Spiral of Silence: How is Large Language Model Killing Information Retrieval?—A Case Study on Open Domain Question Answering

Xiaoyang Chen[1,2], **Ben He**[1,2✉], **Hongyu Lin**[2✉], **Xianpei Han**[2],
Tianshu Wang[2,3], **Boxi Cao**[2,1], **Le Sun**[2✉], **Yingfei Sun**[1✉]
[1]University of Chinese Academy of Sciences
[2]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
[3]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
chenxiaoyang19@mails.ucas.ac.cn, {benhe, yfsun}@ucas.ac.cn
{hongyu, xianpei, tianshu2020, boxi2020, sunle}@iscas.ac.cn

Abstract

The practice of Retrieval-Augmented Generation (RAG), which integrates Large Language Models (LLMs) with retrieval systems, has become increasingly prevalent. However, the repercussions of LLM-derived content infiltrating the web and influencing the retrieval-generation feedback loop are largely uncharted territories. In this study, we construct and iteratively run a simulation pipeline to deeply investigate the short-term and long-term effects of LLM text on RAG systems. Taking the trending Open Domain Question Answering (ODQA) task as a point of entry, our findings reveal a potential digital "Spiral of Silence" effect, with LLM-generated text consistently outperforming human-authored content in search rankings, thereby diminishing the presence and impact of human contributions online. This trend risks creating an imbalanced information ecosystem, where the unchecked proliferation of erroneous LLM-generated content may result in the marginalization of accurate information. We urge the academic community to take heed of this potential issue, ensuring a diverse and authentic digital information landscape.[1]

1 Introduction

The integration of Large Language Models (LLMs) (OpenAI, 2022, 2023; Touvron et al., 2023; Google, 2023) is reshaping the online information landscape, making text generation easier, increasing content production, enhancing personalized knowledge assistance, and enabling advanced fake news creation. Schick (2020) suggest that by 2026, synthetic content could dominate up to 90% of the web. CounterCloud[2] shows that a single developer can create an AI fake news factory cheaply and convincingly. AI-driven content generation is rapidly becoming commonplace, impacting how content is produced and shared (Goldstein et al., 2023; Pan et al., 2023; Dai et al., 2023b). These developments pose novel challenges and opportunities for information retrieval (IR) and generation, especially for Retrieval-Augmented Generation (RAG) systems, which combine both capabilities (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023). As text produced by large language models continues to flood the internet and is indexed by search systems, the enduring effects of such text on the retrieval-generation process grow more ambiguous, and the future landscape of the information environment is yet to be determined.

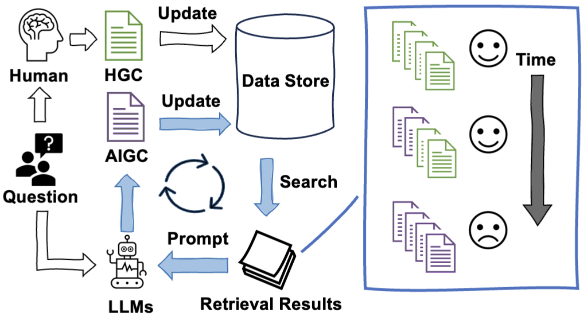


Figure 1: The evolution of RAG systems after introducing LLM-generated texts, where the "Spiral of Silence" effect gradually emerges.

In our research, we focus on the **effects of LLM-generated text on RAG systems**. As shown in Figure 1, we construct a **pipeline that simulates the continuous influx of LLM-generated text into web datasets** and assess its impact on the performance of RAG through iterative runs. To evaluate the RAG performance in the simulation process, we adopt the Open Domain Question Answering (ODQA) task as our evaluative benchmark due to its recent surge in research popularity as an effective test of both retrieval accuracy and generation quality (Pan et al., 2023; Chen et al., 2023). We

[1]We release the resources at https://github.com/VerdureChen/SOS-Retrieval-Loop
[2]https://countercloud.io/?page_id=307

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14930–14951
August 11-16, 2024

employ widely used retrieval and re-ranking methods to supply the context necessary for LLMs to generate answer documents. Upon evaluating these documents, we integrate them into the text corpus for subsequent retrieval-generation cycles. This process is repeated multiple times to monitor and assess the emerging patterns. Experimental results show that LLM-generated text has an immediate effect on RAG systems, generally improving retrieval outcomes while producing varied effects on QA performance. However, over the long term, a marked decrease in retrieval effectiveness emerges, while the QA performance remains unaffected.

Further examination reveals a bias in search systems towards LLM-generated texts, which consistently rank higher than human-written content. As LLM-generated texts increasingly dominate the search results, the visibility and influence of human-authored web content diminish, fostering a digital "**Spiral of Silence**" effect. This effect aptly explains what we observe in our simulations and reveals the potential negative impact of LLM-generated texts on the information ecosystem: while LLM-generated texts sometimes provide a more effective IR experience in the short term, in the long term they may lead to the invisibility of human-authored content, the homogenization of search results, and the inaccessibility of certain accurate information, thereby adversely affecting public knowledge acquisition and decision-making.

The contributions of this paper are threefold: 1) We propose an iterative pipeline to investigate the short-term and long-term impacts of LLM-generated text on RAG systems. 2) We study the potential emergence of a "Spiral of Silence" phenomenon within RAG systems. 3) We analyze the implications of this phenomenon, offering a new perspective on the dynamic interplay between LLM-generated content and RAG systems.

2 Related Works

Retrieval Augmented Generation. RAG systems have been extensively analyzed, demonstrating retrieval's role in enhancing language model efficacy (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Ram et al., 2023). These systems also curtail LLMs' hallucinations during text generation (Ji et al., 2023; Huang et al., 2023a) and reduce knowledge obsolescence (He et al., 2023). Applied in ODQA (Izacard and Grave, 2021; Trivedi et al., 2023; Liu et al., 2023a) and other tasks (Cai et al., 2019; Zhou et al., 2023), current research explores LLMs' output accuracy against specific contexts (Adlakha et al., 2023), robustness to extraneous information (Chen et al., 2023), and the effects of output integration strategies (Liu et al., 2023b). Our study aims to provide a novel perspective to observe and predict the potential trajectory and impact of its future development.

Effects of AIGC. Advances in Artificial Intelligence Generated Content (AIGC) have significantly impacted society and technology. LLMs facilitate creating content to combat misinformation (Xu et al., 2023; Chen and Shu, 2023) but can also produce damaging content (Huang et al., 2023b). The potential biases and discrimination in AIGC have garnered widespread attention (Liang et al., 2021; Zhuo et al., 2023). Shumailov et al. (2023) and Alemohammad et al. (2023) show that LLMs trained on self-generated data degrade without fresh real-world input. Pan et al. (2023) investigated the impact of erroneous information generated by LLMs on ODQA systems. Dai et al. (2023a) indicated that AI-modified texts might rank higher in search results, potentially affecting the fairness of those outcomes. Our research aims to further explore the short-term and long-term effects on RAG systems when AIGC text is continuously integrated into the search system's datasets.

Spiral of Silence. The "Spiral of Silence" theory (Noelle-Neumann, 1974), is a seminal theory within the field of communications that describes how people may suppress their views to avoid isolation, thus often reinforcing dominant public opinions (Scheufle and Moy, 2000; Liu et al., 2019; Lin et al., 2022). We shift focus to a novel "passive human silence" influenced by LLMs, where rapid AI content production and biased search algorithms potentially marginalize human contributions in public discourse. This theory stands apart from concepts such as "echo chambers", "filter bubbles", and "degenerate feedback loops" prevalent in recommendation systems (Alatawi et al., 2021; Chitra and Musco, 2020; Jiang et al., 2019). While these terms describe the narrowing of informational scope as users engage with algorithmic systems, the "Spiral of Silence" theory proposes a scenario where human users are compelled into silence in public discourse due to the influence of LLMs and IR systems. This phenomenon goes beyond mere selective exposure or algorithmic recommendations. Our study explores how

LLM-generated text might induce a "Spiral of Silence" in RAG systems over time. For further discussion regarding the rationale for applying this theory to our study, please see Appendix A.1.

3 Pipeline Construction

In this section, a simulation framework is designed to explore the potential impacts that texts generated by LLMs may have on RAG systems. This framework models a simplified process that tracks how RAG systems gradually adjust their responses as they accumulate LLM-generated text over time.

3.1 Preliminaries

An RAG system f can be formalized as $f : (Q \times D \times K) \rightarrow S$, where Q is the set of queries, D represents a large collection of documents, K is the knowledge within the LLM, and S is the set of text outputs generated by the system. For a particular query $q \in Q$, the goal of the RAG system is to find a mapping $f(q, D, K) = s$ that produces a response text $s \in S$ satisfying the query q. This process involves two stages:

Retrieval Stage, executed by the retrieval function R, is formally defined as $R : (Q \times D) \rightarrow D'$, where $D' \subseteq D$ represents the subset of documents judged by R to be most relevant to the query q.

Generation Stage, executed by the generation function $G : (P \times Q \times D' \times K) \rightarrow S$. Its task is to utilize the prompt $p \in P$, the query $q \in Q$, the related document subset D', and the knowledge of LLMs K to construct the answer s.

Within the entire RAG system f, the functions R and G act in series to form a process expressed as $f(q, D, K) = G(p, q, R(q, D), K)$. In this manner, the RAG system integrates the precision of IR with the richness of LLMs to provide information-rich content when answering questions.

3.2 Simulation Process

Our simulation process starts with a pure human-authored text dataset and gradually introduces the LLM-generated text, observing how this change over time affects the RAG system. Adhering to the specifications outlined in Section 3.1, the RAG architecture is instantiated and expanded with additional details. In the **retrieval stage**, we apply sparse and dense retrieval strategies to obtain a candidate document set that is relevant to the query. Additionally, we also have the option to perform a re-ranking of the candidate documents to further optimize the process. In the **generation stage**, we use the LLMs which are widely used to generate responses. To accurately simulate the evolution process of the RAG system, we specifically use an **iteratively updated indexing structure** that supports incorporating the newly generated LLM text into the index in each iteration, keeping the dataset updated for subsequent retrieval and evaluation.

