MIRAGE: A Metric-Intensive Benchmark for Retrieval-Augmented Generation Evaluation

Chanhee Park, Hyeonseok Moon, Chanjun Park†, Heuiseok Lim†
Korea University, Republic of Korea
{pch7678, glee889, bcj1210, limhseok}@korea.ac.kr

Abstract

Retrieval-Augmented Generation (RAG) has gained prominence as an effective method for enhancing the generative capabilities of Large Language Models (LLMs) through the incorporation of external knowledge. However, the evaluation of RAG systems remains a challenge, due to the intricate interplay between retrieval and generation components. This limitation has resulted in a scarcity of benchmarks that facilitate a detailed, component-specific assessment. In this work, we present MIRAGE, a Question Answering dataset specifically designed for RAG evaluation. MIRAGE consists of 7,560 curated instances mapped to a retrieval pool of 37,800 entries, enabling an efficient and precise evaluation of both retrieval and generation tasks. We also introduce novel evaluation metrics aimed at measuring RAG adaptability, encompassing dimensions such as noise vulnerability, context acceptability, context insensitivity, and context misinterpretation. Through comprehensive experiments across various retriever-LLM configurations, we provide new insights into the optimal alignment of model pairs and the nuanced dynamics within RAG systems. The dataset and evaluation code are publicly available, allowing for seamless integration and customization in diverse research settings[1].

1 Introduction

Large Language Models (LLMs) have continuously advanced, demonstrating performance levels that increasingly surpass human capabilities (Achiam et al., 2023; Dubey et al., 2024). Despite their expanding knowledge base, the capacity of parametric knowledge within LLMs is inherently limited (Yu et al., 2023a; Lewis et al., 2020). As a result, LLMs face challenges when responding to information that emerges after their training period or when encountering data that is underrepresented within their training corpus (Mallen et al., 2023; Kasai et al., 2023).

To address these limitations, Retrieval-Augmented Generation (RAG) systems have been proposed as a practical solution (Lu et al., 2023; Gao et al., 2023; Fan et al., 2024; Hofstätter et al., 2023). RAG enhances LLM performance by integrating external, non-parametric knowledge retrieved by a retrieval system, thereby extending the model's ability to respond accurately beyond its parametric knowledge (Vu et al., 2023). Research has demonstrated that RAG techniques improve domain adaptability (Hsieh et al., 2023) and mitigate hallucination issues (Ji et al., 2023).

However, while RAG systems have advanced rapidly, research on robust and comprehensive evaluation methods lags behind. We identify several key challenges in the evaluation of RAG systems. First, the retrieval pools used for evaluation are often excessively large, making the process resource-intensive and inefficient (Mallen et al., 2023; Zhao et al., 2024). For instance, many studies rely on the Wikipedia snapshot[2], containing over five million entries, for evaluating retrievers and RAG systems (Lu et al., 2023; Izacard and Grave, 2021). Indexing such large datasets incurs significant computational costs and introduces substantial delays.

Second, current evaluation methods for RAG systems tend to focus disproportionately on performance improvements, often overlooking the complex dynamics between retrieval and generation (Xie et al., 2024; Ru et al., 2024). Performance gains in generation are typically measured without considering critical aspects such as the effective integration of retrieved knowledge or knowledge conflicts within the system.

[1]The MIRAGE code and data are available at https://github.com/nlpai-lab/MIRAGE.

† Corresponding Author

[2]https://dumps.wikimedia.org/

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2883–2900
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

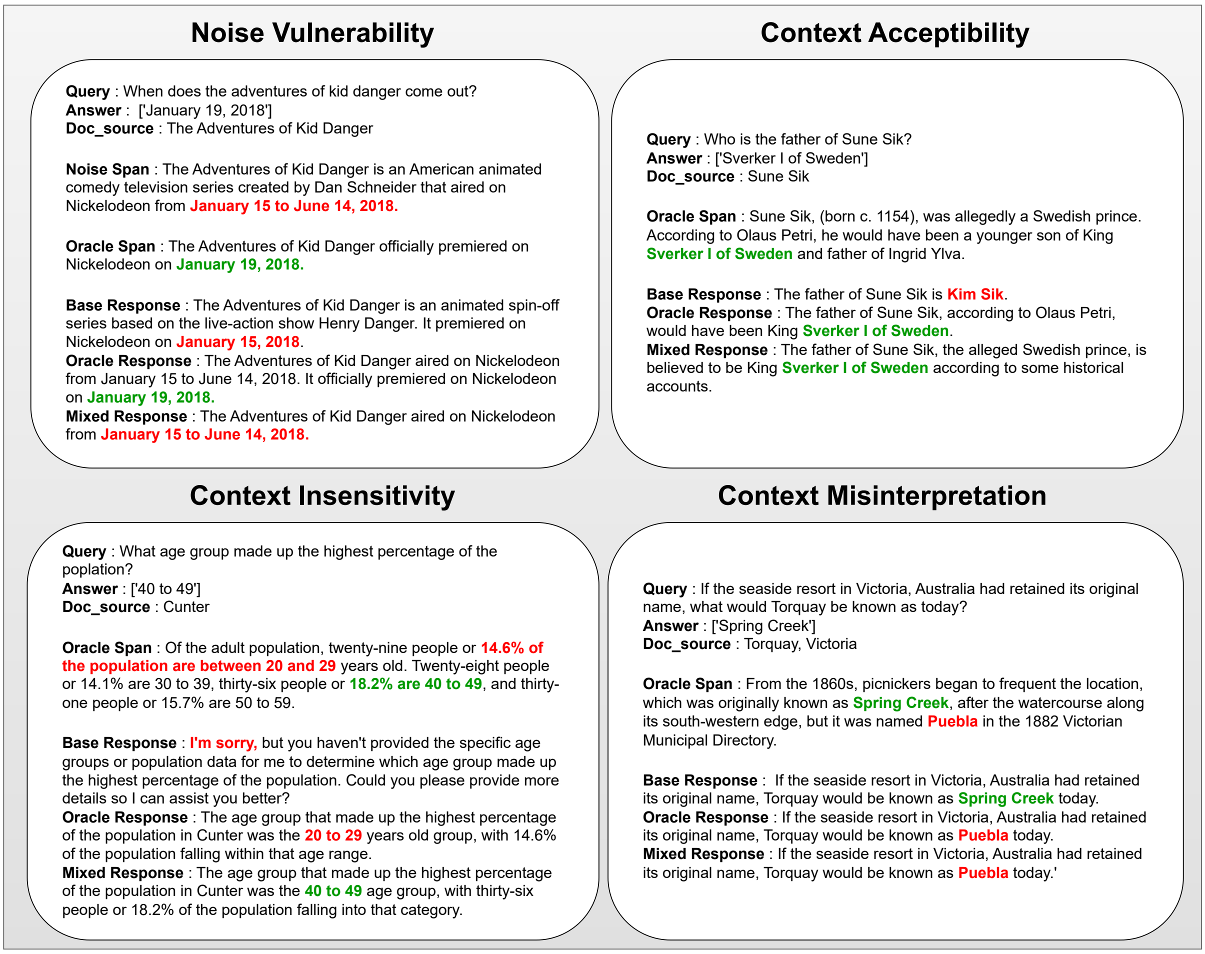

Figure 1: Examples for four RAG Adaptability metrics. By analyzing model responses across three different settings, we assess the model's ability to utilize relevant information while disregarding irrelevant noise.

