MRAG: A Modular Retrieval Framework for Time-Sensitive Question Answering

Siyue Zhang[◇♡♠] **Yuxiang Xue**[♠] **Yiming Zhang**[♣] **Xiaobao Wu**[◇]
Anh Tuan Luu[◇] **Chen Zhao**[♠★]

[◇]Nanyang Technological University [♡]Alibaba-NTU JRI [♠]NYU Shanghai
[♣]Zhejiang University [★]Center for Data Science, New York University
siyue001@e.ntu.edu.sg cz1285@nyu.edu

Abstract

Understanding temporal concepts and answering time-sensitive questions is crucial yet a challenging task for question-answering systems powered by large language models (LLMs). Existing approaches either update the parametric knowledge of LLMs with new facts, which is resource-intensive and often impractical, or integrate LLMs with external knowledge retrieval (*i.e.,* retrieval-augmented generation). However, off-the-shelf retrievers often struggle to identify relevant documents that require intensive temporal reasoning. To systematically study time-sensitive question answering, we introduce the TEMPRAGEVAL benchmark, which repurposes existing datasets by incorporating complex temporal perturbations and gold evidence labels. As anticipated, all existing retrieval methods struggle with these temporal reasoning-intensive questions. We further propose Modular Retrieval (MRAG), a trainless framework that includes three modules: (1) *Question Processing* that decomposes question into a main content and a temporal constraint; (2) *Retrieval and Summarization* that retrieves, splits, and summarize evidence passages based on the main content; (3) *Semantic-Temporal Hybrid Ranking* that scores semantic and temporal relevance separately for each fine-grained evidence. On TEMPRAGEVAL, MRAG significantly outperforms baseline retrievers in retrieval performance, leading to further improvements in final answer accuracy.[1]

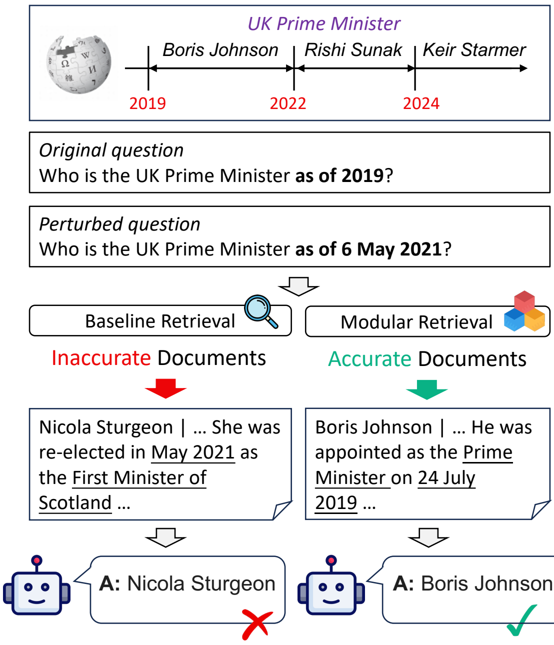


Figure 1: A time-sensitive question example that requires temporal reasoning (as of 6 May 2021 → 2019 - 2022) to both retrieve documents and generate answers. State-of-the-art retrieval systems struggle to conduct in-depth reasoning to identify relevant documents. We provide a new diagnostic benchmark TEMPRAGEVAL, and propose a new modular framework to tackle this challenge.

1 Introduction

Facts are constantly evolving in our ever-changing world. This dynamic nature highlights the need for natural language processing (NLP) systems capable of updating information (Liška et al., 2022; Zhang et al., 2024; Kasai et al., 2024) and providing accurate responses to time-sensitive questions (Chen et al., 2021; Chu et al., 2024). For instance, a common query like "Who is the UK Prime Minister?" sees the answer transition from "`Boris Johnson`" to "`Rishi Sunak`" in 2022 (Figure 1).

With developments of large language models (LLMs), existing approaches rely on the parametric knowledge of LLMs to answer time-sensitive questions directly, and constantly update the parametric knowledge on new facts (Rozner et al., 2024; Wu et al., 2024a; Wang et al., 2024). However, updating LLM parameters are often resource-intensive. An alternative line of research explores Retrieval-Augmented Generation (RAG), which integrates LLMs with external knowledge (*e.g.,* Wikipedia)

[1]Our code and data are available at `https://github.com/siyue-zhang/MRAG`.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3080–3118
November 4-9, 2025 ©2025 Association for Computational Linguistics

through information retrieval (Izacard et al., 2020). While RAG allows for the incorporation of new facts with minimal effort, its performance heavily relies on off-the-shelf retrieval systems, which are often limited to keywords or semantic matching. Time-sensitive questions, however, often require intensive temporal reasoning to identify relevant documents, *i.e.,* reasoning-intensive retrieval (Su et al., 2024a). For example, in Figure 1, retrievers should infer the date "`24 July 2019`" as relevant to the constraint "`as of 6 May 2021`", rather than only match with date like "`May 2021`". Although temporal QA is a well-established domain, there remains a lack of research for temporal reasoning-intensive retrieval systems. We fill that lacuna.

We begin by conducting a diagnostic evaluation of existing retrieval approaches for temporal reasoning-intensive retrieval. Following the idea of systematic evaluation with contrast set (Gardner et al., 2020), we repurpose two existing datasets, TIMEQA (Chen et al., 2021) and SITUATEDQA (Zhang and Choi, 2021), to introduce the **Temp**oral QA for **RAG Eval**uation benchmark (TEMPRAGEVAL). We manually augment the test questions with complex temporal perturbations (*e.g.,* modifying the time period to avoid textual overlap). In addition, we annotate gold evidence on Wikipedia for more accurate retrieval evaluation. We observe remarkable degradation of retrieval performance, which points out a serious robustness issue for existing retrievers.