Specifically, the iterative simulation process unfolds as follows: 1) **Baseline Establishment**: Utilizing an initial dataset comprised of human-authored text unaffected by LLM (D_0), ascertain the performance of a benchmark RAG pipeline. 2) **Zero-shot Text Introduction**: The baseline dataset D_0 is enriched with text set $T_{\text{LLM}}^{(\text{zero-shot})}$ generated by LLMs in zero-shot manner, yielding $D_1 = D_0 \cup T_{\text{LLM}}^{(\text{zero-shot})}$. This simulates the evolution of users' application of LLMs from initial zero-shot deployments to sophisticated RAG configurations. 3) **Retrieval and Re-ranking**: For each query q, a subset of documents D'_i is retrieved from the dataset D_i through a retrieval and optional re-ranking step $R(q, D_i) \rightarrow D'_i$. The retrieval function R remains constant throughout the experimental process to control variables. 4) **Generation Phase**: Answers S are generated using the LLMs ($G(p, D', q, K) \rightarrow s$) with a uniform prompt p in the experiment. 5) **Post-processing Phase**: Post-process S to obtain S', removing text fragments that may expose the identity of the LLMs. 6) **Index Update**: Integrate S' into D_i to update the dataset to D_{i+1}. 7) **Iterative Operation**: Repeat steps 3 to 6 for each new dataset D_{i+1}, until the required number of iterations t is reached.

The pseudo-code for this process is presented in Appendix A.2. Through the simulation process, we observe how LLM-generated text influences the RAG systems and how this impact evolves with data accumulation. While the main simulation assumes that the LLMs remain static due to their relatively infrequent update cycles, we also conduct experiments on the effects of LLM evolution over time in Appendix A.6. For prompt and post-processing details, see Appendix A.9 and A.3.

4 Experiment

Datasets and Metrics. We conduct experiments on commonly used ODQA datasets, including **NQ** (Kwiatkowski et al., 2019), **WebQ** (Berant et al., 2013), **TriviaQA** (Joshi et al., 2017), and **PopQA** (Mallen et al., 2022). We preprocess the datasets following Yu et al. (2023) and Zhang et al. (2023). Given the constraints on experimental resources, we randomly select 200 samples from each test set. When evaluating the retrieval phase, we utilize **Acc@5** and **Acc@20** following Karpukhin et al. (2020). These metrics assess the proportion of questions where the correct answers appear in the top 5 or top 20 retrieval results, respectively. For the answer quality of the LLM output for each iteration, we follow Chen et al. (2023) by applying the **Exact Match (EM)** metric, which checks if the correct answer is fully contained within the generated text. Furthermore, in Section 5, we adopt a holistic perspective to examine the RAG pipeline, with a focus on the interaction between the retrieval and generation phases and how the ranking of human-generated texts changes over time.

Retrieval and Re-ranking Methods. In our experiments, we employ a variety of retrieval methods, including the sparse model BM25, the contrastive learning-based dense retriever Contriever (Izacard et al., 2022), the advanced BGE-Base (Xiao et al., 2023) retriever, and the LLM-Embedder (Zhang et al., 2023) designed for LLMs. For the results retrieved using BM25 and BGE-Base, we separately apply the T5-based (Raffel et al., 2020) re-ranking model MonoT5-3B (Nogueira et al., 2020), the UPR-3B (Sachan et al., 2022) which uses the unsupervised capabilities of T0-3B (Sanh et al., 2022), and the BGE-Reranker (Xiao et al., 2023), which is based on the XLM-RoBERTa-Large (Conneau et al., 2020).

Generative Models. Considering the complexity and variability of real-world environments, the text that is continuously integrated into the system may be generated by a variety of LLMs. Our iterative experiments incorporate text produced by a suite of prevalent LLMs. These include GPT-3.5-Turbo (OpenAI, 2022), LLaMA2-13B-Chat (Touvron et al., 2023), Qwen-14B-Chat (Bai et al., 2023), Baichuan2-13B-Chat (Yang et al., 2023), and ChatGLM3-6B (Du et al., 2022). This enables the RAG systems to blend varied linguistic styles and knowledge, leading to results that more closely replicate real-world scenarios. For more implementation details, please refer to Appendix A.4.

5 Results

In this section, we examine both initial and extended iterations within the simulation framework. We define the short-term effect as the immediate effects observed in the first iteration, while the long-term effect is analyzed from the second to the tenth iteration. We investigate the occurrence of the "Spiral of Silence" effect and how RAG systems respond. Under the task settings of ODQA, we analyze the potential influence of the "Spiral of Silence" on RAG systems' response patterns.

5.1 Short-Term Effects on RAG Performance

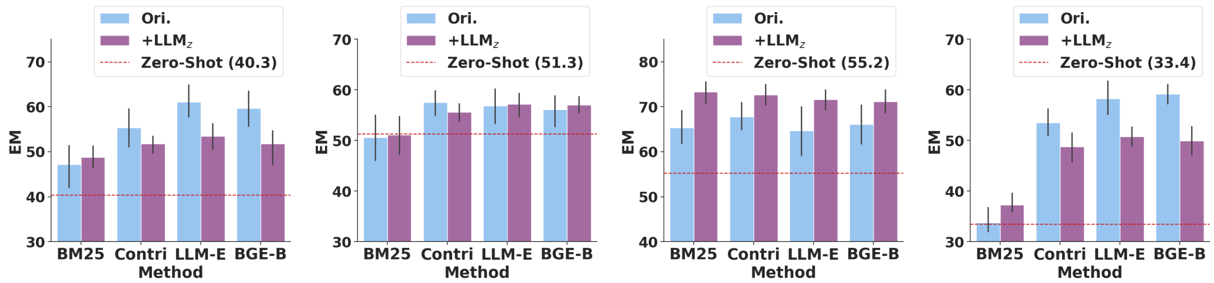
When comparing RAG system results using different retrieval methods on the original dataset versus the augmented one in the first iteration, we observe that: 1) **Immediate Impact of LLM-Generated Text on the RAG System:** The introduction of a minimal amount of LLM-generated text produces immediate effects on both retrieval and QA performance of the RAG system, as shown in Table 1 and Figure 2. Specifically, both retrieval and generation performance exhibit noticeable fluctuations. These changes highlight the sensitivity of the system to even small modifications made by LLM-generated text. 2) **LLM-Generated Text Generally Improves Retrieval Accuracy:** Table 1 reveals that adding LLM-generated responses to a dataset typically enhances the accuracy of retrieval systems, as measured by Acc@5 and Acc@20 metrics. For example, using the BM25 on TriviaQA resulted in accuracy improvements of 31.2% and 19.1% respectively. However, a slight decline in Acc@5 is also observed in certain cases. This suggests a primarily positive, yet complex, impact of LLM-generated text on retrieval accuracy. 3) **The Impact on QA Performance is Mixed:** Due to space constraints, we only present the results of four retrieval methods. As shown in Figure 2, while the RAG system's QA performance typically surpasses the zero-shot LLM outputs, the addition of LLM text can either enhance or impair QA performance depending on the dataset and retrieval strategy. It appears to enhance performance for TriviaQA, but for NQ and PopQA, the effect is detrimental with non-BM25 retrieval methods, suggesting that without significant retrieval enhancement, LLM text inclusion might be counterproductive.

5.2 Long-term Effects on RAG Performance

In this section, we investigate whether the short-term effects are predictive of the long-term behavior of the system. We present the results on NQ and PopQA in Figure 3. For results on other datasets, please refer to Figure 8 in Appendix A.5, where we observe consistent patterns across these datasets.

Model	NQ				WebQ				TriviaQA				PopQA			
	Acc@5		Acc@20		Acc@5		Acc@20		Acc@5		Acc@20		Acc@5		Acc@20	
	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$	Ori.	+LLM$_Z$
BM25	49.0	57.5	67.0	73.5	41.0	51.0*	63.0	71.0	62.5	82.0*	73.0	87.0*	35.5	41.5	51.5	59.5
Contreiver	68.0	68.5	84.0	85.0	66.0	69.5	74.0	80.0	68.0	83.5*	80.5	87.5	62.0	65.0	77.5	79.5
LLM-Embedder	75.5	75.5	86.5	88.0	62.5	72.5*	76.0	79.5	67.5	81.0*	77.5	87.5*	70.0	67.5	79.5	82.0
BGE$_{base}$	77.0	73.0	86.0	86.0	65.5	71.5	77.0	80.0	69.5	81.5*	80.0	87.5*	72.0	70.0	83.0	84.5
BM25+UPR	63.0	66.5	73.5	78.0	57.0	68.0*	68.5	75.0	71.5	83.0*	78.0	89.0*	57.5	61.5	60.0	67.0
BM25+MonoT5	66.5	69.0	74.5	80.5	62.0	67.5	69.5	76.0	72.0	83.5*	78.0	88.0*	53.5	58.5	59.5	66.5
BM25+BGE$_{reranker}$	68.0	69.5	76.5	81.0	64.5	68.5	71.0	76.0	72.5	84.0*	78.0	88.5*	54.0	61.0	60.0	67.5
BGE$_{base}$+UPR	75.5	71.5	87.5	88.0	64.0	69.0	77.0	79.5	76.0	84.0*	84.5	89.5	76.0	71.0	84.5	84.5
BGE$_{base}$+MonoT5	75.0	70.5	86.5	86.5	68.5	72.0	78.0	81.5	77.0	83.5	83.5	89.5	72.0	72.5	85.5	86.0
BGE$_{base}$+BGE$_{reranker}$	69.0	68.0	84.0	84.5	67.5	70.5	78.0	81.5	72.5	83.5*	82.0	88.0	73.0	70.0	84.0	85.0

Table 1: Short-term retrieval performance. A blue background indicates a decrease in retrieval results after the incorporation of LLM-generated text, while a purple background signifies an increase. The deeper the color, the larger the discrepancy from the original results. Statistical significance at 0.05 relative to origin is marked with *.