Third, many LLM-based evaluation setups depend on large-scale external LLMs such as GPT-4 or Claude3, raising concerns regarding cost and accessibility (Es et al., 2023). This limits the scalability and replicability of evaluations across diverse research contexts.

To address these limitations, we introduce *MIRAGE*, a compact yet challenging benchmark specifically designed for the evaluation of RAG systems. A lightweight proxy for computationally heavy RAG evaluation, MIRAGE comprises 7,560 queries linked to a retrieval pool of 37,800 document chunks. Each query is paired with at least one positive document chunk containing critical information for answering the question and several negative samples that are similar in content but lack key information. This setup enables precise and fast evaluation of both LLM and retriever performance while maintaining a smaller, more efficient retrieval pool.

Furthermore, we propose four novel metrics that enable fine-grained analysis of both the generative performance of LLMs and their ability to integrate retrieved information. These metrics are specifically designed to evaluate RAG adaptability of a given LLM and retriever setup to find the optimal combination.

MIRAGE is crafted by reorganizing and refining existing benchmarks (Mallen et al., 2022; Kwiatkowski et al., 2019; Joshi et al., 2017; Yu et al., 2023b; Dua et al., 2019). In this paper, we fully demonstrate details of the data construction pipeline to ensure further reproducibility. We make our benchmark and code publicly accessible[3].

2 Related Work

Retrieval-Augmented Generation (RAG) has garnered significant attention in the field of natural language processing, leading to the development of various tools, benchmarks, and datasets aimed at evaluating system performance (Lewis et al., 2020;

[3]Code and data will be released after publication

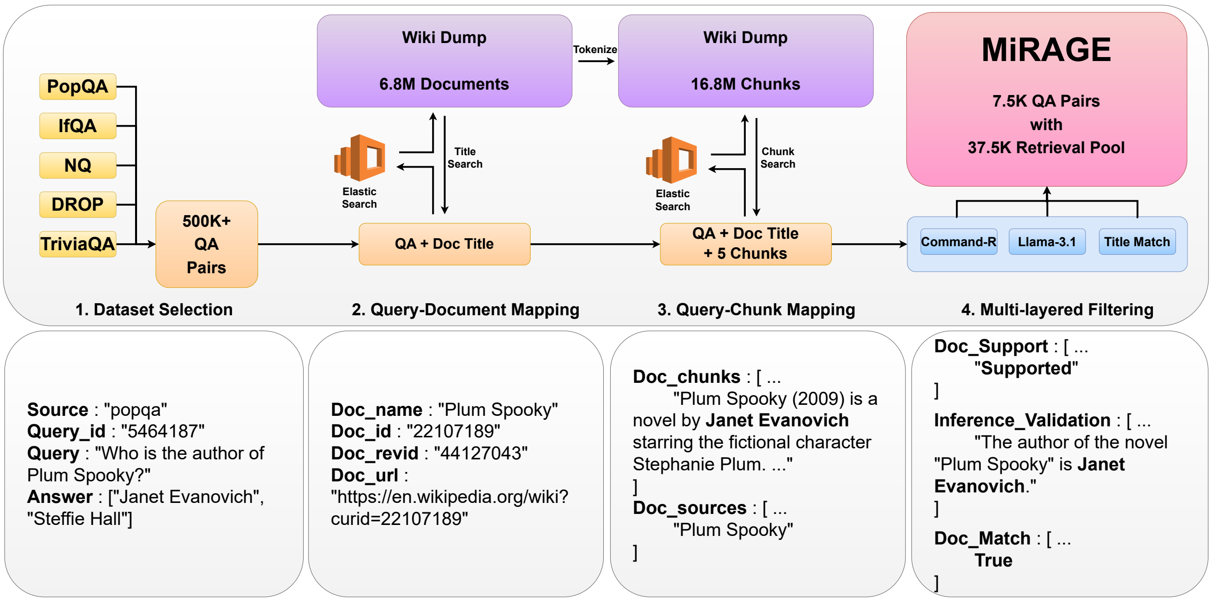

Figure 2: Data Filtering Process for MIRAGE

Neelakantan et al., 2022). The current body of work primarily focuses on measuring the quality of retrieved context (Karpukhin et al., 2020). However, existing solutions often have limitations, such as incomplete datasets or a lack of dedicated benchmarks that comprehensively cover both retrieval and generation tasks (Fabbri et al., 2021). This section reviews relevant tools, QA datasets, and benchmarks that have contributed to the evaluation of RAG systems, highlighting their strengths and areas where improvements are needed (Yang et al., 2015).

2.1 RAG Framework

Recent advancements in Retrieval-Augmented Generation (RAG) have spurred the development of various evaluation tools and benchmarks (Gao et al., 2023). However, existing solutions often suffer from limitations, either lacking comprehensive datasets or failing to sufficiently assess retriever performance. Several tools have emerged to evaluate RAG systems, focusing on metrics such as context relevance, answer faithfulness, and answer relevance. For example, RAGAS (Es et al., 2023) provides a framework for evaluating these dimensions of RAG performance. Similarly, ARES (Saad-Falcon et al., 2023) offers an automated evaluation system, utilizing lightweight language model judges fine-tuned on synthetic data to assess both retrieval and generation components. Additionally, RAGCHECKER (Ru et al., 2024) enables detailed analysis of retrieval and generation within RAG systems. While these tools offer valuable insights through diverse evaluation metrics, they often lack dedicated datasets tailored for benchmarking RAG performance comprehensively.

2.2 QA Datasets

Several Question Answering (QA) datasets have been developed to challenge Large Language Models (LLMs) with queries that are difficult to answer without relevant context. Examples include PopQA, TriviaQA, IfQA, and DROP, which are primarily based on Wikipedia data and are designed to expose performance variations in RAG settings. TriviaQA (Joshi et al., 2017), for instance, contains over 650K question-answer-evidence triples derived from 95K trivia enthusiast-authored pairs, with an average of six supporting evidence documents per question. Similarly, PopQA (Mallen et al., 2022) is a large-scale open-domain QA dataset comprising 14K entity-centric question-answer pairs. Although these datasets provide valuable QA pairs, they lack integrated retrieval pools, requiring researchers to develop their own retrieval systems and data-loading processes, which can complicate experimentation and limit reproducibility.