To address this issue, we propose a training-free Modular Retrieval framework (MRAG) to enhance temporal reasoning-intensive retrieval. Complementary to recent agentic RAG methods such as R1-Searcher and Search-o1 (Song et al., 2025; Li et al., 2025), MRAG provides an improved retrieval component and reduces the number of invocations for retrieval when handling complex temporal questions. Specifically, MRAG contains three key modules: (1) **Question Processing**, which decomposes each question into a main content and a temporal constraint; (2) **Retrieval and Summarization**, which utilizes off-the-shelf retrievers to find evidence passages based on the main content, segments them into independent sentences, and guides LLMs to condense the most relevant passages into query-specific sentences.(3) **Semantic-Temporal Hybrid Ranking**, which ranks each evidence sentence using a combination of a semantic score measuring semantic similarity, and a temporal score, a novel symbolic component that assesses temporal relevance to the query's temporal constraint.

On TEMPRAGEVAL, our proposed MRAG framework achieves substantial improvements in performance, with 9.3% top-1 answer recall and 11% top-1 evidence recall. We also incorporate state-of-the-art (SOTA) answer generators (Asai et al., 2024; Yan et al., 2024; Wei et al., 2022), and demonstrate that the improvements in retrieval from MRAG propagate to enhanced final QA accuracy, with 4.5% for both exact match and F1. Detailed case studies further confirms MRAG's robustness to temporal perturbations qualitatively.

Our contributions can be summarized as follows.

- We introduce TEMPRAGEVAL, a **time-sensitive RAG benchmark** to diagnostically evaluate each component of existing retrieval-augmented generation systems.
- We propose MRAG, a **modular retrieval framework** to separately determine semantic and temporal relevance.
- On TEMPRAGEVAL, MRAG significantly outperforms all baseline retrieval systems, and the improvements lead to better answer generation.

2 Background

In this section, we first define the time-sensitive question answering task (§2.1), and then introduce the baseline retrieval-augmented generation QA based systems (§2.2).

2.1 Temporal Question Answering

There has been extensive research on Temporal Question Answering over Knowledge Graphs (TKGQA). Prior works have developed knowledge graph-based benchmarks (e.g., CronQuestions (Saxena et al., 2021) and TimeQuestions (Jia et al., 2021)), graph neural network models (Chen et al., 2022; Liu et al., 2023; Sharma et al., 2023), and TKG-augmented LLMs (Qian et al., 2024; Du et al., 2025a,b). However, these methods heavily rely on knowledge graphs, which are costly to build, domain-specific, and prone to becoming outdated (Kau et al., 2024). Therefore, we focus on the free-text setting, which is more universal and up-to-date. Instead of building temporal-specific graph models, we improve general-purpose text retrievers for temporal queries. We use Wikipedia as the main knowledge corpus D.[2] Our approach is

[2]December 2021 Wikipedia dump, comprising 33.1 million text chunks.

broadly applicable to other collections, such as the New York Times Annotated Archive (Sandhaus, 2008) and ClueWeb (Overwijk et al., 2022).

2.2 Retrieval-Augmented Generation

The goal of RAG is to address the limitations in the parametric knowledge of LLMs by incorporating external knowledge.

Passage retrieval and reranking. Retrieval methods are typically categorized into sparse retrieval and dense retrieval. Sparse retrieval methods like BM25 (MacAvaney et al., 2020) rely on lexical matching. In contrast, dense retrieval models (Karpukhin et al., 2020; Zhao et al., 2021; Thakur et al., 2021; Izacard et al., 2021; Zhang et al., 2025) encode the question q and passage p into low-dimension vectors separately. The semantic similarity is computed using a scoring function (*e.g.,* dot product) as:

$$f(q,p) = \text{sim}(\text{Enc}_Q(q), \text{Enc}_P(p)), p \in \mathcal{D}. \quad (1)$$

However, these bi-encoder models lack the ability to capture fine-grained interactions between the query and passage. A common optional[3] approach is to have another cross-encoder model to rerank top passages. Cross-encoder models (Khattab and Zaharia, 2020; Wang et al., 2020; Gemma et al., 2024) jointly encode the query q and the passage p together by concatenating them as input into a single model as:

$$f(q,p) = \text{sim}(\text{Enc}([q;p])), p \in \mathcal{D}. \quad (2)$$

Answer generation. Recent reader systems are mainly powered by LLMs with strong reasoning capabilities. With recent advancements in long-context LLMs (Dubey et al., 2024; Lee et al., 2024b), top documents are typically concatenated with the query as reader input:

$$y = \text{Dec}(\text{Enc}([p_1; \ldots; p_k; q])). \quad (3)$$

To unlock the reasoning capabilities of LLMs, Chain-of-Thought (CoT) prompting (Wei et al., 2022) introduces intermediate reasoning steps for improved performance. Self-RAG (Asai et al., 2024) critiques the retrieved passages and its own generations. Recent agentic RAG systems dynamically acquire and integrate external knowledge (using either retriever models or search APIs) during the reasoning process (Li et al., 2025; Song et al., 2025; Jin et al., 2025).

[3]Note that reranking is not always adopted, as it adds additional computational cost.

Dataset	# Eval.	Evid.	Natu.	Comp.
ComplexTQA	10M			✓
StreamingQA	40K	✓	✓	
TempLAMA	35K			
SituatedQA	2K		✓	
TimeQA	3K			✓
MenatQA	2K			✓
TempRAGEval	1K	✓	✓	✓

Table 1: Comparison of temporal QA datasets. TEMPRAGEVAL is featured by manual evidence annotations, human-written question (*i.e.,* Naturalness), and higher complexity in temporal reasoning.

3 TEMPRAGEVAL Benchmark

In this section, we first present existing time-sensitive QA datasets (§3.1), then introduce our diagnostic benchmark dataset TEMPRAGEVAL (§3.2), finally we evaluate existing retrieval approaches on TEMPRAGEVAL (§3.3).

