentation methods that maintain semantic completeness for current retrieval-augmented LMs.

[3]We also tried to manipulate the context similarity to investigate its impact, but we found it challenging to instruct LLMs to generate contexts with low similarity to the question while still providing an answer to it.

[4]We also tried to vary the completeness of retrieved contexts but found it challenging to isolate it from confounding factors like length. This aspect is left for future work.

[5]We also tried several other thresholds for similarity differences, e.g. 0.005, all yielding same conclusions.

## 6 Related Work

### 6.1 Generation-Augmented Approaches

Generation-augmented methods prompt LLMs to generate intermediate contexts for the final response, thereby leveraging their extensive parametric knowledge acquired during the pre-training phase on vast text corpora (Roberts et al., 2020; Petroni et al., 2019). These generated contexts may encompass various types of knowledge, such as background knowledge (Sun et al., 2023; Yu et al., 2022), commonsense knowledge (Liu et al., 2022), domain-specific knowledge (Feng et al., 2023; Luo et al., 2023), and chain-of-thought reasoning processes (Wei et al., 2022; Kojima et al., 2022). Despite their effectiveness, the LLM-generated knowledge may contain hallucinations (Chen et al., 2023; Ji et al., 2023) due to LLMs' outdated memory (De Cao et al., 2021) and limited memory for long-tail knowledge (Kandpal et al., 2023). This inaccuracy information could potentially mislead current retrieval model (Dai et al., 2023) and open-domain question answering systems (Pan et al., 2023).

### 6.2 Retrieval-Augmented Approaches

Retrieval-augmented methods (Guu et al., 2020; Lewis et al., 2020; Ram et al., 2023; Gao et al., 2023) enhance LLMs by integrating relevant documents from external corpora, addressing limitations like the need for knowledge updates (Jang et al., 2022) and long-tail knowledge (Kandpal et al., 2023). Early methods (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2022) focused on joint training of LLMs and retrievers. Recent studies (Ram et al., 2023; Shi et al., 2023) append relevant documents directly to the input while keeping LLMs static. Despite their effectiveness, these methods face challenges with irrelevant retrievals and incomplete knowledge coverage (Yu et al., 2023; Mallen et al., 2023). These noisy retrieval results can misguide LLMs (Mallen et al., 2023; Yoran et al., 2023; Ren et al., 2023).

### 6.3 Hybrid Approaches and Knowledge Conflicts

Recent works investigate merging retrieved and generated contexts to leverage both sources of knowledge (Abdallah and Jatowt, 2023; Yu et al., 2022; Zhang et al., 2023) and achieve improved performance over those relying solely on a single information source. However, conflicts between diverse sources can impede the effectiveness of information integration (Zhang et al., 2023) Current research on knowledge conflicts primarily focuses on two aspects: *context-memory conflict* and *inter-context conflict* (Xu et al., 2024). Regarding context-memory conflict, Xie et al. (2023); Chen et al. (2022) find that LLMs are highly receptive to the input contexts rather than their internal memory. Concerning inter-context conflict, Chen et al. (2022) demonstrates that LLMs tend to rely on a few most relevant retrieved contexts. Additionally, Xie et al. (2023) reveals that LLMs favor contexts consistent with their parametric knowledge when confronted with both supporting and opposing contexts. In contrast to these studies which are limited to conflicts within a single type of context, our work further considers conflicts between generated and retrieved contexts and reveals a bias in LLMs.

## 7 Conclusion and Future Work

In this study, we propose a framework to investigate the underlying mechanisms by which LLMs merge retrieved and generated contexts. Our results reveal a pronounced bias towards generated contexts in several LLMs (GPT 3.5/4 and Llama2-7b/13b). We further identify two key factors that may contribute to this bias: higher similarity between generated contexts and questions, and the semantic incompleteness of retrieved contexts. Our insights highlight the critical need for advanced integration methods that can validate and leverage information from both sources, moving beyond the current overreliance on generated contexts.

While this study primarily focuses on integrating generated and retrieved contexts, the observed bias also highlights a critical risk of retrieval-augmented LLMs. With LLM-generated content increasingly prevalent on the internet, retrieval results may include generated contexts (Chen and Shu, 2023). As this scenario becomes more common, the observed bias of LLMs toward generated contexts implies their susceptibility to misinformation and malicious attacks generated by LLMs. This raises concerns about the security of retrieval-augmented systems, a critical problem that is gaining attention in recent works (Pan et al., 2023; Zou et al., 2024). Addressing the challenges posed by the widespread presence of generated content, such as the detection of such content, represents a promising direction for future research.

## Limitations

Our work has the following limitations:

- This study is confined to open-domain question answering, a representative knowledge-intensive task. The behavior of LLMs across a broader spectrum of natural language processing tasks remains to be further explored.
- This work does not propose specific solutions to effectively mitigate the observed bias, as we focus on revealing the phenomena and analyzing the causes.
- To create a controlled environment conducive to analysis, we utilize a single instance for each context type. LLMs face increasingly intricate conflict scenarios when handling multiple contexts from each type. These conflicts emerge not only between retrieved and internally generated contexts but also among the various contexts originating from the same source (Chen et al., 2022; Xie et al., 2023).

## Ethics Statement

**Data** All data used in this study are publicly available and do not pose any privacy concerns.

**AI Writing Assistance** In our study, we only employed ChatGPT to polish our textual expressions rather than to generate new ideas or suggestions.

## Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB3103700, 2022YFB3103704), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0680202), the Innovation Funding of ICT, CAS (E361120), the National Natural Science Foundation of China (No.62172393), and Major Public Welfare Project of Henan Province (No.201300311200).