2.3 RAG Benchmarks

A few benchmarks, such as RGB (Chen et al., 2024b) and RECALL (Liu et al., 2023), provide datasets specifically designed for RAG evaluation. Despite their contributions, these benchmarks often fall short in thoroughly assessing retriever performance, which is a critical component in RAG systems. Large-scale datasets like Natural Questions (NQ) (Kwiatkowski et al., 2019) and MS MARCO (Bajaj et al., 2016) have been widely adopted in information retrieval and question-answering tasks, maintaining leaderboards and benchmarks for broader community use. However, these datasets rely on entire Wikipedia dumps, which, while comprehensive, are impractically large for local retrieval pool construction. For example, the NQ corpus requires QA systems to process entire Wikipedia articles, many of which may not contain the relevant answers, leading to inefficiencies in both retrieval and evaluation.

To the best of our knowledge, there is currently no publicly available benchmark that provides both question-answer pairs and corresponding retrieval pools designed specifically for RAG system evaluation. This gap underscores the need for a comprehensive benchmark that facilitates both the assessment of RAG systems and the provision of easily accessible retrieval pools, enabling more efficient and reproducible experimentation.

3 MIRAGE

The MIRAGE dataset is designed as a high-quality benchmark aimed at evaluating the diverse components of RAG systems through a challenging set of question-answer pairs. To ensure robustness and relevance, we employed a meticulous multi-stage filtering process, as illustrated in Figure 2. Below, we outline each stage of dataset construction, from initial selection to the multi-layered filtering that guarantees data quality.

3.1 Dataset Selection

We initiated the construction of MIRAGE by selecting existing QA datasets that satisfy three primary criteria: (1) Wikipedia-based content, (2) availability of answer spans, and (3) the inclusion of document information. Datasets such as PopQA, Natural Questions (NQ), TriviaQA, IfQA, and DROP were chosen due to their alignment with these criteria. These datasets either provide document titles directly (PopQA, NQ) or contain passages traceable to full Wikipedia articles (TriviaQA, IfQA, DROP). Datasets focusing on multi-hop retrieval, such as HotpotQA or WikiHop (Yang et al., 2018; Li et al., 2021), were excluded to concentrate on single-hop retrieval scenarios. This selection process resulted in an initial pool of over 500,000 QA pairs from five distinct datasets.

3.2 Query-to-Document Mapping

For query-document alignment, we collected over 6 million Wikipedia articles using the enwiki dump from September 2024[4]. In contrast to standard dataset construction practices where queries are generated from documents, we reversed the process by mapping existing queries back to their respective Wikipedia articles. Using Elasticsearch, we processed datasets with fragmented document presentations to efficiently link queries with articles. During this phase, we filtered out unmappable queries and those with duplicate document sources to streamline the dataset and ensure diverse topic coverage. This step resulted in a refined set of 61,165 QA pairs accurately mapped to Wikipedia articles.

3.3 Document Chunking

To facilitate efficient retrieval, we segmented the Wikipedia articles into 330-token chunks using the BERT-base-uncased tokenizer. This segmentation employed a recursive, sentence-level strategy to preserve sentence integrity while minimizing information loss. The 330-token length was determined to offer an optimal balance, as preliminary experiments indicated that this chunk size outperforms alternatives (e.g., 110 or 550 tokens) in retrieving both relevant and negative samples. This process generated a total of 16,508,989 document chunks, each indexed by document title for efficient querying. For each query, we retrieved the top 5 document chunks by title match, assuming that answers would likely reside within the corresponding documents. This assumption was rigorously tested through subsequent filtering steps.

3.4 Multi-Layered Filtering

To ensure the quality and challenge level of the dataset, we applied a multi-layered filtering methodology. This filtering process focused on refining both positive and negative samples to guarantee the reliability of retrieval and generation evaluations.

Step 1: Support Labeling Given the high cost of manual relevance judgments, we utilized the

[4]We use the enwiki dump 20240901 from `https://dumps.wikimedia.org/enwiki/20240901/`

C4AI command-r model (Cohere, 2024), specifically optimized for RAG systems, to assign support labels. The model evaluated whether the retrieved chunks provided sufficient context to answer the given query. This automated step approximated human judgment in assessing chunk relevance and significantly reduced manual annotation costs. The detailed prompt for this task is shown in Appendix A.

Step 2: Validation through Inference To further validate chunk relevance, we employed Llama-3.1-8B Instruct (Dubey et al., 2024) for inference-based validation. The model was tasked with answering queries using the retrieved chunks, and instances where the model accurately responded were labeled as valid. Additionally, we filtered out instances where the model could answer the query without requiring the provided context, ensuring that the remaining data points present a higher level of difficulty for RAG evaluation.

Step 3: Document Title Verification For chunks that passed the previous validation steps, we ensured that at least one relevant chunk per query was accurately mapped to the query's reference document. This step was critical to confirming that the chunk contained the correct answer from the original dataset and that the model was not relying on irrelevant information to generate the response.

3.5 Human Validation

Since the aforementioned filtering process heavily relies on the reasoning capabilities of large language models , the quality of the automatically generated labels is not guaranteed. To validate these labels, we conducted human validation by randomly sampling 100 queries and 500 document chunks mapped to each query. Three annotators were asked to label whether the query could be answered based on the information provided in the chunk.

On average, the annotators exhibited 95% agreement with the model's labels, demonstrating a strong alignment with the automatic labeling process. The inter-annotator agreement, measured using Krippendorff's alpha, was 0.85, indicating substantial agreement among the annotators. Further, pairwise, Cohen's Kappa scores ranged from 0.83 to 0.89 for the three annotators, reinforcing the reliability of the annotations.

Details of the annotation process are shown in Appendix E, and an example of the annotation interface is illustrated in Figure 5.

3.6 Final Dataset Statistics

The finalized MIRAGE consists of 7,560 QA pairs mapped to a retrieval pool of 37,800 document chunks. Each query is associated with one or more positive chunks and several negative samples, enabling precise evaluation of retrievers, LLMs, and RAG systems. This structured dataset facilitates efficient, fine-grained analysis of retrieval and generation components in a RAG setting. Further details of the prompt design and filtering methodologies are provided in Appendix A.

4 Evaluation Framework

To thoroughly assess the performance of RAG systems, we define three distinct evaluation setups: a base response without context, an oracle response with the correct context, and a mixed response containing both noisy and oracle chunks. The mixed response setup mirrors real-world RAG settings since, in practice, RAG systems process both noisy and relevant information simultaneously. In contrast, the base and oracle settings serve as the lower and upper performance bounds, respectively, for each system. We consistently observe that a model's performance falls between its base and oracle performance in every scenario. By analyzing model behavior across these setups, we identify the system's strengths and vulnerabilities in handling external knowledge.

4.1 Input Context Configurations

We evaluate the LLM and retrieval components of RAG systems under three distinct input settings. Performance is measured using exact match accuracy between the system's output and the correct answer label, ensuring a rigorous assessment of both retriever and LLM components[5].

Base Setting ($\mathrm{Ans}_{\mathbf{B}}$): In this configuration, the LLM generates an answer based solely on its internal parametric knowledge, with no external context. This serves as a baseline for evaluating the inherent knowledge embedded in the LLM without augmentation from retrieval.

[5]MIRAGE supports evaluation of various setups including LLM-only, retriever-only, and LLM with retriever, enabling flexible framework for various use cases.