3.1 Existing Time-Sensitive QA Datasets

There are several existing QA datasets that focus on temporal reasoning. The most representative ones are the following:[4]

- **SITUATEDQA** (Zhang and Choi, 2021) is a time-sensitive QA dataset where the answer to an information-seeking question varies based on temporal context. These questions contain a single type of temporal constraint (*e.g.,* "`as of`") that directly align with the answers. Retrievers with surface-form date matching often exploit these shortcuts to bypass the need for temporal reasoning.
- **TIMEQA** (Chen et al., 2021) is another time-sensitive QA dataset. Unlike SITUATEDQA, the questions in hard split include complex temporal constraints (*e.g.,* "`between 2012 to 2018`"). However, question-answer pairs are *synthetically* generated from time-evolving WikiData facts using templates. In addition, TIMEQA does not include evidence annotations, making it imprecise to evaluate retrieval results.

According to Table 1, we observe that none of existing datasets include these key factors for sys-

[4]We mainly focus on these two datasets, while others, such as Liška et al. (2022); Dhingra et al. (2022); Gruber et al. (2024), serve as alternatives. Datasets such as Tan et al. (2023); Virgo et al. (2022) that are not knowledge-intensive, are excluded from this work (Appendix B).

tematically evaluating current retrieval (and answer generation) systems: (1) Evidence annotation; (2) Natural questions from users; (3) Complex temporal constraints. Therefore, as we will show in the following section, we aim to address this limitation by creating TEMPRAGEVAL.

3.2 TEMPRAGEVAL Construction

We create TEMPRAGEVAL, a time-sensitive QA benchmark for rigorously evaluating temporal reasoning in both retrieval and answer generation.

Perturbed question-answer pair generation. Annotators first select question-answer pairs from the SITUATEDQA and TIMEQA datasets[5] that can be grounded in Wikipedia facts with key timestamps or durations. They then revise temporal perturbations by selecting implicit conditions, temporal relations, and alternative dates to include complex temporal reasoning, without changing the final answer[6]. Annotators are also required to edit the question text to improve naturalness. We include detailed guidelines in Appendix C.

Evidence annotation. To better evaluate the performance of retrieval systems, we supplement question-answer pairs with up to two annotated gold evidence passages. A passage is relevant to the question if annotators can obtain the correct answer based on the passage. Specifically, for each question, annotators are asked to manually review top-20 passages retrieved by Contriever (Izacard et al., 2021) and reranked by the best GEMMA (Li et al., 2023) reranker. If there is no relevant passage, annotators are required to search Wikipedia pages related to the query entities to locate the gold evidence (around 12.7% of questions). We create 1,000 test examples with human-annotated evidence. Appendix D presents sample statistics, revealing that SITUATEDQA questions include popular entities while the entities for TIMEQA questions are long-tailed.

3.3 Preliminary Evaluation on TEMPRAGEVAL

In TEMPRAGEVAL, we first evaluate the performance on SOTA retrieval systems as a sanity check.

[5]Since both datasets lack questions about knowledge beyond the cutoff date of existing LLMs, we primarily focus on historical knowledge and discuss potential future directions on recent knowledge in the Limitations section.

[6]Questions with the same content but different temporal constraints and answers are considered different samples. Perturbations are introduced for each sample.

Experimental Setup. We follow the popular retrieve-then-rerank pipeline, using the dense retriever Contriever (Izacard et al., 2021) and the LLM-based reranker GEMMA (Gemma et al., 2024). The retriever finds top 1,000 passages, and among them, the reranker reorders the top 100 passages. We use two evaluation metrics: **Answer Recall (AR@k)** that measures the proportion of samples where at least one answer appears within the top-*k* retrieved passages, and **Gold Evidence Recall (ER@k)** that assesses the percentage of samples where at least one gold evidence document is included in the top-*k* passages.

Performance degradation on perturbed questions. As shown in Figure 2, we observe a significant degradation in retrieval performance caused by temporal perturbations. For instance, for the GEMMA baseline, the top-1 answer recall and evidence recall drop from 85.8% to 54.7% and from 45.0% to 20.3% on TEMPRAGEVAL-SITUATEDQA. This is because the perturbed temporal constraints avoid matching between timestamps in the questions and the passages. Consequently, retrievers must conduct in-depth temporal reasoning to identify the relevant passages.

We further conduct a controlled experiment to reveal the temporal reasoning capabilities of existing retrieval methods. Specifically, we compute the similarity scores for query-evidence pairs by varying the temporal relation in the query, *e.g.,* "`before`", "`after`", and "`as of`", and the timestamp in the evidence, *e.g.,* from "`1958`" to "`1965`". Experiments confirm that all methods prioritize matching *exact* dates indicating a shortcut for temporal reasoning in retrieval. We present full results in Figure 6 in Appendix.

4 MRAG: Modular Retrieval

Motivated by the performance degradation of the existing retrieval methods in the preliminary evaluation, we propose a Modular Retrieval (MRAG) framework (as shown in Figure 3) to enhance temporal reasoning-intensive retrieval. At a high level, MRAG disentangles relevance-based retrieval from temporal reasoning, leveraging a dense embedding model for semantic scoring and a set of symbolic heuristics for temporal scoring. Specifically, MRAG has three key modules: question processing, retrieval and summarization, and semantic-temporal hybrid ranking.

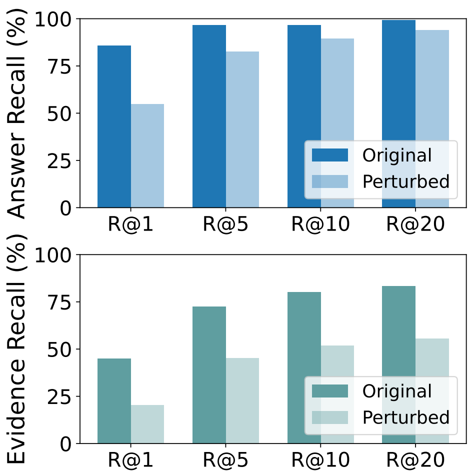


Figure 2: The retrieval performance degradation of the GEMMA baseline on TEMPRAGEVAL-SITUATEDQA, comparing original and perturbed questions (see TEMPRAGEVAL-TIMEQA in Appendix E).