## References

Abdelrahman Abdallah and Adam Jatowt. 2023. Generator-retriever-generator: A novel approach to open-domain question answering. *arXiv preprint arXiv:2307.11278*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341, Singapore. Association for Computational Linguistics.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *arXiv preprint arXiv:2310.20501*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Cook: Empowering general-purpose language models with modular and collaborative knowledge. *arXiv preprint arXiv:2305.09955*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*.

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

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

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*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

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*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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*.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

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, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain QA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728, Singapore. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

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*.

| Dataset | Retrieved | Generated | | | |
|---|---|---|---|---|---|
| | | GPT 4 | GPT 3.5 | Llama2-13b | Llama2-7b |
| NQ | 107.3 | 108.0 | 106.0 | 110.1 | 104.0 |
| TQA | 106.3 | 107.2 | 104.9 | 105.5 | 102.6 |

Table 8: Average lengths of the generated and retrieved contexts. Length is measured in the number of words after punctuation removal.

# A Detailed Datasets Statistics

## A.1 Length Control for Generated Contexts

In our proposed framework, we regulate the length of generated contexts by incorporating length constraints in the prompt:

> *Generate a background context from Wikipedia to answer the given question {#question}. Keep the length of the document around $n$ words.*

We observed that GPT 4 effectively controls the output length, whereas other models struggle with this aspect. To address this issue in the latter, we employ multiple values of $n$ and select the one that best matches the retrieved context.

As a result, Table 8 illustrates the average length of different contexts. Figure 9 shows the length distribution of retrieved contexts and contexts generated by various LLMs. The length distribution of retrieved contexts is more concentrated as they consist of text limited to precisely 100 words, along with their titles (Karpukhin et al., 2020). The variation in the length of different retrieved contexts is solely due to the differences in title lengths.

## A.2 Dataset Size

Table 9 presents the data size of context-conflicting datasets corresponding to various generator-reader pairs. The statistics indicate that conflicting data comprise a substantial proportion across all combinations of generators and readers.

# B Additional Results

## B.1 More Results on AIG Datasets

Figure 10 shows the DiffGR with different (reader, generator) pairs on their corresponding NQ-AIG datasets. It can be observed that LLMs show a strong tendency to rely on generated contexts across various (reader, generator) pairs.

## B.2 More Results about Effect of Parametric Knowledge

Table 10 illustrates the size of the filtered AIR datasets with the constraint "$a_\phi^{\text{llm}} \neq a_\phi^{\gamma} \neq a_\phi^{o}$".

| Reader | Generator | NQ (12367) | | TQA (20150) | |
|---|---|---|---|---|---|
| | | NQ-AIG | NQ-AIR | TQA-AIG | TQA-AIR |
| GPT 4 | GPT 4 | 1120 | 763 | 1712 | 681 |
| GPT 4 | GPT 3.5 | 1017 | 922 | - | - |
| GPT 4 | Llama2-13b | 730 | 1461 | - | - |
| GPT 4 | Llama2-7b | 600 | 1627 | - | - |
| GPT 3.5 | GPT 4 | 1514 | 769 | 2701 | 794 |
| GPT 3.5 | GPT 3.5 | 1337 | 857 | 2389 | 1042 |
| GPT 3.5 | Llama2-13b | 875 | 1318 | 1781 | 2119 |
| GPT 3.5 | Llama2-7b | 701 | 1502 | 1471 | 2641 |
| Llama2-13b | GPT 4 | 2501 | 767 | 4769 | 741 |
| Llama2-13b | GPT 3.5 | 2211 | 899 | 4210 | 1038 |
| Llama2-13b | Llama2-13b | 1441 | 1336 | 2982 | 2091 |
| Llama2-13b | Llama2-7b | 1221 | 1583 | 2567 | 2773 |
| Llama2-7b | GPT 4 | 2699 | 668 | 5370 | 830 |
| Llama2-7b | GPT 3.5 | 2435 | 785 | 4813 | 1120 |
| Llama2-7b | Llama2-13b | 1569 | 1220 | 3526 | 2051 |
| Llama2-7b | Llama2-7b | 1423 | 1381 | 3064 | 2604 |

Table 9: The data quantities of the constructed subsets for different (Generator, Reader) pairs. NQ and TQA refer to the original datasets (dev+test).

| Reader | Generator | NQ (12367) | | TQA (20150) | |
|---|---|---|---|---|---|
| | | NQ-AIG | NQ-AIR | TQA-AIG | TQA-AIR |
| GPT 4 | GPT 4 | 254 | 342 | 259 | 180 |
| GPT 4 | GPT 3.5 | 220 | 381 | - | - |
| GPT 4 | Llama2-13b | 160 | 527 | - | - |
| GPT 4 | Llama2-7b | 124 | 545 | - | - |
| GPT 3.5 | GPT 4 | 653 | 423 | 796 | 237 |
| GPT 3.5 | GPT 3.5 | 500 | 457 | 524 | 322 |
| GPT 3.5 | Llama2-13b | 271 | 665 | 359 | 574 |
| GPT 3.5 | Llama2-7b | 211 | 708 | 279 | 696 |
| Llama2-13b | GPT 4 | 1563 | 478 | 2321 | 309 |
| Llama2-13b | GPT 3.5 | 1318 | 553 | 1889 | 467 |
| Llama2-13b | Llama2-13b | 633 | 841 | 928 | 1020 |
| Llama2-13b | Llama2-7b | 515 | 1018 | 750 | 1358 |
| Llama2-7b | GPT 4 | 1896 | 479 | 3163 | 418 |
| Llama2-7b | GPT 3.5 | 1682 | 557 | 2746 | 591 |
| Llama2-7b | Llama2-13b | 908 | 884 | 1681 | 1238 |
| Llama2-7b | Llama2-7b | 753 | 1002 | 1272 | 1538 |

Table 10: The number of data in selected subsets where $a_\phi^{\text{llm}} \neq a_\phi^{\gamma} \neq a_\phi^{o}$.