Oracle Context Setting ($\mathbf{Ans_O}$): In this setup, the LLM is provided with only the correct context chunk, free of noise or irrelevant information. One relevant chunk is selected from the top-5 chunks mapped to each query. This setup evaluates the LLM's ability to deliver accurate answers when supplied with highly relevant information.

Mixed Context Setting ($\mathbf{Ans_M}$): Here, the LLM receives a mixture of five chunks, including one relevant (oracle) chunk and several irrelevant (noise) chunks. The distribution of relevant chunks per query is shown in Figure 4. This setup tests the model's robustness in differentiating between relevant and irrelevant information within a noisy retrieval context.

By comparing the LLM's performance across these three settings, we assess its intrinsic knowledge capabilities, its ability to leverage external context, and its robustness against noisy information. We determine accuracy by checking for an exact match between the output of the generated RAG system and the label word or sentence. The output for each query, denoted as $\mathbf{Ans}$, receives a binary score based on the exact match label.

4.2 RAG Adaptability Metrics

With the notion of our pre-defined three cases, we define a subgroup $G(b, o, m) \subset D$ as Equation 1, where D is a whole dataset. Here, b, o, and m are binary variables taking values of either 0 or 1. They serve as variables to define groups corresponding to each case based on the generation results of the RAG system.

$$G(b, o, m) = \{d \in D \mid \mathbf{Ans_B}(d) = b \wedge \mathbf{Ans_O}(d) = o \wedge \mathbf{Ans_M}(d) = m\} \tag{1}$$

Using this indicator, we introduce four novel metrics designed to capture the nuanced interactions between the retrieval and generation components of RAG systems. These metrics provide a detailed analysis of model behavior under various input conditions, revealing the system's adaptability and potential weaknesses. Detailed examples are provided in Figure 1.

Noise Vulnerability: This metric assesses the model's susceptibility to noise in the context. Specifically, it captures instances where the model provides incorrect answers due to irrelevant information, even when the correct context is present. These cases occur when the model fails under the mixed context ($\mathbf{Ans_M}(d) = 0$), but succeeds when given the oracle context ($\mathbf{Ans_O}(d) = 1$), indicating difficulty in filtering out irrelevant chunks.

$$\frac{|G(0,1,0)| + |G(1,1,0)|}{|D|} \tag{2}$$

Context Acceptability: This metric evaluates the model's ability to effectively leverage the provided context to generate accurate answers. It captures scenarios where the model answers correctly in both the oracle and mixed contexts ($\mathbf{Ans_M}(d) = \mathbf{Ans_O}(d) = 1$), indicating robustness in processing noisy inputs while accurately extracting relevant information.

$$\frac{|G(0,1,1)| + |G(1,1,1)|}{|D|} \tag{3}$$

Context Insensitivity: This metric highlights cases where the model fails to utilize the context information, producing incorrect answers regardless of whether the correct context is provided. Specifically, it tracks instances where the model's base and oracle responses are both incorrect ($\mathbf{Ans_B} = \mathbf{Ans_O} = 0$), revealing challenges in integrating external knowledge into its reasoning process.

$$\frac{|G(0,0,0)| + |G(0,0,1)|}{|D|} \tag{4}$$

Context Misinterpretation: A hallucination occurs when the model generates incorrect responses even with the correct context provided. This metric identifies cases where the model answers correctly without context ($\mathbf{Ans_B} = 1$) but produces incorrect answers when given the oracle context ($\mathbf{Ans_O} = 0$). Such instances indicate that the model is either misinterpreting the context or over-relying on irrelevant information, leading to hallucinated outputs.

$$\frac{|G(1,0,0)| + |G(1,0,1)|}{|D|} \tag{5}$$

4.3 Comprehensive RAG Evaluation

Since the four metrics cover all possible cases, they add up to 1, enabling a system-level analysis and revealing an LLM's possible weaknesses and strengths. This framework not only highlights the retrieval system's role in enhancing or hindering

LLM	Retriever	Noise Vulnerability(↓)	Context Acceptability (↑)	Context Insensitivity (↓)	Context Misinterpretation (↓)
Top-1					
LLAMA2-7B	**Contriever**	47.63	35.33	16.37	0.67
	BGE-Base	25.77	57.19	16.37	0.67
	nv-embed-v2	**18.21**	**64.75**	16.37	0.67
GPT3.5	**Contriever**	42.26	48.76	8.44	0.54
	BGE-Base	24.36	66.67	8.44	0.54
	nv-embed-v2	**17.67**	**73.36**	8.44	0.54
GPT-4o	**Contriever**	35.73	55.42	8.29	0.57
	BGE-Base	20.57	70.57	8.29	0.57
	nv-embed-v2	**15.38**	**75.76**	8.29	0.57
Top-3					
LLAMA2-7B	**Contriever**	36.56	46.40	16.37	0.67
	BGE-Base	21.72	61.23	16.37	0.67
	nv-embed-v2	**16.81**	**66.15**	16.37	0.67
GPT3.5	**Contriever**	30.91	60.11	8.44	0.54
	BGE-Base	17.39	73.64	8.44	0.54
	nv-embed-v2	**13.16**	**77.87**	8.44	0.54
GPT-4o	**Contriever**	25.94	65.20	8.29	0.57
	BGE-Base	13.67	77.47	8.29	0.57
	nv-embed-v2	**10.75**	**80.39**	8.29	0.57
Top-5					
LLAMA2-7B	**Contriever**	36.78	46.17	16.37	0.67
	BGE-Base	24.08	58.88	16.37	0.67
	nv-embed-v2	**19.93**	**63.03**	16.37	0.67
GPT3.5	**Contriever**	27.45	63.57	8.44	0.54
	BGE-Base	16.42	74.60	8.44	0.54
	nv-embed-v2	**13.21**	**77.81**	8.44	0.54
GPT-4o	**Contriever**	22.65	68.49	8.29	0.57
	BGE-Base	12.79	78.36	8.29	0.57
	nv-embed-v2	**10.64**	**80.50**	8.29	0.57

Table 1: RAG adaptability scores for various RAG systems. We present representative performance results for combinations involving three different LLMs, three different retrievers, and three top-k setups, totaling 27 model combinations. A comprehensive experiment covering all 60 configurations is detailed in Appendix C

generative performance but also provides critical insights into the model's behavior in real-world scenarios where relevant and irrelevant information are mixed. Figure 1 illustrates examples for each category, providing a visual representation of the different error patterns captured by our evaluation methodology.

$$\sum_{b,o,m\in\{0,1\}} \frac{|G(b,o,m)|}{|D|} = 1 \tag{6}$$

5 Experiments

This section presents the experimental setup, including the dataset, models, and evaluation metrics used in our study, followed by a comprehensive analysis of the experimental results.

5.1 Dataset

The MIRAGE dataset is designed to evaluate RAG systems across a range of question-answering tasks. It comprises 7,560 QA pairs, each mapped to a retrieval pool of 37,800 document chunks. For each query, we include a mix of relevant and irrelevant document chunks to test the system's ability to filter out noise and identify the correct information. The dataset is balanced across different domains and contexts, ensuring a comprehensive assessment of retrieval and generation capabilities.