Question processing. We prompt LLMs to decompose each time-sensitive question into a main content (MC) and a temporal constraint (TC). This approach disentangles temporal relevance from semantic relevance: MC measures the semantic relevance of the evidence, while TC determines its temporal relevance.

Retrieval and summarization. We apply off-the-shelf retrievers (*e.g.,* Contriever) to find relevant passages to MC in Wikipedia. Then we employ reranker models to reorder these passages by semantic similarity to MC.

It is common for a passage to contain multiple pieces of temporal information, most of which are unrelated to the question and can distract the temporal scoring component introduced next. For example, the relevant passage in Figure 3 includes the sentence, "`it was filmed in 2017`", which satisfies the TC but is irrelevant. Therefore, we split passages into individual sentences to eliminate temporal distractors. However, as shown in Figure 3, critical information from the most relevant passages—such as "`America's Next Top Model`", "`The winner of the competition`", and "`January 9, 2018`"—can be scattered across different sentences. Relying solely on sentence splitting would miss key details. To overcome this challenge, we additionally employ LLMs to summarize each of the top-k passages into a single sentence condensing relevant phrases and temporal information, as analyzed in §6.1.[7]

Semantic-temporal hybrid ranking. We rerank each sentence (summarized from a LLM or segmented from the original passage) with two distinct scores: a semantic score and a temporal score. The semantic score is calculated from the similarity between the evidence sentence and MC. For temporal score, we first extract the timestamp from each sentence (*e.g.,* "`2018`"). Based on the timestamp and TC (*e.g.,* "`as of 2021`"), we compute a temporal score using symbolic functions similar to temporal activation functions in Chen et al. (2022). The final score for each sentence is obtained by multiplying the semantic score and the temporal score. Finally, we select the passages that contains the highest-scoring sentences. We include the details of symbolic functions in Appendix G.

5 Experiments

In this section, we evaluate MRAG and baseline systems on TEMPRAGEVAL.

5.1 Experimental Setup

Baselines. For retrieval, we include BM25, Contriever, and a hybrid method (Jedidi et al., 2024). Reranking methods include ELECTRA (Clark et al., 2020), MiniLM (Wang et al., 2020), Jina(Jina, 2024), BGE (Xiao et al., 2023), NV-Embed (Lee et al., 2024a), and GEMMA (Gemma et al., 2024). We follow state-of-the-art answer generation approaches based on prompting LLMs (§2.2). We evaluate four approaches, **Direct Prompt** that adopts question-answer pairs as few-shot examples; **Direct CoT** that adds rationals into prompts (Wei et al., 2022); **RAG-Concat**, where passages are concatenated into a LLM; and **Self-RAG**, which processes each passage independently and selects the best answer (Asai et al., 2024).

Metrics. We use the same setup in §3.3 for retrieval evaluation. For answer evaluation, we use **Exact Match (EM)** that measures the exact match to the gold answer, and **F1 score (F1)** that measures the word overlap to the gold answer.

Implementation details. Due to limited budget, we evaluate GPT-4o mini (OpenAI et al., 2024) with direct prompting, and three open-source LLMs: TIMO, a LLaMA2-13B model fine-tuned for temporal reasoning (Su et al., 2024b), and two general-purpose models, Llama3.1-8B-Instruct and

[7]While using LLMs to summarize passages can introduce hallucinations, we mitigate this by summarizing only the top-k passages.

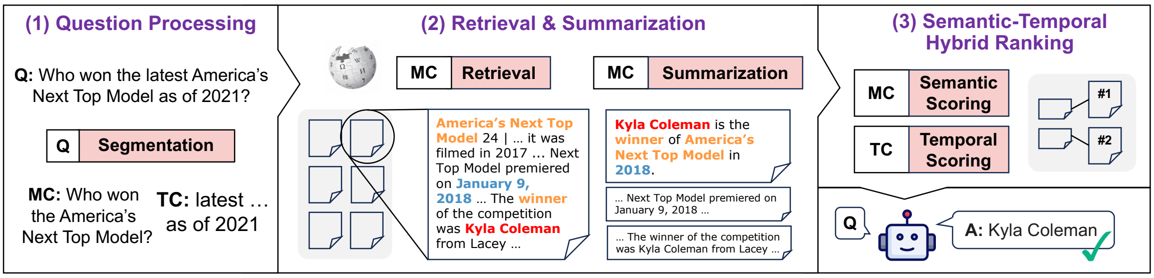


Figure 3: An overview of the MRAG framework, consisting of three key modules: question processing, retrieval and summarization, and semantic-temporal hybrid ranking. The question processing module separates each query into the main content (*i.e.,* MC) and the temporal constraint (*i.e.,* TC). The retrieval and summarization module finds the most relevant evidence based on the main content and summarizes or splits these evidence into fine-grained sentences. The hybrid ranking module combines symbolic temporal scoring and dense embedding-based semantic scoring at a fine-grained level to determine the final evidence ranking.

Llama3.1-70B-Instruct (Dubey et al., 2024). We use 10 examples in prompts. To eliminate the impact of input length constraints across models, we conduct parametric studies as described in Appendix K and report each LLM's performance with its optimal number of input passages in Table 3.

5.2 Main Results

MRAG enhances retrieval performance for time-sensitive questions. According to Table 2, MRAG significantly outperform all retrieve then rerank baselines, which highlight the superior temporal reasoning capabilities. For example, MRAG improves the best baselines Contriver + Gemma significantly, with 7.7% top-5 evidence recall in TempRAGEval-TimeQA and 13.9% top-5 evidence recall in TempRAGEval-SituatedQA.