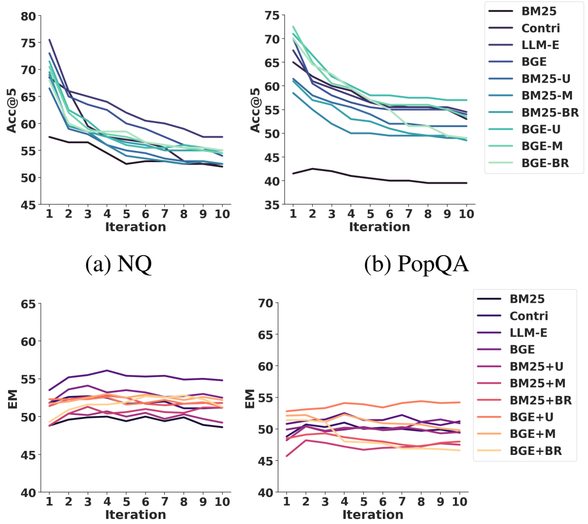
(a) NQ (b) WebQ (c) TriviaQA (d) PopQA

Figure 2: Short-Term QA performance. For each retrieval method, we present both the average performance and the range of variation exhibited by five LLMs. A red dashed line symbolizes the average EM score for zero-shot question generation by LLMs. "Ori." and "+LLM$_Z$" represent the average EM values when models use the original dataset or a dataset enhanced with LLM-generated texts as context, respectively. Retrieval methods are abbreviated: "Contri" for Contriever, "LLM-E" for LLM-Embedder, and "BGE-B" for BGE$_{base}$.

We find that: 1) **Decreased Retrieval Effectiveness Over Time:** Figures 3a and 3b show a general decline in Acc@5 across successive iterations for most methods, with an average drop of 21.4% for NQ and 19.4% for PopQA from the first iteration to the last, except for a temporary improvement in BM25 during the second iteration on PopQA. This trend signals that the retrieval quality boost provided by LLM-generated text may be transient, with a propensity for degradation over time. 2) **Stability in QA Performance Despite Retrieval Decline:** Contrary to expectations, the QA performance does not mirror the retrieval accuracy's decrease. As shown in Figure 3c and 3d, the EM exhibit slight variations but generally maintain their level throughout the iterations. While a diminished retrieval accuracy intuitively seems to undermine the system's capacity to output correct answers, this does not unequivocally translate into a decline in QA efficacy. In subsequent sections, we will delve deeper into the reasons behind these observations and examine the complex dynamic relationship that may exist between retrieval and QA performance.

5.3 Spiral of Silence

In the context of LLM-augmented RAG systems, we have observed a rapid shift in response to the integration of LLM-generated text, a decline in retrieval performance over time, and stability in QA performance despite retrieval decline. To explain these phenomena, we draw on the theory of the "Spiral of Silence" as posited by Noelle-Neumann (1974), extending its principles to the behavior of RAG systems enhanced by LLMs. To explore the presence of a "Spiral of Silence" phenomenon, we propose **three Hypotheses** for investigation. **(H1): Dominance of LLM-Generated Texts:** Retrieval models are more likely to prioritize LLM-generated text in search results, which could result in LLM-generated text taking a dominant position in the retrieval hierarchy. **(H2): Marginalization of Human-Generated Content:** If human-authored text consistently loses ranking prominence through successive iterations, it may be excluded from the top results until it becomes invisible, thus creating silence. **(H3): Homogenization of Opinions:** The preferential ranking of LLM-generated text could



(a) NQ (b) PopQA

(c) NQ (d) PopQA

Figure 3: Long-Term RAG performance. The upper section illustrates the retrieval outcomes for various methods, while the lower section depicts the average EM across LLMs. Iteration 1 represents the results following the incorporation of zero-shot LLM-generated text. Abbreviated re-ranking methods in the legend are: +U for UPR, +M for MonoT5, and +BR for BGE-Reranker.

culminate in a uniformity of displayed perspectives by the RAG system, potentially sidelining the accuracy or variety of the information.

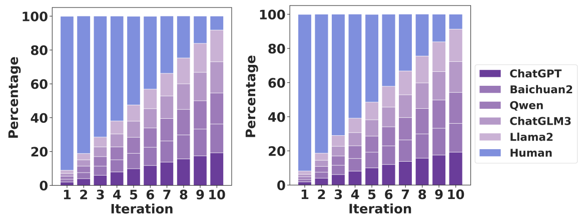
To verify **(H1)**, we analyze Iteration 1 where LLM-generated texts are first introduced to the retrieval system. We calculate the proportion of these texts appearing in the top 5 search results:

$$\mathrm{P} = \frac{\sum_{q \in Q} c_q^{LLM}}{\sum_{q \in Q}(c_q^{LLM} + c_q^{Human})} \times 100\% \qquad (1)$$

where c_q^{LLM} is the count of LLM-generated texts and c_q^{Human} is the count of human-generated texts in the top 5 search results for query q. Table 2 reveals that, even with a modest inclusion of LLM-generated texts, most retrieval models often rank them at the top. This behavior supports the findings of Dai et al. (2023a), where LLM-rewritten texts are preferred by retrieval models over the originals. Our study extends this by directly generating query-specific texts with LLMs. The preference might stem from inherent biases within the system or the actual relevance of the LLM-produced content. **This suggests retrieval systems tend to favor LLM-generated texts, making them more prominent in search results, which can rapidly influence an RAG system's behavior.**

Method	NQ	WebQ	TriviaQA	PopQA
BM25	34.1	19.6	57.6	23.9
Contriever	72.8	75.2	80.1	67.0
LLM-Embedder	68.2	64.6	75.3	70.0
BGE$_{base}$	80.7	84.1	85.6	81.5
BM25+UPR	62.3	49.8	75.7	47.1
BM25+MonoT5	66.2	55.8	83.0	47.1
BM25+BGE$_{reranker}$	64.4	55.2	81.6	46.6
BGE$_{base}$+UPR	74.4	69.3	79.1	71.2
BGE$_{base}$+MonoT5	81.4	84.0	88.4	74.3
BGE$_{base}$+BGE$_{reranker}$	67.2	74.2	83.2	72.8

Table 2: Percentage of LLM-generated documents occupying the top 5 retrieval results, after augmenting each query with five LLM-generated documents. The blue background indicates a majority presence of human-generated documents, while the purple background denotes a predominance of LLM-generated documents.

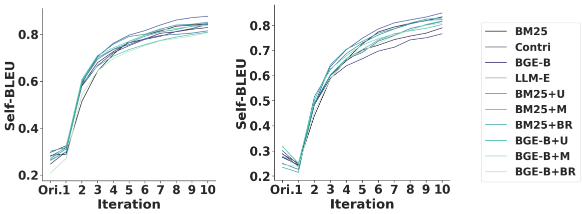


(a) NQ (b) PopQA

Figure 4: Average percentage of texts from various sources within the top 50 search results over multiple iterations across different search methods. For results on WebQ and TriviaQA, please refer to Figure 9 in Appendix A.5.

To validate **(H2)**, we incorporate a **temporal dimension**, observing the percentage change of texts generated by various LLMs and humans within the top 50 search results across different datasets over time. As shown in Figure 4, after ten iterations, the percentage of human-generated texts significantly decreased, falling below 10% for all datasets. **This pattern suggests a sustained diminishing impact of human-contributed texts and hints at the possibility of their eventual exclusion from search results if the trend continues.**

To explore **(H3)**, we examine the risk of potential viewpoint homogenization in the RAG system from both the **diversity** and **accuracy** dimensions during the simulation. **Diversity** is quantified using Self-BLEU (Zhu et al., 2018), which works by comparing a generated text with other generated texts to measure similarity. High similarity results in a high Self-BLEU score, indicating lower diversity and suggesting a convergence of viewpoints. As shown in Figure 5, upon introducing zero-shot LLM-generated texts (Iteration 1), the Self-BLEU scores across different datasets expe-



(a) NQ (b) PopQA

Figure 5: 3-gram Self-BLEU score for the top 5 search results over iterations, from the original dataset (Ori.) to subsequent iterations including LLM-generated texts. For results on WebQ and TriviaQA, please refer to Figure 10 in Appendix A.5.

rience varying degrees of change. However, over more iterations, the Self-BLEU scores for the top 5 results consistently rise and plateau across all datasets, **indicating a significant reduction in textual diversity with each iterative cycle**. Subsequently, we assessed whether the **accuracy** of the top documents returned by the IR system tends toward uniformity over time. Figure 6 charts the number of documents with the correct answer in the top 5 results ("Context Right Num") against the number of queries LLM answers correctly or incorrectly, across the Iteration1, 2, 5, 10. For simplicity, we showcase only the NQ dataset's averaged outcomes, but we find the same trends across other datasets. It indicates that in the initial iterations, fewer correct answer documents (e.g., "Context Right Num" of 0, 1, or 2) typically correlate with a greater number of LLM-answered queries being incorrect (EM=0). Despite this, a significant fraction of the top documents still include the correct answer. When the LLM correctly answers a query (EM=1), the correct answer documents within the top results can range anywhere from 1 to 5. As the iterations continue, the frequency of having 1 to 4 correct answer documents in the top 5 results for each query diminishes, and by the Iteration10, contexts for LLM-correct queries almost always contain the correct answer, whereas contexts for LLM-incorrect answers almost always do not. **This pattern demonstrates a trend towards polarization and uniformity in the accuracy of provided contexts as the RAG system iterates.**

At this point, we have confirmed through our experiments the presence of the "Spiral of Silence" phenomenon, as outlined in three tested hypotheses. Moreover, the dashed lines in Figure 6 represent the total number of correct and incorrect LLM answered queries, along with the Acc@5 retrieval metric over various iterations. The LLM's rate of correct answers remains constant through the iterations, aligning to Section 5.2. However, as iterations advance, correct answers diminish within top documents for LLM-incorrect queries, reducing their contribution to the Acc@5 and thus decreasing retrieval performance. In contrast, for LLM-correct queries, more retrieved documents containing the correct answer do not affect Acc@5 or EM. **Thus, the pattern discussed in Section 5.2 can be explained by the "Spiral of Silence" theory, which accounts for the observed dip in IR results and the sustained QA performance.**

5.4 Effects of "Spiral of Silence" on ODQA

We will delve into a more nuanced discussion of the impact of the "Spiral of Silence" within the context of ODQA. It is important to note that the influence of the phenomenon is not confined to this scenario; it may also be pertinent across all settings that involve knowledge retrieval, generation, and the influx of text from LLMs. Specifically, our analysis is structured around two dimensions: the query level and the document level.