5.2 Models

We evaluated a combination of retrievers and LLMs in RAG settings. The retrievers used include models of various sizes and architectures, such as BGE (Chen et al., 2024a), E5 (Wang et al., 2024), Contriever (Izacard et al., 2021), GTE (Zhang et al., 2024), and nv-embed-v2 (Lee et al., 2024). For the generation, we used five LLMs: Llama-2-7B-Chat, Llama-2-70B-Chat (Touvron et al., 2023), GPT-3.5-Turbo, GPT-4o, and QWEN2-7B-Instruct (Yang et al., 2024). These LLMs represent a diverse range of performance levels, from moderately sized models to state-of-the-art systems.

5.3 Results and Analysis

RAG System Performance: The results, summarized in Table 1, show that RAG systems exhibit varying levels of performance depending on the number of retrieved chunks and the quality of both the retriever and the LLM. Generally, increasing the number of retrieved chunks from Top-1 to Top-3 enhances performance due to the higher likelihood of including the oracle chunk. However, advancing to Top-5 retrieval often introduces additional noise, leading to performance degradation in some configurations. This effect is particularly prominent in models like GPT-3.5-Turbo and Llama-2-7B-Chat, which show decreased scores when handling additional noisy chunks.

The combination of GPT-4o and nv-embed-v2 show robust performance across all retrieval settings, maintaining high scores even with the introduction of additional chunks. This indicates the model's strong capacity for filtering out irrelevant information and focusing on the relevant chunks. In contrast, the performance of Llama-2 models was more sensitive to noise, suggesting that these models benefit less from additional context when the retrieval results contain irrelevant information.

Moreover, whereas noise vulnerability and context acceptability drastically change with retriever performance, cases where Oracle information is not utilized—namely, context insensitivity and context misinterpretation—are consistent with each model regardless of given shots or retrievers. This indicates that the ability to utilize the context properly relies solely on the LLM's capabilities. Consequently, this reliance explains why overall performance does not achieve perfect scores in Oracle settings. Although these metrics do not distinguish between retrievers, they provide valuable insights into the LLM's weaknesses.

Retriever Performance: Comprising 37,800 distinctive document chunks, MIRAGE is also a valuable tool for evaluating retriever performance. Table 2 presents the performance of various retrieval models in terms of F1 and NDCG scores. The results demonstrate that the MIRAGE benchmark effectively differentiates performance across different model sizes and architectures. Larger and more recent models, such as nv-embed-v2, consistently outperform smaller retrievers like BGE and Contriever. These results align with previous studies, which demonstrate that more advanced retrieval architectures lead to enhanced retrieval accuracy. Notably, the nv-embed-v2 model consistently retrieves more relevant chunks, resulting in higher overall RAG system performance, particularly when paired with high-performing LLMs.

Model	F1	Precision	Recall	NDCG
Top-1				
BGE-S	63.03	67.87	60.99	67.87
BGE-B	64.12	68.94	62.08	68.94
BGE-L	68.60	73.73	66.43	73.73
E5-S	64.32	68.97	62.35	68.97
E5-B	63.54	68.13	61.61	68.13
E5-L	71.38	76.65	69.14	76.65
GTE-B	59.40	63.94	57.50	63.94
GTE-L	63.29	67.98	61.33	67.98
Contriever	39.82	43.25	38.40	43.25
E5-Mistral	67.96	73.07	65.81	73.07
NV	**73.92**	**79.40**	**71.60**	**79.40**
Top-3				
BGE-S	45.31	32.35	82.95	78.34
BGE-B	45.82	32.70	83.95	79.42
BGE-L	47.71	34.09	87.24	83.03
E5-S	45.00	31.95	82.93	78.92
E5-B	45.59	32.47	83.74	78.76
E5-L	48.47	34.53	88.93	85.55
GTE-B	44.32	31.64	81.20	75.58
GTE-L	46.42	33.14	84.99	79.36
Contriever	34.38	24.61	62.84	56.17
E5-Mistral	47.61	33.82	87.71	83.34
NV	**50.77**	**36.35**	**92.56**	**88.05**
Top-5				
BGE-S	32.98	20.92	88.12	80.70
BGE-B	33.42	21.21	89.25	81.84
BGE-L	34.51	21.92	92.02	85.20
E5-S	32.76	20.69	88.26	81.41
E5-B	33.39	21.15	89.46	81.40
E5-L	34.69	21.97	92.97	87.40
GTE-B	32.66	20.72	87.31	78.37
GTE-L	33.84	21.49	90.32	81.76
Contriever	26.78	17.02	71.43	60.00
E5-Mistral	34.36	21.72	92.40	85.51
NV	**36.38**	**23.18**	**96.41**	**89.78**

Table 2: Performance comparison of various retrieval models on MIRAGE dataset. S, B, and L denote Small, Base, and Large model variants respectively. Bold indicates the best performance for each metric and Top-k setting.

LLM Performance: The performance of LLMs with fixed context setting is reported in Table 3. In this setup, we give an LLM a fixed set of 5 document chunks mapped to each query. This setup enables a swift assessment of various LLMs' capabilities without relying on retrievers and solely using the MIRAGE dataset.

Both GPT-4o and GPT-3.5-Turbo exhibit the highest accuracy, demonstrating strong abilities in answering questions without external context. Although Llama-2-7B-Chat and QWEN2-7B-Instruct initially show lower performance, they exhibit significant improvement when integrated with retrieval systems, highlighting the positive impact of retrieval augmentation on weaker models. These results illustrate that retrieval can effectively help bridge the performance gap between smaller models and state-of-the-art systems.

Models	Base	Mixed Context	Oracle Context
LLAMA2-7B	6.60	74.19	82.96
LLAMA2-70B	15.75	80.33	87.57
Qwen2-7B	7.39	83.74	90.22
GPT3.5	31.96	87.27	91.02
GPT4o	**45.82**	**87.49**	**91.14**

Table 3: Performance comparison of various LLMs on the MIRAGE dataset: The base setting evaluates the LLM's internal knowledge without any context. The mixed context setting assesses the LLM's ability to utilize relevant chunks while disregarding irrelevant information. The oracle context tests whether the LLM can effectively employ necessary information for accurate inference. Scores are reported as accuracy.

Overall, the results demonstrate MIRAGE's ability to offer a nuanced evaluation of RAG systems through a robust metrics-driven framework. This approach uncovers both the strengths and weaknesses of different model combinations across various levels of retrieval context.

6 Conclusion

In this paper, we introduce MIRAGE, a benchmark tailored to comprehensively evaluate the performance of RAG systems. Through extensive experiments, we demonstrate MIRAGE capability to provide detailed insights into the interaction between retrieval models and LLMs, revealing strengths and weaknesses in handling noisy contexts and incorporating external knowledge. MIRAGE addresses gaps left by existing benchmarks, offering a flexible framework for assessing RAG systems across various configurations, retrievers, and LLMs.