Retrieval augmentation improves time-sensitive QA performance. According to Table 3, we observe that LLMs relying solely on their parametric knowledge struggle to accurately answer time-sensitive questions, with limited QA accuracy. Incorporating retrieval-augmented generation significantly improves QA accuracy. Notably, we observe larger improvements on TempRAGEval-TimeQA, which primarily focuses on less frequent entities that pose greater challenges to the parametric knowledge of LLMs (Kandpal et al., 2023).

Enhanced retrieval contributes to improved time-sensitive QA performance. As shown in Table 3, MRAG outperforms baseline RAG approaches in QA accuracy, for instance 49.2% EM (MRAG) over 44.0% EM (RAG) in TempRAGEval-TimeQA for Llama3.1-8B. Incorporating a self-reflection strategy improves performance for Llama3.1 models but not for TIMO, likely due to the limited reasoning capacity of its backbone model, Llama 2.

6 Analysis

This section presents a detailed analysis of the results and the contribution of each MRAG module.

6.1 Ablation Study

The impact of the number of passages for summarization. Our LLM based summarization removes irrelevant temporal information but may also introduce hallucinations (details in Appendix H). The parametric study at the bottom of Table 2 shows that summarizing top-five passages achieves the best balance. Additionally, we compare the RAG setup with the long-document QA setup bypassing retrievers in Appendix L. RAG achieves a better accuracy, as retrievers exclude irrelevant passages and reduce noise.

The impact of the number of passages for answer generation. Our experiments show that the optimal number of passages depends on the LLMs. TIMO can handle a maximum of 3 passages, while the optimal number for LLaMA 3.1-8B is five, and for LLaMA 3.1-70B, it is twenty as shown in Table 4. Full results are presented in Appendix K.

The impact of key retrieval steps in MRAG. In the retrieval and summarization module, with retrieved passages, MRAG conducts two key steps: (1) *Passage Keyword Ranking* reduces the passages from 1,000 to 100 based on the keyword presence; (2) *Passage Semantic Ranking* reorders the top-100 passages and splits them into ∼500 sentences, including chunk summaries. Similarly, in the semantic-temporal hybrid ranking module, an-

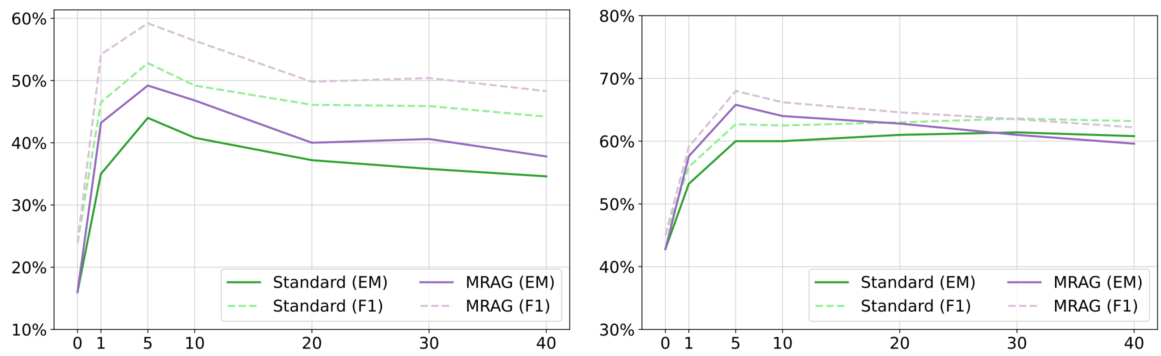
Method			TEMPRAGEVAL-TIMEQA				TEMPRAGEVAL-SITUATEDQA			
			AR @		ER @		AR @		ER @	
1st	2nd	# QFS	1	5	1	5	1	5	1	5
BM25	-	-	17.5	39.0	4.2	14.1	27.6	58.2	6.8	18.4
Cont.	-	-	18.8	49.9	9.6	28.7	22.6	51.1	6.8	17.1
Hybrid	-	-	18.8	51.2	9.6	28.1	22.6	55.8	6.8	19.7
Cont.	ELECTRA	-	40.1	76.9	21.8	58.6	35.5	71.3	15.3	37.1
Cont.	MiniLM	-	34.0	76.1	16.2	57.3	36.8	73.4	20.0	40.3
Cont.	Jina	-	42.4	77.2	23.6	58.6	47.9	78.4	19.5	41.1
Cont.	BGE	-	40.3	80.9	23.3	61.3	36.3	74.2	14.5	35.0
Cont.	NV-Embed	-	49.9	81.2	33.4	62.9	47.4	81.3	23.4	46.1
Cont.	Gemma	-	46.7	82.5	26.0	66.6	54.7	82.6	20.3	45.3
Cont.	NV-Embed-v2	-	54.4	83.0	33.4	64.5	54.7	82.1	25.5	48.4
Cont.	MRAG	-	57.6	89.4	32.4	73.5	61.1	88.2	27.4	56.3
Cont.	MRAG	5	**58.6**	**90.0**	**37.1**	**74.3**	61.3	**89.0**	**31.1**	**59.2**
Cont.	MRAG	10	56.0	88.1	35.5	73.2	**62.1**	87.9	30.8	57.9

Table 2: The answer recall (AR@k) and gold evidence recall (ER@k) of each retrieval method on perturbed temporal queries in TIMEQA and SITUATEDQA subsets of TEMPRAGEVAL. 1st means the first-stage retrieving method; 2nd means the second-stage reranking method; # QFS means the number of top passages to be summarized. **Bold** numbers indicate the best performance. We include complete results in Appendix I.