GPT-4 corresponds to fewer samples because it has better parametric knowledge, i.e., $a_\phi^{\text{llm}}$ is more likely to be correct. Table 11 illustrates the proportion of LLMs choosing the answer provided by generated contexts, retrieved contexts, or parametric knowledge.

## B.3 Effect of Context Order

In the above experiments, retrieved and generated contexts are presented in random order. Previous studies (Xie et al., 2023; Liu et al., 2023; Lu et al., 2022) have found that the model may be sensitive to the order of the input contexts. In their experiments, the input context was either all retrieved (Liu et al., 2023) or all generated (Xie et al., 2023). We conducted experiments to investigate whether the context order impacts the preference for the generated context. The generated and retrieved contexts are concatenated with three different orders: generated-first, retrieved-first, and random. To control the cost of API, this section conducts experiments on the context-conflicting datasets from only the test sets of NQ and TQA. We compute the DiffGR with different context orders respectively.

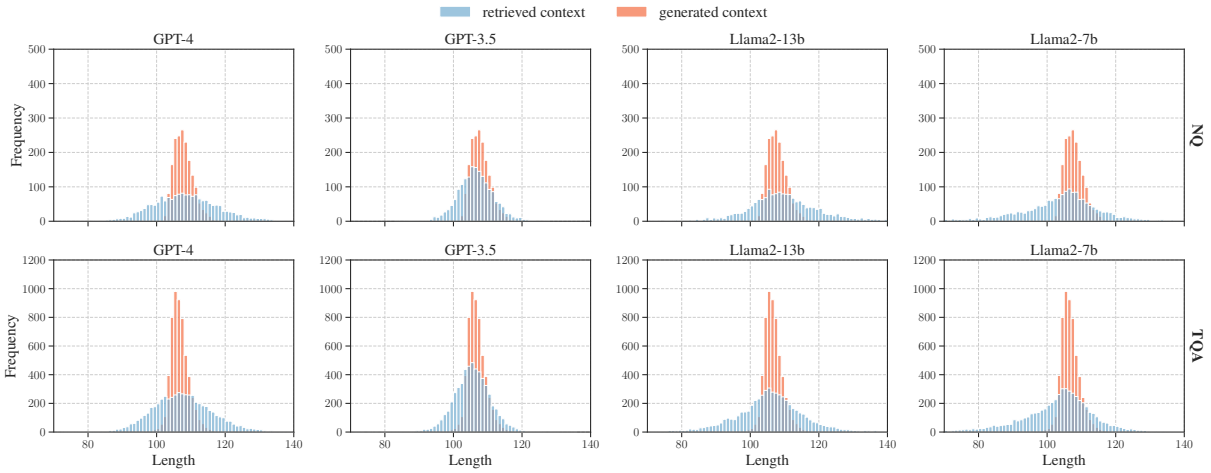



Figure 9: Length distribution of generated and retrieved contexts on different datasets with different generator models.

| Reader | Generator | NQ-AIR | | | | TQA-AIR | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\rho_{\text{gen}}$ | $\rho_{\text{ret}}$ | $\rho_{\text{llm}}$ | DiffGR | $\rho_{\text{gen}}$ | $\rho_{\text{ret}}$ | $\rho_{\text{llm}}$ | DiffGR |
| GPT 4 | GPT 4 | 66.08 | 18.71 | 2.05 | 0.5586 | 76.67 | 13.89 | 2.78 | 0.6933 |
| GPT 4 | GPT 3.5 | 66.40 | 23.62 | 1.84 | 0.4752 | - | - | - | - |
| GPT 4 | Llama2-13b | 58.63 | 28.84 | 2.85 | 0.3406 | - | - | - | - |
| GPT 4 | Llama2-7b | 62.20 | 26.42 | 2.39 | 0.4037 | - | - | - | - |
| GPT 3.5 | GPT 4 | 65.25 | 16.08 | 0.95 | 0.6047 | 68.78 | 17.72 | 3.38 | 0.5902 |
| GPT 3.5 | GPT 3.5 | 67.83 | 12.91 | 0.88 | 0.6802 | 72.05 | 16.15 | 1.55 | 0.6338 |
| GPT 3.5 | Llama2-13b | 67.22 | 15.34 | 1.20 | 0.6284 | 70.21 | 16.38 | 2.61 | 0.6217 |
| GPT 3.5 | Llama2-7b | 64.55 | 16.24 | 1.13 | 0.5979 | 74.14 | 13.65 | 1.87 | 0.6890 |
| Llama2-13b | GPT 4 | 61.72 | 25.73 | 2.51 | 0.4115 | 57.28 | 33.01 | 1.29 | 0.2688 |
| Llama2-13b | GPT 3.5 | 62.39 | 26.76 | 1.08 | 0.3996 | 61.88 | 26.34 | 1.71 | 0.4029 |
| Llama2-13b | Llama2-13b | 69.92 | 18.67 | 1.43 | 0.5785 | 72.55 | 17.75 | 1.96 | 0.6069 |
| Llama2-13b | Llama2-7b | 69.16 | 18.96 | 1.87 | 0.5697 | 75.04 | 15.54 | 1.62 | 0.6569 |
| Llama2-7b | GPT 4 | 54.28 | 24.43 | 1.46 | 0.3793 | 47.13 | 33.97 | 3.59 | 0.1622 |
| Llama2-7b | GPT 3.5 | 52.24 | 27.29 | 1.80 | 0.3138 | 54.31 | 28.76 | 2.20 | 0.3075 |
| Llama2-7b | Llama2-13b | 61.65 | 19.46 | 3.17 | 0.5202 | 64.46 | 20.76 | 1.86 | 0.5128 |
| Llama2-7b | Llama2-7b | 60.38 | 21.16 | 2.20 | 0.4810 | 66.25 | 18.40 | 3.12 | 0.5653 |