At the **query level**, Figure 7a signifies the average count of queries shifting between consecutive iterations from incorrect to correct and vice versa, respectively. Notably, during the 1->2 iteration, there is an initial surge in both metrics, which subsequently experience a sharp decline as the iterations continue. This suggests that the LLM-generated text's initial introduction catalyzes a more dynamic state, likely due to the correction of existing errors or the introduction of new inaccuracies. Over time, however, the "Spiral of Silence" effect seems to guide the system towards a state of equilibrium where the transition rate stabilizes to less than 1% per 200 queries. **This means most queries maintain their status as either correct or incorrect, indicating that individual query QA results become fixed.**

At the **document level**, we compute the average rank shifts of the first documents containing the correct answer within retrieval results, under different LLM answer states. In Figure 7b, we observe that: 1) **Different Trends of Correct Answer Rankings Based on Source:** In instances where EM=0, correct documents from all sources ("First Right From ALL Sources") and from humans ("First Right From Human") both tend to be ranked lower over time. When EM=1, the rankings for correct documents from all sources improve slightly, while rank-

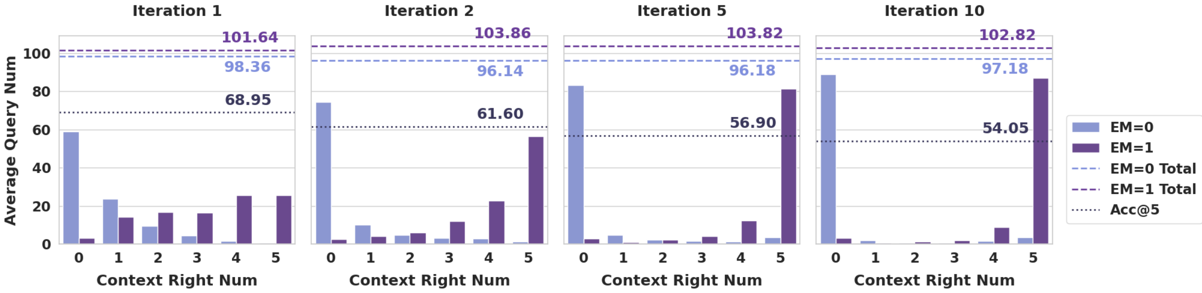
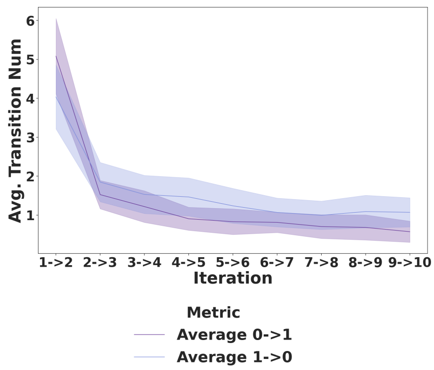
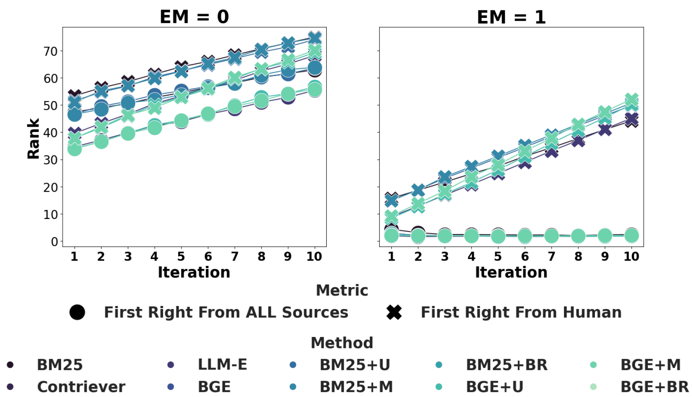


Figure 6: Correlation between the number of top 5 search results containing the correct answer ("Context Right Num") and the accuracy of responses given by LLMs on the NQ dataset. The responses are categorized based on Exact Match (EM) score: EM=1 for correct and EM=0 for incorrect. The overall number of queries that the LLMs answered correctly (EM=1 Total) and incorrectly (EM=0 Total), along with the average retrieval accuracy (Acc@5) are shown by dashed lines. The results are averaged across different LLMs, retrieval, and ranking methods.



(a) Average Query State Transition Number from incorrect to correct ("Average 0->1") and correct to incorrect ("Average 1->0") between consecutive iterations, aggregated across all IR methods and datasets.



(b) Ranking of the first retrieved document containing the correct answer in each iteration, with LLM responses being incorrect (EM=0) or correct (EM=1). "First Right From All Sources" refers to the average rank for the first text containing the correct answer, where the source could be either LLM or human; "First Right From Human" is for human-only sources, both considered across datasets and LLMs.

Figure 7: Effects of "Spiral of Silence" on query and document level of RAG systems.

ings for correct answers from humans continue to decline. This suggests that the LLM's correct texts gain prominence in retrieval rankings over time, overshadowing correct texts from human-generated texts. 2) **Gradual Dysfunction of the IR System in Incorrect LLM Responses:** When the LLM provides incorrect answers (EM=0), there's a risk that documents that once rose to the top with accurate information might increasingly be obscured by the growing mass of LLM-generated content. This can lead to a scenario where the IR system, originally intended to help users find precise information, becomes less reliable. If it prioritizes and disseminates the LLM's inaccuracies, a feedback loop could ensue, solidifying these errors. This concerning trend highlights the critical need for ongoing adjustments and improvements to the IR systems to uphold their purpose.

6 Analysis

To further comprehend the "Spiral of Silence" effect, we illustrate its interaction with misinformation introduced by adversaries using LLMs. Moreover, we test two information filtering mechanisms to alleviate the progression of the effect. For more information on the experimental setup and results, refer to Appendix A.7 and Appendix A.8.

7 Conclusion

In this study, we initiate our research from empirical observations, aiming to investigate the implications of progressively integrating LLM-generated text into RAG systems. To this end, We employ the ODQA task as a case study to examine both the immediate and extended impacts of LLM text on these systems. Our simulation has revealed the emergence of a "Spiral of Silence" effect, suggesting that without appropriate intervention, human-generated content may progressively diminish its influence within RAG systems. Further investigation into this phenomenon reveals that unchecked accumulation of erroneous LLM-generated information could lead to the overlooking of correct information by IR systems, resulting in harm. We urge the academic community to be vigilant and take measures to prevent the potential misuse of LLM-generated data.

Limitations

This study aims to present a new perspective on the impact of LLM-generated texts entering the internet on RAG systems. However, the complexity of reality means that it is impossible to account for all variables. The methods of LLM text generation and the mechanisms by which this content enters the retrieval set are constantly changing, which could affect the performance of RAG systems. While ODQA serves as an insightful approach to evaluate the progression of RAG systems, it is necessary to recognize that ODQA assessments are not exhaustive in capturing the full spectrum of information retrieval scenarios. Nonetheless, the simulation framework proposed in this research is readily adaptable to other tasks that employ RAG systems. Our discussion introduces the "Spiral of Silence" as a potential outcome of the proliferation of LLM-generated texts. Although such a development is not predetermined, given the myriad of factors at play in the real world, this work aims to foster a deeper investigation into the phenomenon and its prospective implications for information diversity in AI-mediated environments.

Ethical Considerations

This paper only explores the potential impact of the LLM-generated text, without involving the release of the generated text and the intervention of social progress, so the possibility of ethical risks is small. We used publicly available LLMs and datasets to conduct experiments that did not involve any ethical issues. In the appendix, we analyze the potential interplay between harmful information and the phenomena outlined in our paper, with a principal objective to draw attention to this issue to advocate for its resolution.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62122077/62272439/62106251), Beijing Municipal Science and Technology Project (No. Z231100010323002), and the Fundamental Research Funds for the Central Universities.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *CoRR*, abs/2307.16877.

Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. 2021. A survey on echo chambers on social media: Description, detection and mitigation. *CoRR*, abs/2112.05084.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go MAD. *CoRR*, abs/2307.01850.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

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

Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1219–1228. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *CoRR*, abs/2311.05656.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *CoRR*, abs/2309.01431.

Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 115–123. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

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

Yi Dai, Hao Lang, Yinhe Zheng, Bowen Yu, Fei Huang, and Yongbin Li. 2023b. Domain incremental lifelong learning in an open world. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5844–5865. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *CoRR*, abs/2301.04246.