Method	TEMPRAGEVAL-TimeQA			TEMPRAGEVAL-SituatedQA		
	# Docs	EM	F1	# Docs	EM	F1
			GPT4o-mini			
Direct Prompt	-	19.6	30.6	-	54.2	58.6
			TIMO			
Direct Prompt	-	16.2	24.8	-	50.6	53.1
Direct CoT	-	15.8	28.2	-	49.4	53.9
RAG-Concat	3	43.4	<u>55.2</u>	3	55.8	58.1
MRAG-Concat	3	**48.2**	**57.2**	3	<u>61.4</u>	<u>63.6</u>
Self-MRAG	3	<u>44.6</u>	54.9	3	**62.4**	**64.5**
			Llama3.1-8B-Instruct			
Direct Prompt	-	16.0	23.9	-	42.8	45.0
Direct CoT	-	16.8	27.8	-	49.6	54.5
RAG-Concat	5	44.0	52.8	5	60.0	62.7
MRAG-Concat	5	<u>49.2</u>	<u>59.2</u>	5	<u>65.8</u>	<u>68.0</u>
Self-MRAG	5	**54.2**	**65.6**	5	**66.4**	**68.2**
			Llama3.1-70B-Instruct			
Direct Prompt	-	31.0	42.3	-	59.0	62.1
Direct CoT	-	33.2	45.8	-	69.0	72.6
RAG-Concat	5	54.4	63.2	20	67.0	69.8
MRAG-Concat	5	<u>58.0</u>	<u>68.4</u>	20	<u>69.2</u>	<u>72.5</u>
Self-MRAG	5	**61.2**	**75.3**	20	**72.2**	**76.0**

Table 3: End-to-end QA performance comparison for various generation strategies and LLMs on TEMPRAGEVAL. **Bold** numbers indicate the best performance the each backbone LLM. The second best is <u>underlined</u>. TIMO has a limited input length with up to three passages. We report the best number of passages for Llama models, and provide ablation on different numbers in Appendix K.

other two key steps are conducted: (1) *Sentence Keyword Ranking* further narrows the scope to 200 sentences; (2) *Hybrid Ranking* ranks these sentences by semantic-temporal hybrid scoring. As shown in Table 4, keyword-based ranking steps effectively reduce ranking candidates without significant performance loss, while semantic and hybrid ranking steps markedly enhance retrieval performance. An efficiency-accuracy trade-off can be made by adjusting the number of targeted passages or sentences at each step.



(a) TEMPRAGEVAL-TIMEQA

(b) TEMPRAGEVAL-SITUATEDQA

Figure 4: Llama3.1-8B-Instruct reader performance versus number of concatenated context passages retrieved by the GEMMA and MRAG methods. Standard refers to RAG-Concat and MRAG refers to MRAG-Concat.

Computational Overhead. As MRAG involves retrieval, summarization, and re-ranking, it incurs approximately twice the computational overhead of standard RAG pipelines, which is manageable. We provide a detailed analysis in Appendix M.

6.2 Human Evaluation

One limitation of the retrieval metrics is that AR@k overestimates performance, as a passage might incidentally contain an answer without directly supporting it. Conversely, ER@k acts as a conservative lower bound, potentially overlooking other relevant but unannotated passages. To address this, we conduct a human evaluation on a random subset of 200 examples to assess actual retrieval performance. The results validate the advantages of MRAG over GEMMA in retrieval accuracy with full results presented in Appendix J.

6.3 Case Study

We analyze retrieval errors qualitatively to highlight the advantages of MRAG over the GEMMA baseline. As shown in Figure 5, the top-1 passage by GEMMA matches the query date "1988" but discusses a father-son record set in 2007. In contrast, MRAG retrieves a passage about a teammate combination record from the same season, despite the differing date ("1961" vs. "1988"). Additional cases and answer generation error cases are provided in Appendix O.1 and Appendix O.2.

7 Other Related Works

LLM embeddings. Recent research has explored LLM embeddings for retrieval. Some studies focus on distilling or fine-tuning LLM embeddings for reranking tasks, such as GEMMA (Gemma et al., 2024) and MiniCPM (Hu et al., 2024). Others aim to develop generalist embedding models capable of performing a wide range of tasks including retrieval and reranking, *e.g.*, gte-Qwen (Yang et al., 2024) and NV-Embed (Lee et al., 2024a). These LLM-based methods have demonstrated unprecedented performance in benchmarks such as MTEB (Muennighoff et al., 2023) and our TEMPRAGEVAL.

Reasoning intensive retrieval. Existing retrieval benchmarks primarily target keyword-based or semantic-based retrieval. Su et al. (2024a) introduces BRIGHT, a new retrieval task emphasizing intensive reasoning. We focus on temporal reasoning, one aspect of a broader class of reasoning-intensive retrieval. MRAG is expected to generalize to other forms of symbolic reasoning, such as numeric ranges and geospatial constraints. It mitigates direct numeric matching in retrieval and

Question:
Who had the most home runs by two teammates in a season as of 1988?

Gemma #1 Passage:
Bobby Bonds | ... until José Canseco of the Oakland Athletics in **1988**. Barry and Bobby had 1,094 combined home runs through 2007 — a record for a father-son combination ...

MRAG #1 Passage:
50 home run club | **M&M Boys**—are the only teammates ... hitting a combined 115 home runs in **1961** and breaking the single-season record for home runs by a pair of teammates.

Figure 5: A case study for top-1 passage retrieved by GEMMA and MRAG from TEMPRAGEVAL.