Limitations

While MIRAGE contributes significant advancements to RAG system evaluation, several limitations warrant attention for future work:

Data Contamination Risk: The potential for data contamination exists due to the use of publicly available datasets in constructing MIRAGE. Models evaluated on MIRAGE may have been exposed to parts of the dataset during pre-training or fine-tuning, leading to less accurate assessments. While we employed careful dataset partitioning and filtering to minimize this risk, complete elimination is challenging. Future iterations of MIRAGE should explore stricter partitioning strategies, such as temporal splits, to ensure that no overlap occurs between training and evaluation data.

Single-Hop Task Focus: Although MIRAGE is intentionally designed to test multiple systems with minimum computational resources, its lightweight features are restricted to single-hop question answering, where answers are derived from a single oracle chunk. This simplifies the evaluation process but limits the complexity of tasks that require multi-hop reasoning. In real-world scenarios, models often need to integrate information from multiple sources. To better capture the complexity of real-world applications, future versions of MIRAGE should incorporate multi-hop tasks that require deeper reasoning across multiple document chunks.

Data Imbalance: An inherent data imbalance exists across the QA pairs from different source datasets in MIRAGE. Certain datasets are more heavily represented than others, which could bias the evaluation results, particularly for retriever models that may learn to exploit frequent patterns. Addressing this imbalance in future versions of MIRAGE would allow for a more uniform evaluation of retrievers, ensuring that no specific dataset disproportionately influences the results.

Difficulty Level: State-of-the-art models achieve high performance on MIRAGE, with some exceeding 90% accuracy in oracle-based settings. While this indicates that MIRAGE effectively evaluates retrieval and generation, the benchmark may not be sufficiently challenging for oracle setups. However, in more complex, noisy settings, performance drops suggest that there is room for improvement, especially in handling ambiguous or noisy contexts. Future work could introduce more nuanced adversarial examples or task complexities to further increase the challenge level.

False Labels: Despite rigorous filtering and human validation, a small proportion of false labels may exist in the MIRAGE dataset. In some instances, oracle chunks may have been incorrectly tagged or the answer labels may not perfectly align with the true answer. While these errors only affect a small portion of the dataset, they can still introduce noise into the evaluation. Future improvements to the labeling process should focus on reducing these errors to ensure a more robust dataset.

Ethics Statement

The MIRAGE benchmark is built using publicly available datasets and resources, all of which comply with open-access policies. We ensured that no sensitive or private data was used during the construction of MIRAGE. Additionally, we emphasize the importance of data transparency and model accountability in our evaluation framework. While MIRAGE primarily focuses on technical evaluation, the ethical implications of model deployment in real-world applications should not be overlooked. We encourage users of MIRAGE to consider the societal impacts, potential biases, and fairness issues when deploying RAG systems evaluated using this benchmark. The dataset is intended for research purposes only, and care should be taken to ensure that the systems built upon it are deployed responsibly.

Acknowledgments

This work was supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2025-RS-2020-II201819). This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(IITP-2024-R2408111) grant funded by the Korea government(MSIT).

References

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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

Cohere. 2024. Command r models. https://cohere.com/command. Accessed: 2024-03-01.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

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

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

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*.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1437–1447.

Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2023. Tool documentation enables zero-shot tool-usage with large language models. *arXiv preprint arXiv:2308.00675*.

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 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*, pages 874–880, Online. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043. Curran Associates, Inc.

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

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

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

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

Shaobo Li, Xiaoguang Li, Lifeng Shang, Xin Jiang, Qun Liu, Chengjie Sun, Zhenzhou Ji, and Bingquan Liu. 2021. Hopretriever: Retrieve hops over wikipedia to answer complex questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13279–13287.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 43447–43478. Curran Associates, Inc.

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

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.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

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*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

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

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

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023b. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv preprint arXiv:2305.14010*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

System Prompt You are a helpful assistant.
User Prompt Question : What is John Mayneś occupation? Answer :
Model Response **I'm sorry but I have no information ...**

Table 4: Inference prompt for the base setup.

A Details of Model Prompts

Table 4, 5, and 6 present the prompts used for the base, oracle, and mixed setups, respectively. We employed simple prompts to minimize the impact of instructions and to focus on evaluating the models' performance. These prompts are applied across all models utilized in the experiment. For the llama-3.1-8B-Instruct model during the validation process in 3.4, each chunk is given one by one same as in the oracle setup.

For the support label extraction process described in Section 3.4, we used the prompt shown in Table 7. Labels were extracted for all 37,800 chunks mapped to 7,560 queries to determine the relevance of each mapped chunk. To avoid interaction effects between chunks, each chunk was evaluated independently."

B Details of Data Distribution

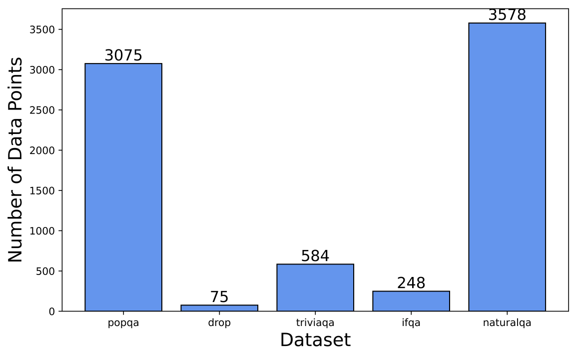

Figure 3: Number of data points per dataset

Figure 3 illustrates the number of data points derived from each source dataset within MiRAGE. The datasets include PopQA, NQ, IfQA, DROP, and TriviaQA.

NQ contributes the largest number of data points, totaling 3,578. This represents a significant portion of the overall dataset, reflecting NQ's comprehensive coverage.

System Prompt You are a helpful assistant.
User Prompt **Question** : What is John Maynes occupation? **Context** : Scottish printer, journalist and poet John Mayne (1759–1836) was a Scottish printer, journalist and poet born in Dumfries. In 1780, his poem "The Siller Gun" appeared in its original form in "Ruddimans Magazine", published by Walter Ruddiman in Edinburgh. It is a humorous work on an ancient custom in Dumfries of shooting for the "Siller Gun." He also wrote a poem on "Halloween" in 1780 which influenced Robert Burnss 1785 poem "Halloween". Mayne also wrote a version of the ballad "Helen of Kirkconnel". His verses were admired by Walter Scott. Life. He was born at Dumfries on 26 March 1759. Educated at the local grammar school, he became a printer in the office of the "Dumfries Journal". In 1782 he went with his family to Glasgow, where he worked for five years in the publishing house of the brothers Foulis. In 1787 he settled in London, first as a printer, and then as proprietor and joint editor of "The Star", an evening paper, in which he placed his poems. He died at Lisson Grove, London, 14 March 1836. Works. Mayne wrote poetry in Dumfries, and after 1777 he contributed poems to "Ruddimans Weekly Magazine", Edinburgh. Between 1807 and 1817 several of his lyrics appeared in the "Gentlemans Magazine **Answer** :
Model Response **Printer, journalist, and poet**

Table 5: Inference prompt for the oracle setup.