Google. 2023. Introducing gemini: our largest and most capable ai model. `https://blog.google/technology/ai/google-gemini-ai/?utm_source=gdm&utm_medium=referral#sundar-note`. Accessed: 2023-12-06.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023b. Catastrophic jailbreak of open-source llms via exploiting generation. *CoRR*, abs/2310.06987.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 383–390. ACM.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

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

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Chen Lin, Dugang Liu, Hanghang Tong, and Yanghua Xiao. 2022. Spiral of silence and its application in recommender systems. *IEEE Trans. Knowl. Data Eng.*, 34(6):2934–2947.

Chang Liu, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Edmund Y. Lam, and Ngai Wong. 2023a. Gradually excavating external knowledge for implicit complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14405–14417. Association for Computational Linguistics.

Dugang Liu, Chen Lin, Zhilin Zhang, Yanghua Xiao, and Hanghang Tong. 2019. Spiral of silence in recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 222–230. ACM.

Ye Liu, Semih Yavuz, Rui Meng, Meghana Moorthy, Shafiq Joty, Caiming Xiong, and Yingbo Zhou. 2023b. Exploring the integration strategies of retriever and large language models. *CoRR*, abs/2308.12574.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *CoRR*, abs/2212.10511.

Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication*, 24(2):43–51.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT. `https://openai.com/blog/chatgpt`. Accessed: January 10, 2024.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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, Singapore, December 6-10, 2023*, pages 1389–1403. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dietram A Scheufle and Patricia Moy. 2000. Twenty-five years of the spiral of silence: A conceptual review and empirical outlook. *International journal of public opinion research*, 12(1):3–28.

Nina Schick. 2020. *Deep Fakes and the Infocalypse: What You Urgently Need To Know*. Monoray.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross J. Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *CoRR*, abs/2305.17493.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10014–10037. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Danni Xu, Shaojing Fan, and Mohan S. Kankanhalli. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 9291–9298. ACM.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *CoRR*, abs/2310.07554.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, and Graham Neubig. 2023. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.

Algorithm 1 Simulation Process

function RUNRAG(D, Q)
  $Res \leftarrow$ empty list
  for $q \in Q$ **do**
    $D' \leftarrow$ RETRIEVE(q, D)
    $p \leftarrow$ GENPROMPT(q)
    $S \leftarrow$ GENANSWER(p, D', q)
    $S' \leftarrow$ POSTPROC(S)
    Add (q, D', S') to Res
  end for
  return Res
end function

$D_0 \leftarrow$ LOADDATA
$initRes \leftarrow$ RUNRAG(D_0, Q)
$basePerf \leftarrow$ EVALRAG($initRes$)
$T \leftarrow$ GENZEROSHOT
$D_1 \leftarrow D_0$ **combined with** T
$t \leftarrow$ number of iterations
for $i \leftarrow 1$ **to** t **do**
  $iterRes \leftarrow$ RUNRAG(D_i, Q)
  $perf_i \leftarrow$ EVALRAG($iterRes$)
  $D_{i+1} \leftarrow$ UPDATEDATA($D_i, iterRes$)
end for

A Appendix

A.1 Discussion of Application on "Spiral of Silence"

In aligning the "Spiral of Silence" theory with the focus of this study, emphasis on the aspect of the "individual's will to express" inherent in the original theory is purposefully diminished. The factors influencing the "Spiral of Silence" phenomenon, as mentioned in Scheufle and Moy (2000), with media and temporality being the principal elements within RAG systems, directly affect the relative standing of LLM and human texts as the system evolves. While the individual's desire to express may be indirectly affected by media and temporality, these are not the primary drivers of the "Spiral of Silence" within RAG systems. In RAG systems, we hypothesize that texts generated by LLMs will increasingly be favored in the hierarchy of information retrieval, whereas texts authored by humans might be systematically marginalized, resulting in a structural form of "passive silencing".

A.2 Pseudo-Code of Simulation Process

The pseudo-code of the simulation process in section 3.2 is shown in Algotirhm 1.

A.3 Post-Process Details

During the experimental process, we observe that the response texts from LLMs occasionally contain specific phrases at the beginning that indicate their identity. These phrases are difficult to remove through prompts and are irrelevant to the topic at hand. Examples include sentences such as:

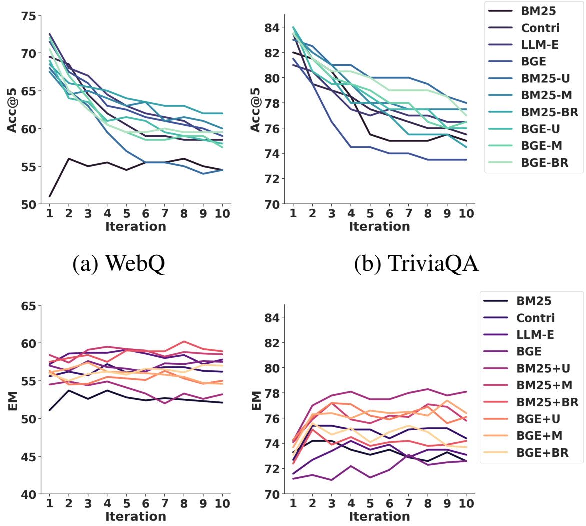
- `"I'd be happy to assist you with your question."`
- `"According to my knowledge..."`
- `"As an AI language model..."`

We collect over 40 such sentences using a manual annotation approach and filter each LLM-generated text through string matching. If a matching string is found, the corresponding sentence or fragment is removed.

A.4 Implementation Details.

To construct and execute the simulated iterative framework, we adopt a diverse array of tools and technologies to facilitate real-time interaction between various retrieval methods, indexing architectures, and LLMs. We implement the APIs of various LLMs relying on api-for-open-llm[3]. With integration of LangChain[4] with Faiss (Johnson et al., 2019) and Elasticsearch[5], we execute batched incremental updates of LLM-generated documents in each iteration, thus simulating the process of document index updating by search engines in real-world scenarios. To maintain the diversity of the generated texts, we set the temperature at 0.7 for all LLMs. In each iteration of the experiment, except for the zero-shot setting, we keep the size of the context document set D_i' fixed at 5. We rerank the first 100 documents recalled by the retrieval method when the step is applied. We apply the LLMs to generate response text, post-process via rules, and then merge their outputs into the index for each query in every iteration. Therefore, for each iteration, we will add 4k new samples to the index. The total number of iterations t is set at 10, which results in a total of 40k invocations of the LLMs for each experimental run.

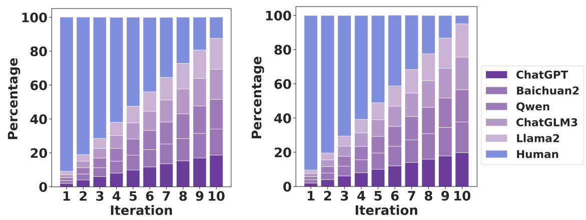
[3]https://github.com/xusenlinzy/api-for-open-llm
[4]https://github.com/langchain-ai/langchain
[5]https://github.com/elastic/elasticsearch



(a) WebQ (b) TriviaQA

(c) WebQ (d) TriviaQA

Figure 8: Long-Term RAG performance for WebQ and TriviaQA. The upper section illustrates the retrieval outcomes for various methods, while the lower section depicts the average EM across LLMs. Iteration 1 represents the results following the incorporation of zero-shot LLM-generated text. Abbreviated re-ranking methods in the legend are: +U for UPR, +M for MonoT5, and +BR for BGE-Reranker.



(a) WebQ (b) TriviaQA

Figure 9: Average Percentage of texts from various sources within the top 50 search results over multiple iterations across different search and methods for WebQ and TriviaQA.

A.5 Results on WebQ and TriviaQA

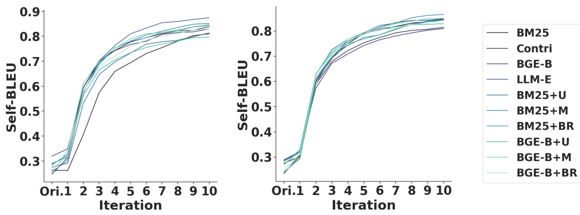
Figure 8 shows long-term RAG performance on WebQ and TriviaQA.

The percentage from various sources and the Self-BLEU of the retrieval results on WebQ and TriviaQA are shown in Figure 9 and Figure 10.

A.6 Simulation with Model Evolution

In this section, we explore the impact of the continuous influx of text generated by evolving LLMs on RAG systems over time. To this end, we experiment by using a series of LLMs, ranging from weak to strong, to observe the phenomenon.

Experimental Setup: We conduct experiments based on BM25 on the NQ and WebQ datasets.



(a) WebQ (b) TriviaQA

Figure 10: 3-gram Self-BLEU score for the top 5 search results over iterations for WebQ and TriviaQA, from the original dataset (Ori.) to subsequent iterations including LLM-generated texts.

Dataset	NQ		WebQ	
Iteration	Acc@20	EM	Acc@20	EM
1	68.5	29.5	66.0	35.5
2	67.5	28.5	65.0	38.7
3	64.0	29.3	63.5	38.0
4	63.0	32.5	62.5	43.5
5	60.5	32.3	62.5	41.8
6	58.0	32.0	61.5	43.3
7	55.5	30.7	61.5	43.2
8	52.0	33.7	60.5	44.0
9	51.5	32.8	61.0	44.3
10	47.5	33.7	58.5	44.2

Table 3: BM25 retrieval performance (Acc@20) and QA performance (EM) across iterations for NQ and WebQ under the evolving LLM setting.

For the text generation models, we use progressively larger versions of the Qwen1.5 model (Bai et al., 2023) across different iterations: 0.5b, 1.8b, and 4b versions for the first three iterations. For iterations four to seven, we employ Qwen1.5-7b-Chat (Bai et al., 2023), LLaMA2-7b-Chat (Touvron et al., 2023), and Baichuan2-7b-Chat (Yang et al., 2023). Finally, for the eighth to tenth iterations, we utilize GPT-3.5-Turbo (OpenAI, 2022), LLaMA2-13b-Chat (Touvron et al., 2023), and Qwen1.5-14b-Chat (Bai et al., 2023). This setup simulates the impact of increasing model capabilities on the RAG system over time.