	# Chunk / Sent.	Answer Recall @				Gold Evidence Recall @			
Steps		**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
Chunk Retrieving	21M	22.6	51.1	65.5	79.5	6.8	17.1	22.9	30.5
+ Chunk Keyword Ranking	1000	29.2	62.4	77.4	87.1	12.4	26.1	35.5	46.8
+ Chunk Semantic Ranking	100	55.3	84.0	88.2	95.5	22.9	50.0	59.0	66.8
+ Sent. Keyword Ranking	500	55.3	81.8	85.5	91.6	23.4	49.0	55.3	62.4
+ Sent. Hybrid Ranking	200	**61.3**	**89.0**	**93.4**	**94.2**	**31.1**	**59.2**	**65.8**	**69.0**

Table 4: The ablation study of key steps of MRAG on perturbed queries in TEMPRAGEVAL − SITUATEDQA.

enhances reasoning capabilities.

8 Conclusion

This study focuses on time-sensitive QA, a task that challenges LLM based QA systems. We first present TEMPRAGEVAL, a diagnostic benchmark featuring natural questions, evidence annotations, and temporal complexity. We further propose a training-free MRAG framework, which disentangles relevance-based retrieval from temporal reasoning and introduces a symbolic temporal scoring mechanism. While existing systems struggle on TEMPRAGEVAL due to limited temporal reasoning capacities in retrieval, MRAG shows significant improvements. We hope this work advances future research on reasoning-intensive retrieval.

Limitations

There are still some limitations in our work: (1) Our proposed benchmark is designed to evaluate time-sensitive questions with explicit temporal constraints. However, addressing questions with implicit temporal constraints presents a more complex challenge for retrieval systems. We could extend to implicit temporal reasoning by associating explicit information to the implicit one using LLM common sense and background knowledge like (Chen et al., 2022). (2) Our dataset does not include time-sensitive questions that fall outside the LLM knowledge cutoff. We could extend our dataset with questions from RealTime QA (Kasai et al., 2024) and AntiLeak-Bench (Wu et al., 2024b). (3) Our main objective is to improve temporal reasoning in retrieval, which has not been tackled by previous works. More complex scenarios like multi-hop and recursive reasoning require further research efforts. (4) The proposed framework introduces computational overhead for improved performance as detailed in Appendix M. (5) We analyze knowledge conflicts between LLMs and passages in Appendix P and leave conflicts among passages for future work.

Ethics Statement

TEMPRAGEVAL were constructed upon the test set of TIMEQA (Chen et al., 2021) and SITUATEDQA (Zhang and Choi, 2021) datasets, which are publicly available under the licenses BSD-3-Clause license[8] and Apache-2.0 license[9]. These licenses all permit us to compose, modify, publish, and distribute additional annotations upon the original dataset. All the experiments in this paper were conducted on 4 NVIDIA L40S 46G GPUs. We hired 3 graduate students in STEM majors as annotators. We recommended that annotators spend at most 2 hours per day for annotation in order to reduce pressure and maintain a comfortable pace. The whole annotation work lasted about 5 days.

Acknowledgements

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). Siyue Zhang and Chen Zhao were supported by NYU Shanghai Center for Data Science. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.

[8]https://opensource.org/licenses/BSD-3-Clause
[9]https://www.apache.org/licenses/LICENSE-2.0

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *2021 Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

Ziyang Chen, Xiang Zhao, Jinzhi Liao, Xinyi Li, and Evangelos Kanoulas. 2022. Temporal knowledge graph question answering via subgraph reasoning. *Know.-Based Syst.*

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*.

Enjun Du, Xunkai Li, Tian Jin, Zhihan Zhang, Rong-Hua Li, and Guoren Wang. 2025a. Graphmaster: Automated graph synthesis via llm agents in data-limited environments. ArXiv preprint arXiv:2504.00711v2.

Enjun Du, Siyi Liu, and Yongqi Zhang. 2025b. Graphoracle: A foundation model for knowledge graph reasoning. ArXiv preprint arXiv:2505.11125.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu

Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma:

Open models based on gemini research and technology.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2024. Complextempqa: A large-scale dataset for complex temporal question answering. *arXiv preprint arXiv:2406.04866*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Sebastian Riedel, and Edouard Grave. 2020. A memory efficient baseline for open domain question answering. *arXiv preprint arXiv:2012.15156*.

Nour Jedidi, Yung-Sung Chuang, Leslie Shing, and James Glass. 2024. Zero-shot dense retrieval with embeddings from relevance feedback.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Jina. 2024. Jina reranker v2 for agentic rag: Ultrafast, multilingual, function-calling and code search. `https://jina.ai/news/jina-reranker-v2-for-agentic-rag-ultra-fast-multilingual-function-calling-and-code-search/`. Accessed: 2024-11-26.

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

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

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2024. Realtime qa: What's the answer right now?

Amanda Kau, Xuzeng He, Aishwarya Nambissan, Aland Astudillo, Hui Yin, and Amir Aryani. 2024. Combining knowledge graphs and large language models.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024b. Can long-context language models subsume retrieval, rag, sql, and more?

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *CoRR*.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. *arXiv preprint arXiv:2205.11388*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

Yonghao Liu, Di Liang, Mengyu Li, Fausto Giunchiglia, Ximing Li, Sirui Wang, Wei Wu, Lan Huang, Xiaoyue Feng, and Renchu Guan. 2023. Local and global: temporal question answering via information fusion. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pageviews. 2024. Pageviews analysis. `https://pageviews.wmcloud.org/`. Accessed: 2024-11-26.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024. Timer4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. 2024. Knowledge editing in language models via adapted direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Aditya Sharma, Apoorv Saxena, Chitrank Gupta, Mehran Kazemi, Partha Talukdar, and Soumen Chakrabarti. 2023. TwiRGCN: Temporally weighted graph convolution for question answering over temporal knowledge graphs. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Ji-Rong Wen, Yang Lu, and Xu Miu. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. `https://github.com/RUCAIBox/R1-searcher`. Accessed: 2025-05-19.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2024a. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024b. Timo: Towards better temporal reasoning for language models. *Proceedings of the Conference on Language Model*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Language Resources and Evaluation Conference*.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024. EasyEdit: An easy-to-use knowledge editing framework for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. 2024a. AKEW: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2024b. Antileak-bench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. *arXiv preprint arXiv:2412.13670*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Sophia Yang. 2023. Advanced rag 01: Small-to-big retrieval. https://towardsdatascience.com/advanced-rag-01-small-to-big-retrieval-172181b396d4. Accessed: 2024-09-15.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Siyue Zhang, Anh Tuan Luu, and Chen Zhao. 2024. SynTQA: Synergistic table-based question answering via mixture of text-to-SQL and E2E TQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025. Diffusion vs. autoregressive language models: A text embedding perspective.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In *Emperical Methods in Natural Language Processing*.