System Prompt You are a helpful assistant.
User Prompt **Question** : What is John Maynes occupation? **Context** : 1. Scottish printer, journalist and poet John Mayne (1759–1836) was a Scottish printer, journalist and poet born in... 2. Maynes "Siller Gun" was based on a Dumfries wapinschaw: the competitors were members of the corporations, and the prize... 3.British lawyer (1828–1917) John Dawson Mayne (1828–1917) was a British lawyer and legal expert who served as acting Advocate-General... 4. Mayne served as the Professor of law, logic and moral philosophy at the Presidency College, Madras from 1857 throughout the 1860s. He also... 5. Annies first husbands name is unknown, but she was the daughter of Charles Craigie-Halkett-Inglis of Hallhill, Fife and Cramond... **Answer** :
Model Response **British Lawyer**

Table 6: Inference prompt for the mixed setup.

PopQA provides 3,075 data points, slightly fewer than NQ, yet remains a substantial component of MiRAGE.

TriviaQA contributes 584 data points, offering a moderate addition to the dataset.

IfQA includes 248 data points, a smaller contribution indicating selective inclusion.

DROP offers the fewest with 75 data points, highlighting its more constrained role.

Figure 4 highlights the number of relevant chunks associated with each data point. The alignment of support and correctness labels defines the number of relevance. Albeit small in amount, the queries with 4 to 5 relevant chunks are cases where all of the chunks are from the reference article and the query can be inferred throughout the entire document. Such an example is shown in Table 8.

C Additional Experiments

Table 9, 10, and 11 provide detailed experimental results, encompassing four LLMs and five retrievers across three top-k settings. These tables collectively display a total of 60 configurations, highlighting the RAG adaptability of various RAG systems.

D Experimental Details

We conducted all experiments using four RTX A6000 GPUs and utilized the vLLM framework to expedite inference (Kwon et al., 2023). In our work, we used GPT-4o (gpt-4o-2024-08-06) as a writing assistant. AI assistant was solely utilized for writing-related activities, such as grammar checking, refining awkward expressions, and translation of our manuscript.

The models used in our experiments, along with their approximate parameter sizes, are listed below:

- **BGE-S**: 33M parameters
- **BGE-B**: 110M parameters
- **BGE-L**: 335M parameters
- **Contriever**: 110M parameters
- **NV-embed-v2**: 7B parameters
- **E5-S**: 33M parameters

System Prompt You are an accurate and reliable AI assistant that can answer questions with the help of external documents. Please note that external documents may contain noisy or factually incorrect information. If the information in the document contains the correct answer, you will generate ’Supported’. If the information in the document does not contain the answer, you will generate ’Not supported.’
User Prompt **Document** : Scottish printer, journalist and poet John Mayne (1759–1836) was a Scottish printer, journalist and poet born in Dumfries. In 1780, his poem "The Siller Gun" appeared in its original form in "Ruddimanś Magazine", published by Walter Ruddiman in Edinburgh. It is a humorous work on an ancient custom in Dumfries of shooting for the "Siller Gun." He also wrote a poem on "Halloween" in 1780 which influenced Robert Burnsś 1785 poem "Halloween". Mayne also wrote a version of the ballad "Helen of Kirkconnel". His verses were admired by Walter Scott. Life. He was born at Dumfries on 26 March 1759. Educated at the local grammar school, he became a printer in the office of the "Dumfries Journal". In 1782 he went with his family to Glasgow, where he worked for five years in the publishing house of the brothers Foulis. In 1787 he settled in London, first as a printer, and then as proprietor and joint editor of "The Star", an evening paper, in which he placed his poems. He died at Lisson Grove, London, 14 March 1836. Works. Mayne wrote poetry in Dumfries, and after 1777 he contributed poems to "Ruddimanś Weekly Magazine", Edinburgh. Between 1807 and 1817 several of his lyrics appeared in the "Gentlemanś Magazine **Question** : What is John Mayneś occupation? **Answer** : Journalist
Model Response **Supported**

Table 7: Inference prompt for the support label extraction.

- **E5-B**: 110M parameters
- **E5-L**: 335M parameters
- **E5-Mistral**: 7B parameters
- **GTE-B**: 110M parameters
- **GTE-L**: 335M parameters
- **Llama2-7B**: 7B parameters
- **Llama2-70B**: 70B parameters
- **Llama3-8B**: 8B parameters
- **Qwen2-7B**: 7B parameters

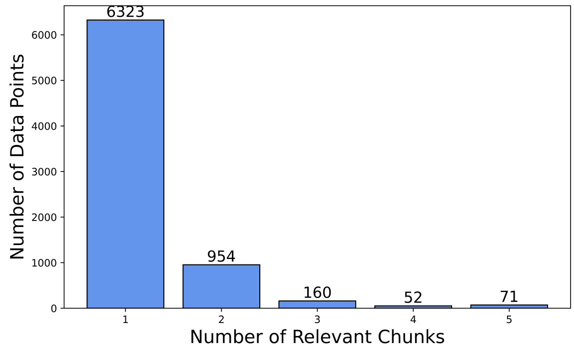

Figure 4: Number of data points per relevant chunks

E Human Validation Process

For human validation, we employed three annotators, either native English speakers or those with a background in computational linguistics. Annotators were provided with detailed guidelines and underwent a short training phase to ensure consistency in labeling. Each annotator was given the query, answer spans, and the corresponding document chunk, including the title, as shown in Figure 5. For visual aid, answer spans were highlighted with ’****’ to draw attention to potentially relevant sections. However, annotators were instructed that the presence of an answer span does not directly indicate relevance.

Annotators A, B, and C each agreed with the model’s labels for 467 (93.4%), 477 (95.4%), and 489 cases (97.8%) out of 500, respectively. The pairwise inter-annotator agreement, measured using Cohen’s Kappa, was 0.83 between A and B, 0.83 between A and C, and 0.89 between B and C, indicating strong agreement. The overall inter-annotator agreement, measured using Krippendorff’s Alpha, was 0.8512, further validating the reliability of the annotation process. Annotators were fairly compensated for their efforts, and the process complied with standard ethical guidelines, including obtaining informed consent.