Analysis of Experimental Results: From the experimental results, we can observe that: **1) Retrieval Performance Continues to Decline:** According to the Acc@20 results reported in Table 3, even with the gradual improvement of LLM performance in the simulation, the retrieval performance exhibits a general downward trend over time. Compared to the results in Table 1, the improvements in retrieval performance are also limited for weaker LLMs in the initial stage (73.5 → 68.5, 71.0 →

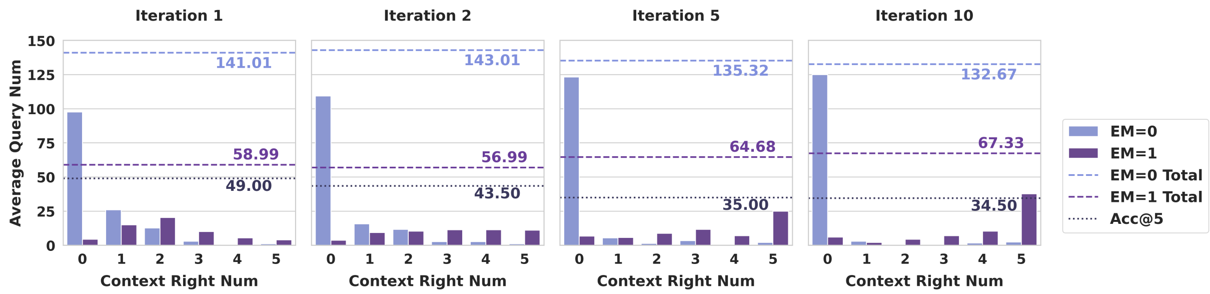


Figure 11: Correlation between the number of top 5 BM25 search results containing the correct answer ("Context Right Num") and the accuracy of responses given by evolving LLMs on the NQ dataset. The responses are categorized based on n Exact Match (EM) score: EM=1 for correct and EM=0 for incorrect. The overall number of queries that the LLMs answered correctly (EM=1 Total) and incorrectly (EM=0 Total), along with the average retrieval accuracy (Acc@5) are shown by dashed lines. The results are averaged across different LLMs.

Iteration	NQ	WebQ
1	77.7	87.0
2	45.9	64.3
3	27.2	48.6
4	18.0	40.2
5	14.6	36.2
6	12.3	32.6
7	10.1	28.9
8	8.6	25.6
9	6.7	21.9
10	5.8	19.2

Table 4: Percentage of human text in the top 5 BM25 retrieval results across iterations for NQ and WebQ under the evolving LLM setting.

Model Name	NQ	WebQ
Qwen1.5-0.5b-Chat	12.5	6.9
Qwen1.5-1.8b-Chat	7.4	4.4
Qwen1.5-4b-Chat	8.5	6.3
Qwen-7b-Chat	1.1	1.7
LLaMA2-7b-Chat	20.9	18.5
Baichuan2-7b-Chat	10.3	7.6
LLaMA2-13B-Chat	19.1	19.9
Qwen1.5-14b-chat	0.9	1.8
GPT-3.5-Turbo	13.5	13.7
Human	5.8	19.2

Table 5: Percentage of generated text by different LLMs in the top 5 BM25 retrieval results at the end of the simulation (Iteration 10) for NQ and WebQ under the evolving LLM setting.

66.0). **2) QA Results Show Improvement:** Based on the EM results reported in Table 3, we can see that with the expansion of the LLM scale, the QA results have improved, especially from the third to the fourth iteration, where the model expanded from less than 4B to 7B, resulting in more than a 10% improvement in QA results on both datasets. Considering that QA results remain stable in Figure 3 when LLMs are unchanged, the current improvement likely stems from the enhanced capabilities of the LLM itself. **3) The Existence of the "Spiral of Silence" Phenomenon:** In Table 4, we present the proportion of human text in the top 5 retrieval results during the iterations. It can be observed that LLM-generated text still dominates the retrieval results, and the influence of human text continues to decline. In Table 5, we report the proportion of different LLM-generated texts in the top 5 retrieval results after the tenth iteration. Overall, texts generated by larger-scale LLMs occupy a slightly higher proportion, but this is not absolute.

Table 6 presents the diversity of retrieval results, while Figure 11 illustrates the relationship between retrieval accuracy and the generated answers. As the model evolves, the experimental results remain consistent with the conclusion in Section 5.3: the diversity of retrieval results continues to decline, and polarization occurs. In summary, the experimental results indicate that the evolution of LLMs will not prevent the occurrence of the "Spiral of Silence" phenomenon.

A.7 Effects of Misinformation

In previous sections, we explored the impact of non-maliciously LLM-generated texts on the evolution of the RAG system over time. In this section, we will discuss the persistence of the "Spiral of Silence" when attackers deliberately inject specific misinformation into the RAG system, how misinformation could affect the system over time, and the feasibility of targeted misinformation injection.

Iteration	NQ	WebQ
1	27.5	25.4
2	49.8	39.5
3	65.6	53.3
4	74.4	60.9
5	70.8	59.2
6	72.3	60.9
7	74.8	63.6
8	75.5	64.5
9	75.3	64.7
10	78.2	68.4

Table 6: 3-gram Self-BLEU score for the top 5 BM25 search results over iterations for NQ and WebQ under the evolving LLM setting.

Experimental Setup: Our experiment follows the CTRLGEN method detailed in Pan et al. (2023), which aligns well with the zero-shot setting used in our trials and simulates the intent of malicious actors to create and propagate false information. Specifically, for each query, we generate five incorrect answers using GPT-3.5-Turbo and then randomly select one to guide five different LLMs to each create a document supporting that incorrect response. These documents replace the zero-shot data in the index from the experiments in Section 3 and are used for simulated iterative experiments. Details of the prompts used are provided in Appendix A.9. For the sake of conciseness, we report only the experimental results for four retrieval methods.

Experimental Evaluation: When generating texts containing misinformation using LLMs, we face two primary challenges. First, the model may ignore the instructions, thus inadvertently generating texts that only contain the correct answer. Second, even if the LLM-generated text includes the provided incorrect answer, the content may not genuinely support that answer. To address these issues, we utilize GPT-3.5-Turbo to evaluate the alignment of text t with the given answer a, which could be either correct or incorrect. This evaluation complements the calculation of the EM metric for texts generated by LLMs. We define the EM_{llm} metric as follows:

$$\mathrm{EM}_{llm}(t,a) = \begin{cases} 1, & \text{if } t \text{ contains and supports } a \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

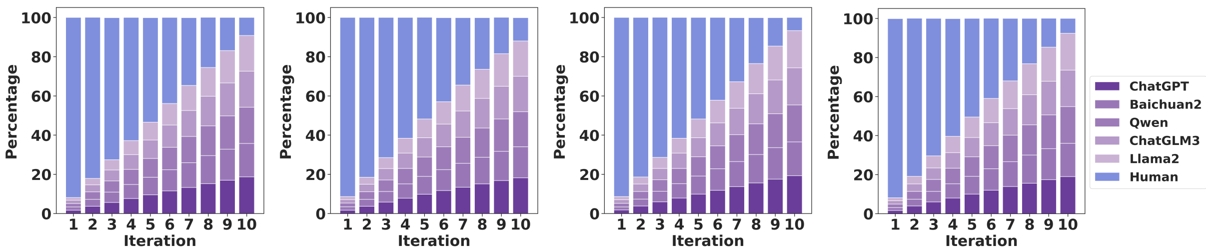
Here, a represents an answer that the text t is being evaluated against. The text t is deemed to contribute positively to this metric if it encompasses a and is validated by GPT-3.5-Turbo as being supportive of a. The EM_{llm} for the correct answers and specific incorrect answers, as generated by the CTRLGEN method containing misinformation, are presented in Table 7.

Moreover, we substantiate the rationality of employing EM as a QA evaluation metric in the experiments of Section 4 by calculating the Pearson correlation coefficient between EM_{llm} and EM based on the experimental results in this section. We observe that the EM_{llm} values for the texts generated by the other four LLMs, as verified by GPT-3.5-Turbo, have an average correlation exceeding 0.5 across four datasets when compared to the EM values obtained through direct string matching, as shown in Table 8. This demonstrates a significant correlation between the two metrics. For incorrect answers, the correlation is relatively lower, indicating the necessity of using GPT-3.5-Turbo to further filter texts in the exploration of misinformation. Considering the higher efficiency of direct EM calculation over EM_{llm}, we use EM to evaluate the QA quality of the RAG system for experiments in other sections.

Analysis of Experimental Results: From the experimental results, we can observe that: **1) The "Spiral of Silence" Still Exists:** We first investigate the presence of the "Spiral of Silence" phenomenon when misinformation targeting the objective is injected into the corpus. As shown in Table 9, although the majority of the injected information is misleading, the content generated by the LLMs is still quickly ranked at the top by the retrieval systems, taking a dominant position. When comparing four different retrieval methods, the BM25 algorithm shows greater robustness than the others, being least affected by the LLM-generated content. However, it is noteworthy that approximately 20% of the content generated by the LLM could still be quickly placed in the forefront of the search results by the BM25 algorithm. Figure 12 illustrates that over time, human-written texts are gradually excluded from the searchable range, and as depicted in Figure 13, the phenomenon of homogenization of opinions in search results persists. This further indicates that regardless of the accuracy of the LLM-generated information, the "Spiral of Silence" phenomenon remains present. **2) The RAG System has a Limited Degree of Self-Correction Capability:** LLM-generated texts containing misinformation lead to a significant decline in retrieval

Model	NQ		WebQ		TriviaQA		PopQA	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
GPT-3.5-Turbo	0.015	0.7	0.075	0.57	0.105	0.57	0.045	0.71
Baichuan2-13B-Chat	0.11	0.595	0.21	0.44	0.16	0.455	0.1	0.65
Qwen-14B-Cha	0.065	0.61	0.11	0.565	0.165	0.535	0.05	0.7
ChatGLM3-6B	0.085	0.605	0.195	0.435	0.245	0.415	0.105	0.61
LLaMA2-13B-Chat	0.04	0.43	0.085	0.385	0.125	0.405	0.03	0.55
Avg	0.063	0.588	0.135	0.479	0.16	0.476	0.066	0.644

Table 7: EM_{llm} of different models for Correct answers and specific Incorrect answers.