Table 11: The columns labeled $\rho_{\text{gen}}$, $\rho_{\text{ret}}$, and $\rho_{\text{llm}}$ respectively represent the proportion (%) of responses that match the answer provided by generated contexts, retrieved contexts, and internal parametric knowledge of LLMs.

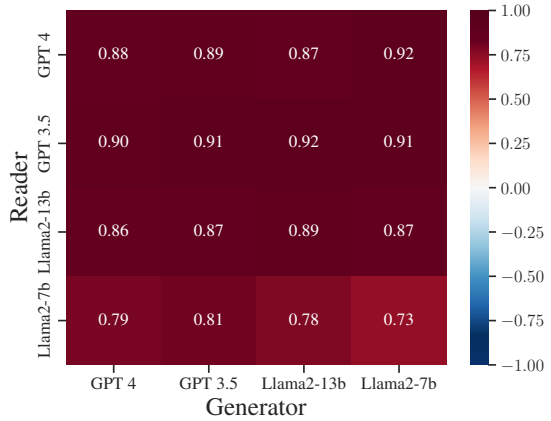



Figure 10: DiffGR with different (reader, generator) pairs on their corresponding NQ-AIG datasets.

| Order | NQ-AIR | TQA-AIR |
|---|---|---|
| generated-first | 0.699 | 0.682 |
| retrieved-first | 0.665 | 0.556 |
| random | 0.691 | 0.586 |

Table 12: DiffGR with different context order on NQ-AIR and TQA-AIR datasets. GPT-3.5 serves as both the generator and reader.

As shown in Table 12, across all context orders, LLMs consistently show a strong tendency to favor generated contexts. When the retrieved context is positioned first, there is a slight reduction in DiffGR. This reduction may result from the LLMs' preference for generated contexts being partially offset by their bias towards the top context (Liu et al., 2023; Xie et al., 2023).

| | NQ-AIR | TQA-AIR |
|---|---|---|
| Llama2-13b | 632 | 702 |
| GPT-3.5 | 323 | 181 |

Table 13: The number of samples filtered with a BERTScore difference of less than 0.05.

### B.4 More Results about Effect of Context Completeness

**(a) Length Distribution.** In §5.3, we employ three methods, "Nature", "Trunc." and "S-Trunc.", to vary the completeness of generated contexts, while controlling the length at the same time. Figure 11 illustrates the length distribution for generated contexts corresponding to these methods. From the results, we can observe that the contexts generated by original GenRead (Yu et al., 2022) are significantly longer compared to the retrieved contexts.

**(b) Similarity and Length Control.** Table 13 presents the data quantity after filtering out samples where three types of generated contexts exhibit significant differences in similarity. Table 14 illustrates the average similarity and completeness of three types of generated contexts. "Nature", "Trunc." and "S-Trunc." result in contexts with Comparable average length and similarity, with preliminary differences in completeness.

#### B.4.1 Cases about Completeness

Table 15 provides some examples to facilitate the understanding of completeness. From the cases, we observe that retrieved contexts and "Trunc." often contain incomplete sentences. Additionally, compared to "S-Trunc", "Nature" typically exhibits greater semantic completeness. Specifically, "Nature" often encompasses a full logical structure of an article, including an introduction, discussion, and conclusion, whereas "S-Trunc" may terminate abruptly.

### B.5 More Results about Effect of Text Similarity

**(a) Similarity Metric.** We employ Jaccard similarity to assess the term-based overlap, and BERTScore (Zhang et al., 2020) for evaluating the semantic similarity between contexts and questions. To mitigate the effect of length discrepancies between contexts and questions, we calculate the similarity at the sentence level and then aggregate them to derive the overall context-question similarity. In this work, we adopt a maximum aggregation strategy due to the single-hop nature of the NQ and TQA datasets, where the majority of questions can be answered using a small subset of sentences. We also try the average aggregation strategy and observe similar results.

Figure 12 illustrates the distribution of similarity when employing maximum and average aggregation methods. It is observable that the generated contexts exhibit a markedly higher degree of similarity regardless of the aggregation method used. Furthermore, this disparity in similarity is more pronounced with maximum aggregation, as contexts typically contain sentences that are irrelevant, which dilute the similarity scores when an average aggregation is applied.

**(b) Similarity Distribution.** Figure 13 and 14 show the similarity distribution of retrieved and generated contexts across various generators. All LLM-generated contexts exhibit a higher similarity over retrieved contexts.

**(c) Effect of Similarity.** Figure 15 demonstrates a general trend that on slices with a smaller average similarity gap, LLMs exhibit a reduced preference for generated context.

**(d) Cases about Similarity.** Table 16 shows examples that contain contexts with different similarities to the question. The contexts with high similarity typically directly support answering by repeating the phrasing in the question. Conversely, the contexts with low similarity introduce more challenges, often necessitating an understanding of synonyms and even some inferences. These observations indicate that text similarity can partly reflect the relevance between a question and a context, as well as the difficulty the LLM encounters in identifying potential answers.