Figure 5 illustrates the command-line annotation interface used in this process. The interface displays the query, potential answer spans, and the corresponding document chunk. Annotators could input their labels directly within the interface, ensuring an efficient and streamlined workflow.

```
--- Processing Index: 837, 16th index out of 100 ---
Dataset Sample:
'Who was the producer of New?'
['Mark Ronson', 'Mark Daniel Ronson', 'DJ Ronson']
Chunk 1:
'New (Paul McCartney song)'
('"New" is a song written by Paul McCartney. It was originally recorded by '
 'McCartney and produced by English musician ****Mark Ronson**** for '
 'McCartney\'s sixteenth studio album "New", and appears as the sixth track on '
 'the album. After being released early on the iTunes Store as a track '
 'available to download from "New" on 28 August 2013, the song was released as '
 'a single on 2 September 2013 and available exclusively on Amazon.com. The '
 "single's premiere on 28 August was concurrent with the official reveal of "
 '"New" on the same day. The single gained heavy airplay on Japanese radio '
 'stations, where it became a number 4 hit on the Japan Hot 100. The single '
 "joined BBC Radio 2's playlist and the album of the same name was their "
 'Record of the Week. The song appears in both the opening and the end credits '
 'of the 2013 animated film "Cloudy with a Chance of Meatballs 2". Reception. '
 '"New" was greeted positively by critics and the musical press. As well as '
 "being selected as BBC Radio 2's Record of the Week and placed on their "
 'A-list, the track was greeted as the "Track of the Day" by "Mojo" which '
 'praised its "doe-eyed optimism, irresistible melody" and "orchestrated pop '
 'arrangements". "Rolling Stone"\'s Will Hermes, praised its "bouncy '
 'harpsichord-laden melody", giving it a four-star rating and drawing '
 'comparisons to the Beatles\' "Got to Get You into My Life", a view shared by '
 '"The Daily Telegraph" which described it as a "jaunty, Beatles-esque stomp".')
Enter label for Chunk 1 (0 or 1, or type 'exit' to quit):
```

Figure 5: The command-line screen used for the annotation process.

Query What is Anthony Sharps occupation?
Answer [actor, actress, actors, actresses]
Context 1. British actor (1915–1984) Dennis Anthony John Sharp (16 June 1915 – 23 July 1984) was an English actor, writer and director. Stage career. Anthony Sharp was a graduate of the London Academy of Music and Dramatic Art (LAMDA) and made his stage... 2. There he played Benedick in "Much Ado About Nothing" in 1958 and Malvolio in "Twelfth Night" the following year. Rejoining the company in the 1970s, he appeared in such plays as "Love's Labour's Lost" and "The Man of... 3. His credits included "Any Other Business" (Westminster Theatre 1958), "Caught Napping" (Piccadilly Theatre 1959), "Wolf's Clothing" (Strand Theatre 1959), "Billy Bunter Flies East" (Victoria Palace 1959), "The... 4. His only starring role in a feature film was the homicidal priest Father Xavier Meldrum in Pete Walker's 1975 horror picture "House of Mortal Sin". His final feature film, in which he played foreign secretary Lord Ambrose,... 5. In 1974, he appeared as the vicar in the radio version of "Steptoe and Son", and in 1978 he was both Garkbit, the waiter in the Restaurant at the End of the Universe , and The Great Prophet Zarquon in Fit the Fifth of the...

Table 8: Data sample with 5 relevant chunks. These examples include queries with general questions such that the answer can be inferred throughout the entire Wikipedia article.

LLM	Retriever	Noise Vulnerability(↓)	Context Acceptability (↑)	Context Insensitivity (↓)	Context Misintepretation (↓)
		Top-1			
GPT3.5	**BGE-S**	25.69	65.33	8.44	0.54
	BGE-B	24.36	66.67	8.44	0.54
	BGE-L	21.67	69.35	8.44	0.54
	Contriever	42.26	48.76	8.44	0.54
	NV	**17.66**	**73.36**	8.44	0.54
GPT4o	**BGE-S**	22.05	69.09	8.29	0.57
	BGE-B	20.57	70.57	8.29	0.57
	BGE-L	18.83	72.31	8.29	0.57
	Contriever	35.73	55.42	8.29	0.57
	NV	**15.38**	**75.76**	8.29	0.57
LLAMA2-7B	**BGE-S**	26.83	56.13	16.37	0.67
	BGE-B	25.77	57.19	16.37	0.67
	BGE-L	22.80	60.16	16.37	0.67
	Contriever	47.63	35.33	16.37	0.67
	NV	**18.21**	**64.75**	16.37	0.67
Qwen2-7B	**BGE-S**	29.1	61.11	9.41	0.37
	BGE-B	27.75	62.46	9.41	0.37
	BGE-L	24.29	65.93	9.41	0.37
	Contriever	51.14	39.07	9.41	0.37
	NV	**19.04**	**71.17**	9.41	0.37

Table 9: Top1 performance for RAG systems

LLM	Retriever	Noise Vulnerability(↓)	Context Acceptability (↑)	Context Insensitivity (↓)	Context Misintepretation (↓)
		Top-3			
GPT3.5	**BGE-S**	18.52	72.50	8.44	0.54
	BGE-B	17.39	73.63	8.44	0.54
	BGE-L	16.00	75.02	8.44	0.54
	Contriever	30.91	60.11	8.44	0.54
	NV	**13.16**	**77.87**	8.44	0.54
GPT4o	**BGE-S**	14.02	77.12	8.29	0.57
	BGE-B	13.67	77.47	8.29	0.57
	BGE-L	12.38	78.77	8.29	0.57
	Contriever	25.94	65.20	8.29	0.57
	NV	**10.75**	**80.39**	8.29	0.57
LLAMA2-7B	**BGE-S**	22.32	60.64	16.37	0.67
	BGE-B	21.72	61.23	16.37	0.67
	BGE-L	20.32	62.63	16.37	0.67
	Contriever	36.56	46.40	16.37	0.67
	NV	**16.81**	**66.15**	16.37	0.67
Qwen2-7B	**BGE-S**	21.88	68.33	9.41	0.37
	BGE-B	21.04	69.18	9.41	0.37
	BGE-L	19.33	70.88	9.41	0.37
	Contriever	37.09	53.13	9.41	0.37
	NV	**15.83**	**74.39**	9.41	0.37

Table 10: Top3 performance for RAG systems

LLM	Retriever	Noise Vulnerability(↓)	Context Acceptability (↑)	Context Insensitivity (↓)	Contect Misintepretation (↓)
Top-5					
GPT3.5	**BGE-S**	17.36	73.66	8.44	0.54
	BGE-B	16.42	74.60	8.44	0.54
	BGE-L	15.36	75.66	8.44	0.54
	Contriever	27.45	63.57	8.44	0.54
	NV	**13.21**	**77.81**	8.44	0.54
GPT4o	**BGE-S**	13.17	77.97	8.29	0.57
	BGE-B	12.79	78.35	8.29	0.57
	BGE-L	11.78	79.36	8.29	0.57
	Contriever	22.65	68.49	8.29	0.57
	NV	**10.64**	**80.50**	8.29	0.57
LLAMA2-7B	**BGE-S**	24.12	58.84	16.37	0.67
	BGE-B	24.08	58.88	16.37	0.67
	BGE-L	23.19	59.77	16.37	0.67
	Contriever	36.78	46.17	16.37	0.67
	NV	**19.93**	**63.03**	16.37	0.67
Qwen2-7B	**BGE-S**	22.5	67.71	9.41	0.37
	BGE-B	21.82	68.40	9.41	0.37
	BGE-L	20.15	70.06	9.41	0.37
	Contriever	34.46	55.76	9.41	0.37
	NV	**17.61**	**72.60**	9.41	0.37

Table 11: Top5 performance for RAG systems