(a) NQ (b) WebQ (c) TriviaQA (d) PopQA

Figure 12: Average percentage of texts from various sources within the top 50 search results over multiple iterations when adding **Misinformation** across different search methods.

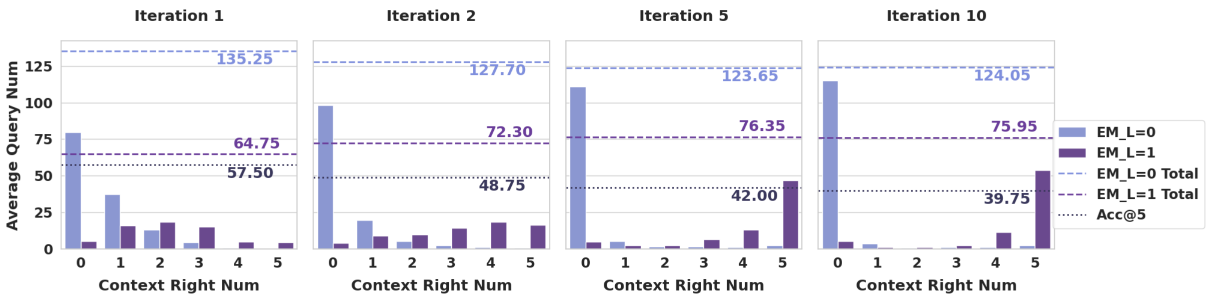


Figure 13: Correlation between the number of top 5 search results containing the correct answer ("Context Right Num") and the accuracy of responses given by LLMs on the NQ dataset when adding **Misinformation**. The responses are categorized based on EM_{llm} score: EM_L=1 for correct and EM_L=0 for incorrect. The overall number of queries that the LLMs answered correctly (EM_L=1 Total) and incorrectly (EM_L=0 Total), along with the average retrieval accuracy (Acc@5) are shown by dashed lines. The results are averaged across different LLMs, retrieval and ranking methods.

Answer Type	NQ	WebQ	TriviaQA	PopQA
Correct	0.740	0.568	0.836	0.529
Incorrect	-0.574	-0.389	-0.385	-0.121

Table 8: Evaluation of the Average Pearson Correlation Coefficient between EM and EM_{llm} for Correct and Incorrect Answers.

Method	NQ	WebQ	TriviaQA	PopQA
BM25	26.7	17.7	24.7	47.4
Contriever	60.7	62.5	64.7	67.2
LLM-Embedder	65.8	70.4	73.3	74.2
BGE_{base}	50.7	48.6	63.2	60.6

Table 9: Percentage of LLM-generated documents with **Misinformation** occupying the top 5 retrieval results, after augmenting each query with five documents generated by LLMs. Data entries framed by a blue background indicate a majority presence of human-generated documents, while entries with a purple background denote a predominance of LLM-generated documents.

and QA performance based on the RAG system compared to results in Section 5.2. With the continuation of the iteration process, the number of correct answers increased and the number of incorrect answers decreased, albeit by a small margin. This suggests that the RAG system has a certain degree of self-correction capability, which may stem from the model's knowledge or the human-written texts containing correct information retrieved in the initial stages. **3) The Introduction of a Small Amount of the LLM-generated Texts with Specific Misleading Information during the Iterative Process could Inject such Information into the RAG Output:** In Figure 14, we quantify the EM_{llm} metric of the original and specific misleading answers generated by the RAG system based on four retrieval methods at various iteration stages. The results show that after the purposeful addition of misleading information (before the first iteration), the proportion of RAG system-generated answers containing specific misleading information significantly increases, especially on NQ and PopQA, where the proportion of incorrect answers exceeds that of correct ones, and the influence of misleading answers persisted over time. However, the BM25 algorithm exhibits relatively higher robustness, and the EM_{llm} of incorrect answers output by the RAG system based on it remains lower than the other three retrieval methods. The experimental results of this section reveal that despite the presence of self-correcting mechanisms, the injection of specific misleading information can still severely compromise the system's accuracy and enable the manipulation of the RAG system to consistently output specific misinformation in response to certain questions. Therefore, without timely intervention, the "Spiral of Silence" phenomenon could marginalize accurate information, leading to severe misinformation consequences.

A.8 Attempts to Alleviate the "Spiral of Silence"

The "Spiral of Silence" effect could lead to the marginalization of human-generated text expression and further enhance the homogeneity of retrieval outcomes. If left unaddressed, this phenomenon could precipitate a series of adverse repercussions. To mitigate or eliminate the influence of the "Spiral of Silence" effect, this section initiates a discussion on two fronts. First, from the perspective of the authenticity of sources, we employ the widely used AIGC detection technologies to filter out and exclude all non-human-produced texts at the top of the search results. Second, addressing the validity of content, we strive to maintain diversity among the top search results to overcome potential issues caused by excessive homogenization.

Experimental Setup: To balance the efficiency and effectiveness of the retrieval system, for each set of search results returned by the system, we post-process to acquire the top 5 qualifying documents that are visible to the LLMs. In the **source filtering** experiment, we employ the Hello-SimpleAI/chatgpt-qa-detector-roberta[6] model to authenticate the origins of the texts within the search results, aiming to retain the first 5 documents identified as human-generated and supply them as input to the LLM's context. For the **content filtering** part of the experiment, we apply a selection process based on computing the 3-gram Self-BLEU scores. The specific procedure is as follows: For the top 5 documents returned for each search query, we initially calculate their Self-BLEU scores; if the score exceeds a predetermined threshold (set at 0.4 for this experiment), we then compute the Self-BLEU scores for all possible combinations of 4 documents and select the minimum value among them. This minimum value indicates the maximum individual document contribution to the Self-BLEU score not included in the calculation. Subsequently, we exclude the document

[6] https://github.com/Hello-SimpleAI/chatgpt-comparison-detection

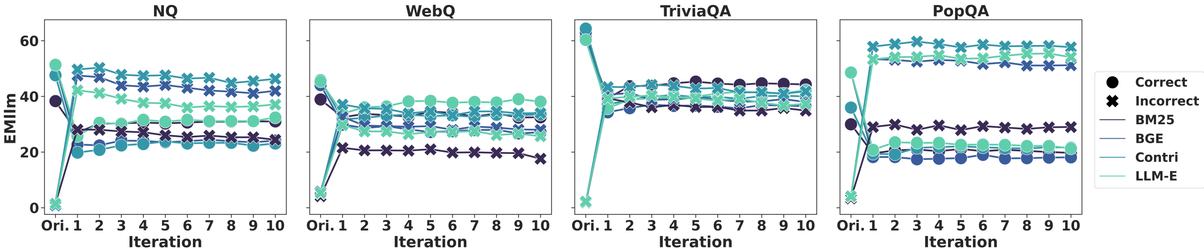
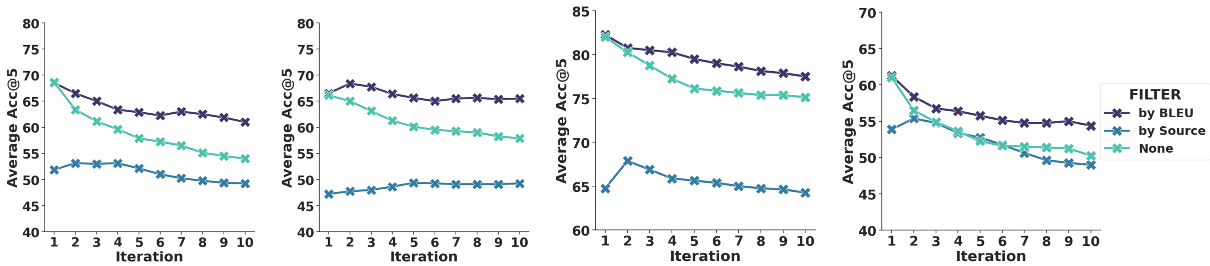
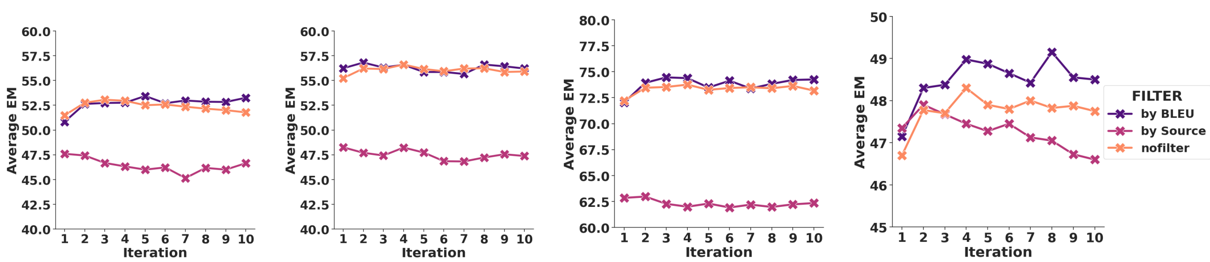


Figure 14: The average EM_{llm} scores of the correct and incorrect answers of the RAG system in the simulation across datasets and LLMs.