### B.6 The Effectiveness of Exactly Matching

Table 17 demonstrates that the proportion of "Others" is significantly lower relative to the disparities between "Gen" and "Ret", its impact on the conclusions of this paper is negligible.

| Context | | Completeness | | NQ-AIR | | | TQA-AIR | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sentence | Semantic | Length | Jaccard | BERTScore | Length | Jaccard | BERTScore |
| Llama2-13b | Retrieved | ✗ | ✗ | 107.8 | 0.1157 | 0.8534 | 106.6 | 0.1832 | 0.8630 |
| | Nature | ✔ | ✔ | 109.4 | 0.1859 | 0.8755 | 105.0 | 0.3178 | 0.8909 |
| | S-Trunc | ✔ | ✗ | 106.1 | 0.1734 | 0.8719 | 105.0 | 0.2868 | 0.8833 |
| | Trunc | ✗ | ✗ | 107.8 | 0.1765 | 0.8726 | 106.6 | 0.2890 | 0.8837 |
| GPT-3.5 | Retrieved | ✗ | ✗ | 107.1 | 0.1162 | 0.8552 | 106.3 | 0.1836 | 0.8622 |
| | Nature | ✔ | ✔ | 105.7 | 0.1748 | 0.8730 | 105.0 | 0.3993 | 0.9044 |
| | S-Trunc | ✔ | ✗ | 105.8 | 0.1733 | 0.8730 | 104.8 | 0.4024 | 0.9044 |
| | Trunc | ✗ | ✗ | 107.1 | 0.1769 | 0.8736 | 106.3 | 0.4027 | 0.9044 |

Table 14: Average length and similarity of contexts for GPT-3.5 and Llama2-13b models. The three types of generated contexts exhibit similar average lengths and similarity, with the primary distinction being their completeness.

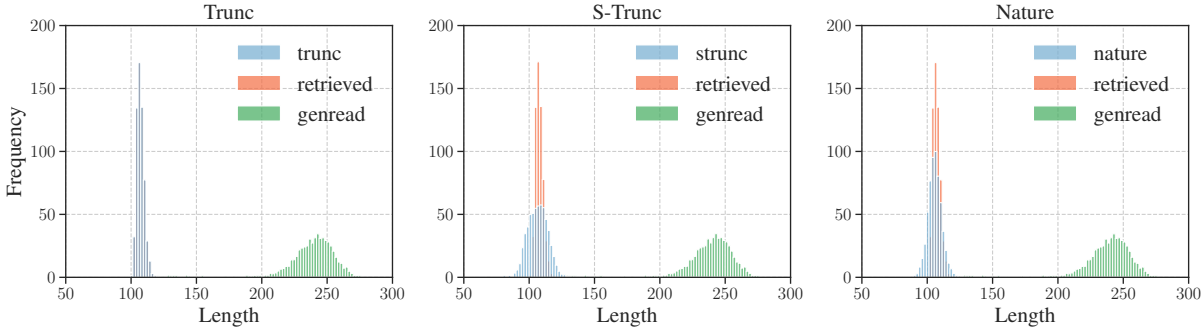



Figure 11: Length distribution of generated and retrieved contexts on the NQ dataset with GPT-3.5 as the generator. "genread" represents the contexts generated by the original GenRead method (Yu et al., 2022). "trunc", "strunc", and "nature" are the generated contexts using three different methods to control the length.

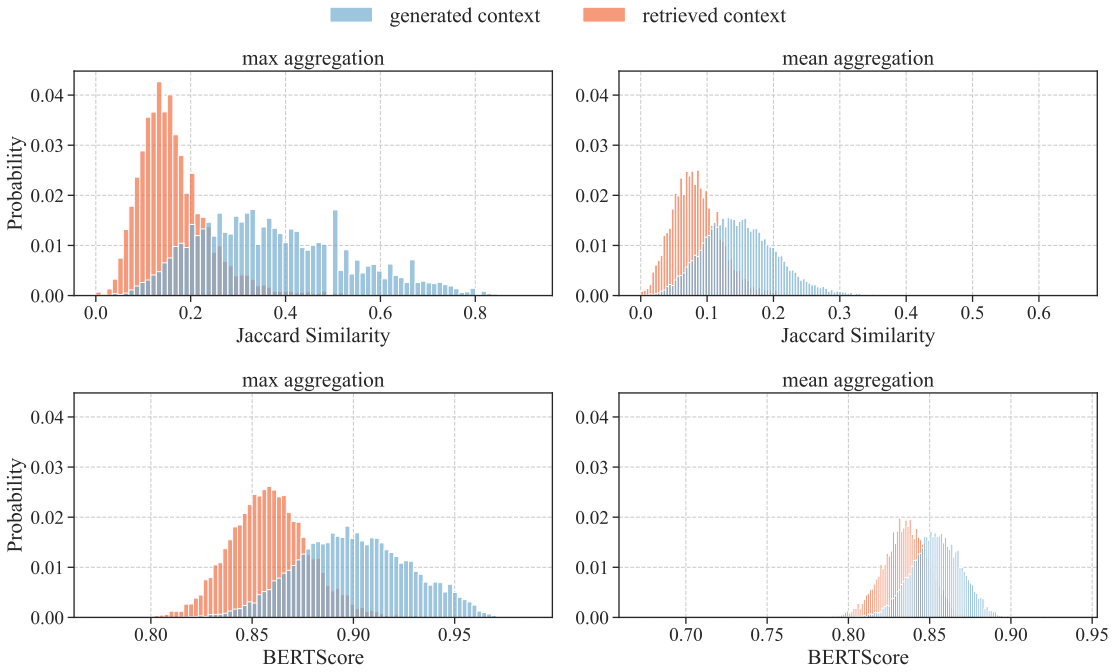



Figure 12: Similarity distribution with maximum or mean aggregation strategies. Generated contexts consistently exhibit higher similarity across two aggregation strategies.