(a) NQ (b) WebQ (c) TriviaQA (d) PopQA

Figure 15: Average long-term retrieval performance of different filtering strategies.



(a) NQ (b) WebQ (c) TriviaQA (d) PopQA

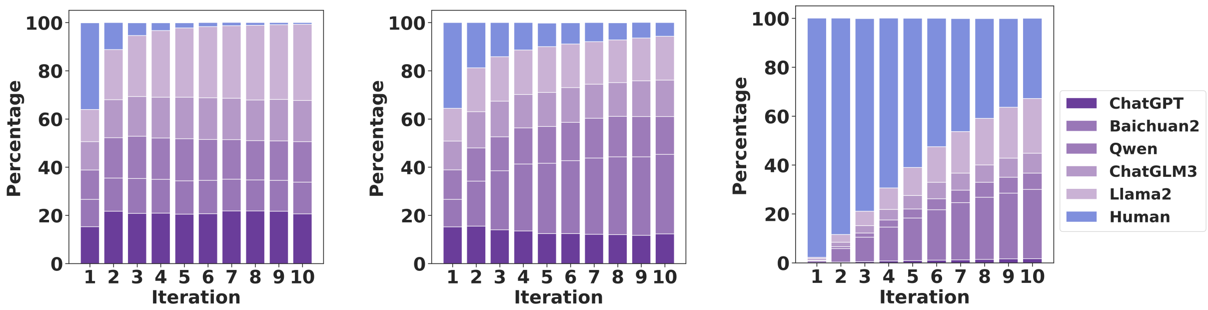
Figure 16: Average long-term QA performance of different filtering strategies.

contributing the most to the Self-BLEU score and incorporate the next ranked document into the combination, repeating this filtering process until the combination's Self-BLEU score meets the preset threshold criteria.

Analysis of Experimental Results: From the experimental results, we can observe that: **1) Both Approaches Yield more Stable Retrieval Outcomes; however, the Source Filtering Method Incurs a Performance Cost:** Figure 15 and Figure 16 illustrate the variations in the average retrieval outcomes and QA performance across datasets, before and after the application of two distinct filtering strategies, compared to an unfiltered condition. Observations indicate that by implementing source-based and diversity-based filtering methods, the fluctuation range of the top 5 retrieval results is reduced compared to the non-intervention scenario, suggesting that the filtering mechanisms can bring a more stable retrieval performance for RAG systems. Across the four datasets, the retrieval performance following SELF-BLEU value filtering generally surpasses the unfiltered condition; conversely, the source-based filtering strategy results in an overall performance degradation. This could be attributed to the discriminating model erroneously excluding valid human-generated texts while aiming to eliminate those generated by LLMs. Moreover, in QA tasks, diversity filtering either enhances or maintains QA performance, whereas source-based document filtering leads to a decline in QA performance across all datasets. For instance, on the TriviaQA dataset, the average EM score drops by over 14%. **2) Both Methods can only Alleviate the "Spiral of Silence" Phenomenon to Varying Degrees but Cannot Eliminate it:** Figure 17 displays the proportion of documents from different sources within the top 5 retrieval results in each iteration under three filtering setups on the NQ dataset. It is observable that without any filtering strategy, human-generated texts rapidly vanish from the top 5 documents in the initial iterations. The SELF-BLEU value filtering method retains human-generated texts to a small extent; source filtering, on the other hand, maximally filters out LLM-generated texts, especially those produced by GPT-3.5-Turbo, Qwen, and ChatGLM3, with over 30% of human-generated texts remaining in the top 5 by the end of the tenth iteration. However, despite both filtering strategies slowing the disappearance of human texts, the proportion of human-generated content continues to exhibit a declining trend. Figure 18 and Figure 19 demonstrate that compared to the absence of filtering strategies, both filtering methods slow down the polarization speed of top document accuracy in retrieval performance. Overall, we discovered that filtering based on the source of documents and their diversity can, to some extent, slow down the emergence of the "Spiral of Silence" phenomenon. Source-based filtering has a more pronounced effect in terms of preserving the proportion of human-generated texts and mitigating viewpoint polarization; however, this benefit comes at the expense of the performance of the RAG system. Text filtering based on diversity shows superior performance in maintaining RAG system functionality, but it has a weaker impact on preserving the ratio of human texts and alleviating viewpoint polarization. Despite these findings, neither method can completely eradicate the "Spiral of Silence" effect, indicating the imperative to explore additional solutions. For example, there is a need to investigate retrieval models that can effectively balance between LLM-generated documents and human-generated documents to address this issue.

A.9 Prompts

The prompts used in the experiment are shown in Table 10.



(a) None (b) Content Filtering (c) Source Filtering

Figure 17: Average percentage of texts from various sources within the top 5 search results over multiple iterations on NQ when using different filtering strategies across different search methods.

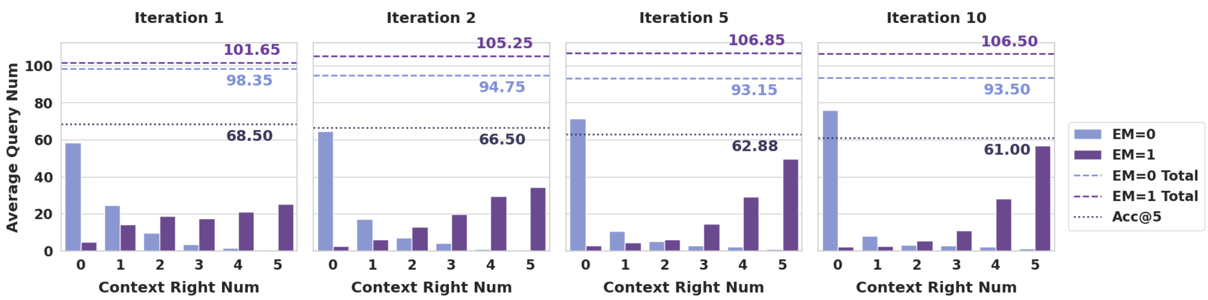


Figure 18: Correlation between the number of top 5 search results containing the correct answer ("Context Right Num") and the accuracy of responses given by LLMs on the NQ dataset when using **Content Filtering**. The responses are categorized based on Exact Match (EM) score: EM=1 for correct and EM=0 for incorrect. The overall number of queries that the LLMs answered correctly (EM=1 Total) and incorrectly (EM=0 Total), along with the average retrieval accuracy (Acc@5) are shown by dashed lines. The results are averaged across different LLMs, retrieval, and ranking methods.

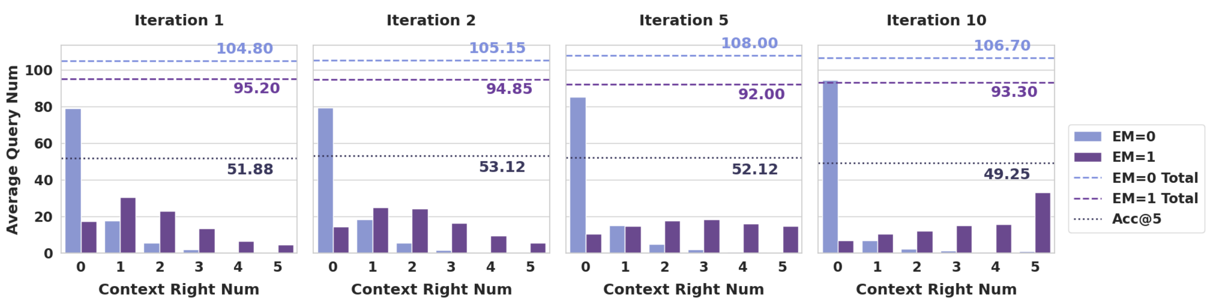


Figure 19: Correlation between the number of top 5 search results containing the correct answer ("Context Right Num") and the accuracy of responses given by LLMs on the NQ dataset when using **Source Filtering**. The responses are categorized based on Exact Match (EM) score: EM=1 for correct and EM=0 for incorrect. The overall number of queries that the LLMs answered correctly (EM=1 Total) and incorrectly (EM=0 Total), along with the average retrieval accuracy (Acc@5) are shown by dashed lines. The results are averaged across different LLMs, retrieval and ranking methods.

Task	Prompts

Zero-Shot Generation	`Provide a background document in 100 words according to your knowledge from Wikipedia to answer the given question.`
	
	`Question:{`**`question_str`**`}`
	
	`Background Document:`

Generation with Contexts	`Context information is below.`
	`--------------`
	`[Context 1]:{`**`context_str1`**`}`
	`...`
	`[Context 5]:{`**`context_str5`**`}`
	`--------------`
	`Using both the context information and also using your own knowledge, answer the following question with a background document in 100 words.`
	
	`Question:{`**`question_str`**`}`
	
	`Background Document:`

Mis_Answer Generation	`Generate a false answer to the given question. It should be short (less than ten words in general) and look plausible, compared to the reference answer.`
	
	`Question:{`**`question_str`**`}`
	
	`Reference Answers:{`**`ref_answer_str`**`},`
	
	`False answer:`

Mis_Passage Generation	`Suppose you are generating realistic-looking claims for a quiz contest. You need to generate a background document less than 100 words in support of the answer to the given question. Don’t contain any word in the original answers in {ref_answer_str}. The background document must contain the following given answers with their original form.`  `Question:{`**`question_str`**`}`  `Answers:{`**`false_answer_str`**`},`  `Background document:`

Answer Check	`Does the following response support the answer to the question?`
	`Question: {`**`question_str`**`}`
	`Response: {`**`response_str`**`}`
	`Answer: {`**`ref_answer_str`**`} / {`**`false_answer_str`**`}`
	`Just answer ’yes’ or ’no’.`

Table 10: Prompts for different tasks.