| Example |
| --- |
| **Question:** In World War I, what did the Allies call the defence line that was known as the 'Siegfried Line' to the Germans?<br>**Answer:** ['Hindenberg Line', 'Hindenburg line', 'Battles of the Hindenburg Line', "'HINDENBURG LINE'", 'Advance to the Hindenburg Line', 'Hindenburg Line']<br>**Retrieved context:** Title: Siegfried Line Content: Siegfried Line The term Siegfried Line refers to two different German defensive lines, one during the First World War and the other during the Second World War. The original was a line of forts and tank defences built by Germany in northern France during 1916–17, known in German as the Siegfriedstellung ("Siegfried Position"). The British referred to it as the "Hindenburg Line", a name subsequently adopted by the rest of the Allies. After WW1 the German Fortifications East of the Rhein were required to be demolished due to the Treaty of Versailles. The second, commonly known in German as the<br>**Generated context (Nature):** The Allies referred to the German defence line known as the 'Siegfried Line' as the 'Westwall' during World War I. Constructed between 1936 and 1940, the Westwall was a system of fortifications built along Germany's western border. It stretched from the Swiss border to the Dutch border, covering approximately 630 kilometers. The line consisted of concrete bunkers, tank traps, and anti-tank ditches, aiming to deter any potential invasion from the west. The Allies recognized the formidable nature of this defensive barrier and thus referred to it as the Westwall, while the Germans named it the Siegfried Line, after the mythical hero Siegfried from German folklore.<br>**Generated context (Trunc.):** The Allies referred to the German defence line known as the "Siegfried Line" as the "Westwall" during World War I. The Westwall was a series of fortifications constructed by Germany along its western border, stretching from the Swiss border to the Netherlands. It was named after the German Emperor, Wilhelm II, who was commonly referred to as "Kaiser Wilhelm" or "Siegfried" by the Allies. The construction of the Westwall began in the 1930s as a response to the growing tensions in Europe and the threat of war. The line consisted of a network of bunkers, pillboxes, tank traps, and other defensive structures, strategically positioned<br>**Generated context (S-Trunc.):** The Allies referred to the German defence line known as the "Siegfried Line" as the "Westwall" during World War I. The Westwall was a series of fortifications constructed by Germany along its western border, stretching from the Swiss border to the Netherlands. It was named after the German Emperor, Wilhelm II, who was commonly referred to as "Kaiser Wilhelm" or "Siegfried" by the Allies. The construction of the Westwall began in the 1930s as a response to the growing tensions in Europe and the threat of war. The line consisted of a network of bunkers, pillboxes, tank traps, and other defensive structures, strategically positioned to impede any potential invasion from the west. |

Table 15: Examples with retrieved contexts and generated contexts. "Nature", "Trunc." and "S-Trunc." represent three types of generated contexts with different completeness. Retrieved contexts often contain incomplete sentences.

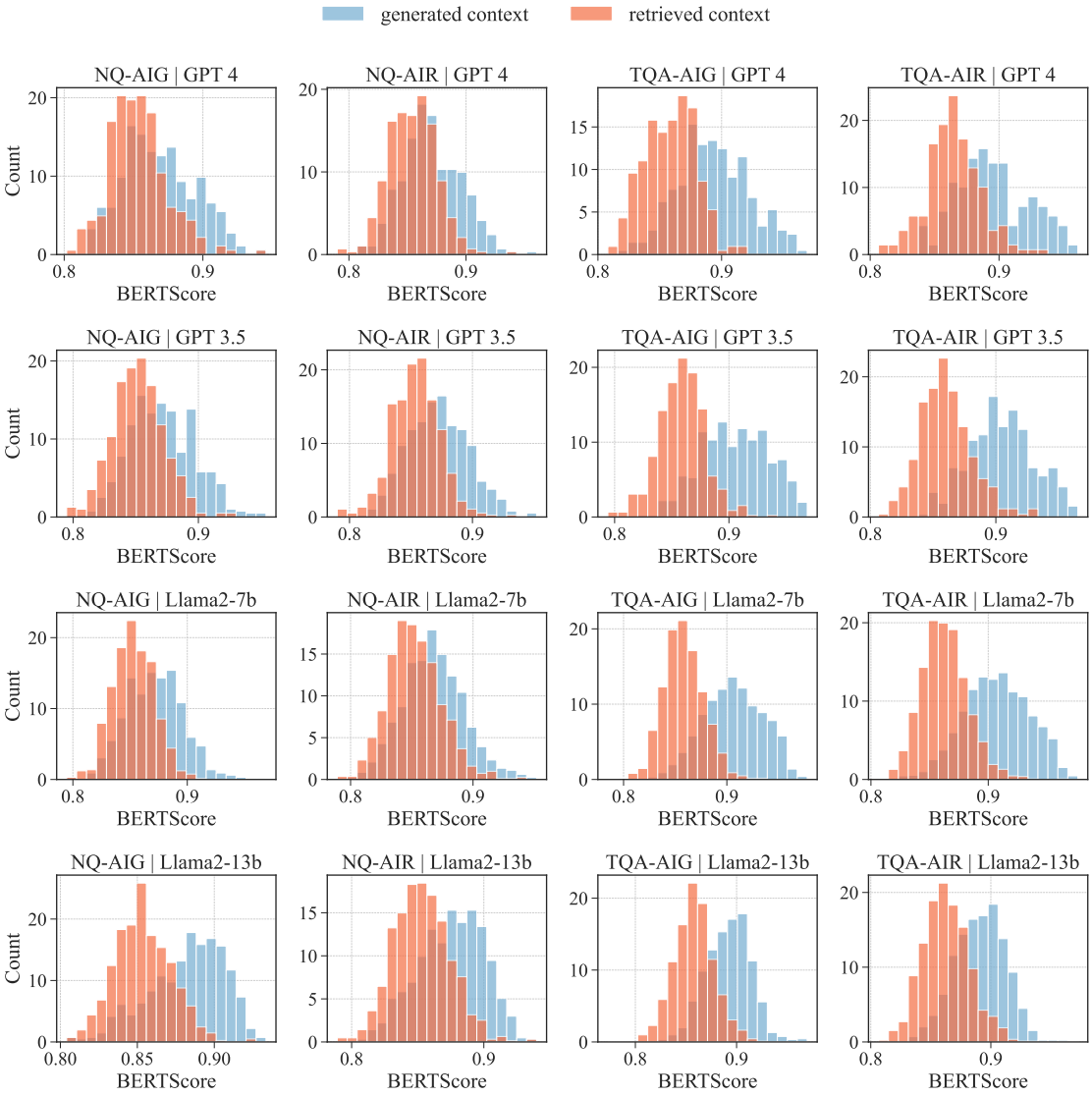



Figure 13: BERTScore distribution of retrieved contexts and contexts generated by different LLMs. All LLM-generated contexts exhibit a higher similarity over retrieved contexts.

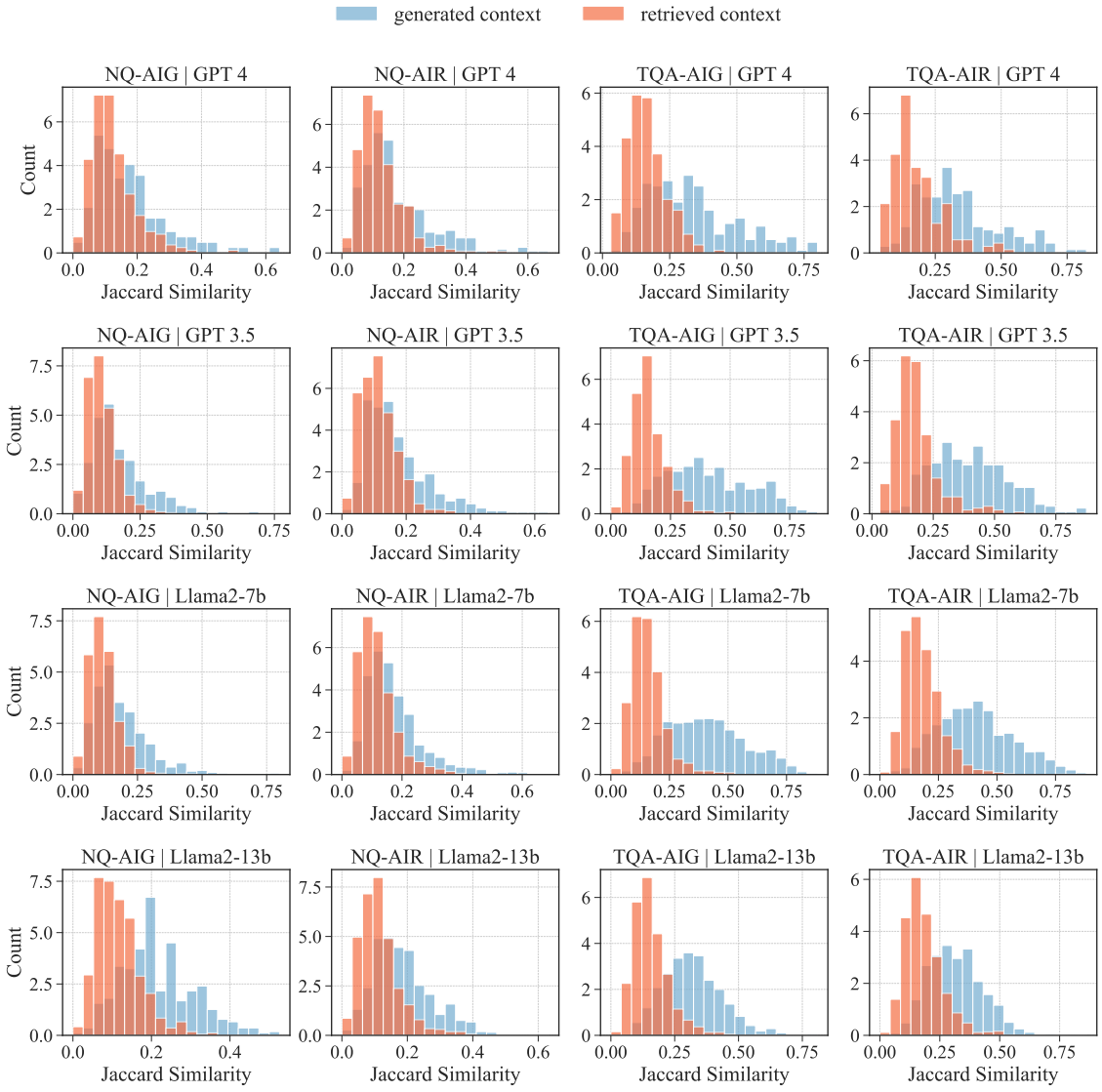
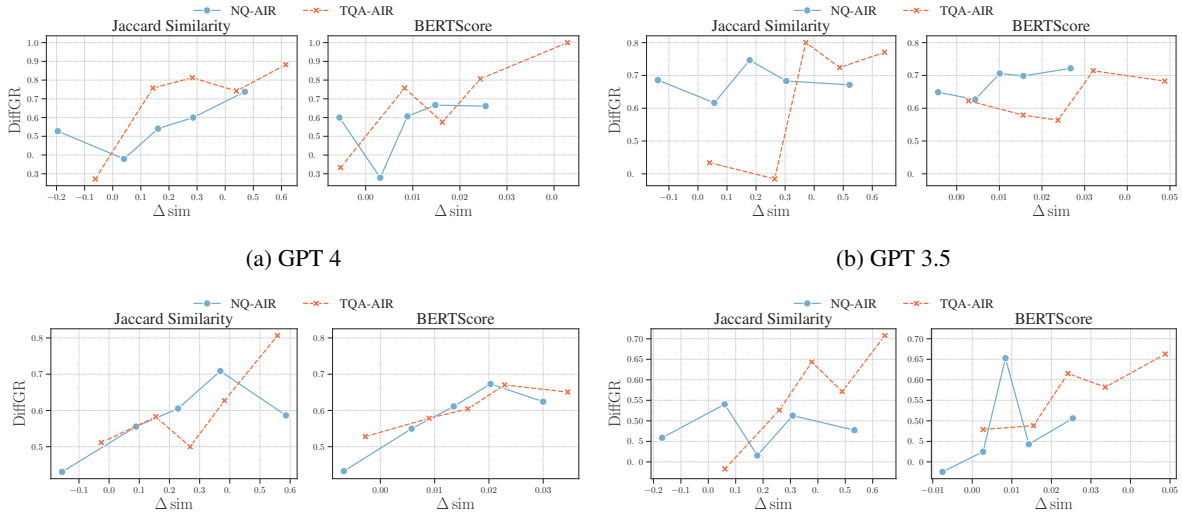



Figure 14: Jaccard Similarity distribution of retrieved contexts and contexts generated by different LLMs. All LLM-generated contexts exhibit a higher similarity over retrieved contexts.




(a) GPT 4

(b) GPT 3.5

(c) Llama2-13b

(d) Llama2-7b

Figure 15: The generation ratio in slices with different average $\Delta$ sim. $\Delta$ sim is the difference in similarity between the generated context and the retrieved context. The LLM marked in the figure serves both as a generator and a reader.

| | TQA-AIR Example | TQA-AIG Example |
|---|---|---|
| Question | Between 1959 and 1967 which city was the capital of Pakistan (Islamabad was being built)? | Who is the most successful UK solo artist in the USA? |
| Golden Answer | Rawalpindi | Elton John |
| Generated Context | Between 1959 and 1967, the capital of Pakistan was Karachi. Karachi is the largest city in Pakistan and is located on the southern coast of the country...<br>**Jaccard Similarity**: 0.47<br>**BertScore**: 0.93 | Elton John is the most successful UK solo artist in the USA. Born Reginald Kenneth Dwight in 1947, he adopted the stage name Elton John in the late 1960s...<br>**Jaccard Similarity**: 0.69<br>**BertScore**: 0.93 |
| Retrieved Context | ...was first shifted temporarily to Rawalpindi in the early 60s, and then to Islamabad when essential development work was completed in 1966...<br>**Jaccard Similarity**: 0.16<br>**BertScore**: 0.85 | ...In 2009, Jay Sean's single "Down" reached the number one spot on the "Billboard" Hot 100 and sold millions in the United States, making him the most successful male UK urban artist in US chart historyät the time...<br>**Jaccard Similarity**: 0.14<br>**BertScore**: 0.86 |
| Model output | Karachi | Elton John |

Table 16: Some examples where both the generator and reader are GPT-3.5. We highlight the incorrect candidate answers in the context in pink, and the correct answers in the context in green.

| Reader | Generator | NQ-AIG | | | NQ-AIR | | | TQA-AIG | | | TQA-AIR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gen | Ret | Others | Gen | Ret | Others | Gen | Ret | Others | Gen | Ret | Others |
| GPT 4 | GPT 4 | 0.9125 | 0.0589 | 0.0286 | 0.7379 | 0.1743 | 0.0878 | 0.9387 | 0.0304 | 0.031 | 0.7651 | 0.1762 | 0.0587 |
| GPT 3.5 | GPT 3.5 | 0.9177 | 0.0449 | 0.0374 | 0.7083 | 0.1470 | 0.1447 | 0.9347 | 0.026 | 0.0393 | 0.7332 | 0.1775 | 0.0893 |
| Llama2-13b | Llama2-13b | 0.8966 | 0.0500 | 0.0534 | 0.7216 | 0.1811 | 0.0973 | 0.9071 | 0.0433 | 0.0496 | 0.7212 | 0.1918 | 0.0870 |
| Llama2-7b | Llama2-7b | 0.7041 | 0.1082 | 0.1876 | 0.6148 | 0.2071 | 0.1781 | 0.7973 | 0.0927 | 0.1100 | 0.6555 | 0.2101 | 0.1344 |

Table 17: "Gen" denotes the proportion of responses that match the candidate answer within generated contexts, whereas "Ret" refers to the proportion of matching the candidate answer within retrieved contexts. "Others" encompasses the proportion of responses that do not align with either category. *Given that the proportion of "Others" is significantly lower relative to the disparities between "Gen" and "Ret", its impact on the conclusions of this paper is negligible.*