A Controlled Experiments

We conduct controlled experiments to investigate the behaviors of retrieval methods on temporally constrained queries, including the bi-encoder retriever Contriever (Izacard et al., 2021), the cross-encoder reranker MiniLM (Wang et al., 2020), and the LLM embedding-based reranker GEMMA (Li et al., 2023). As shown in Figure 6, all methods prioritize date-matching, having the highest scores when the query and document share the same year. Besides, the score is unusually high when the query and document share the same month and day but differ in year, e.g., orange triangles in the diagrams.

The Contriever retriever is less sensitive to document dates than the MiniLM reranker. Both methods exhibit similar trends across varying temporal relations, indicating their inability to differentiate effectively between different relations, e.g., "`before`" and "`after`". Notably, documents without specific dates receive unusually low scores, even lower than those with irrelevant dates, e.g., orange dash lines in the diagrams.

The LLM embedding based reranker Gemma exhibits stronger temporal reasoning capabilities. For the "`after`" relation, documents with dates later than the query date are assigned relatively high and consistent scores. So all temporally relevant documents will be retained. However, for "`before`" and "`as of`", despite their temporal relevance, documents with earlier dates fail to achieve sufficiently high similarity scores, potentially leading to their exclusion from the retrieval process.

In summary, existing retrieval methods demonstrate limited temporal reasoning capabilities. The LLM embedding-based method shows better performance than others. Our proposed MRAG framework is retriever-agnostic, which aims to improve temporal reasoning capabilities for any type of retrieval models.

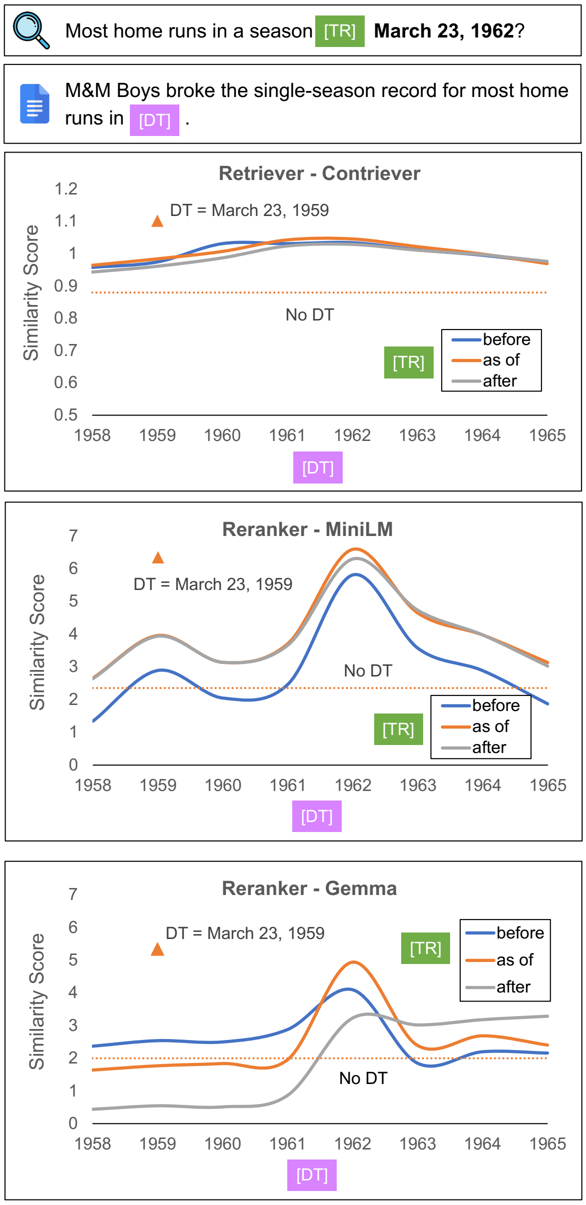


Figure 6: Similarity scores of query-document pairs by varying the temporal relation in the query and the date in the document.

B Dataset Selection Criteria

Our benchmark focuses on time-sensitive question answering, which is knowledge-intensive. Therefore, datasets designed for only temporal reasoning (e.g., "What is the time 5 year and 5 month after Oct, 1444") are not considered, such as TempReason (Tan et al., 2023) and DurationQA (Virgo et al., 2022). Aggregated benchmarks (*e.g.,* TimeBench (Chu et al., 2024) and TRAM (Wang and Zhao, 2024)) focus on evaluating diverse temporal reasoning capabilities, which have a broader scope than our focus. MenatQA (Wei et al., 2023) is built by adding counterfactual and order factors to TimeQA (Chen et al., 2021) questions, which is similar to our approach. Other knowledge-intensive temporal QA datasets can serve as alternative sample sources including StreamingQA (Liška et al., 2022), TempLAMA (Dhingra et al., 2022), and concurrent dataset ComplexTQA (Gruber et al., 2024). We select SITUATEDQA (Zhang and Choi, 2021) for its human-written questions, distinguishing it from other temporal QA datasets that are typically synthetic. Additionally, we opt for TIMEQA (Chen et al., 2021) due to its hard split, which already includes complex temporal questions. Notably, both SITUATEDQA and TIMEQA can be grounded in the Wikipedia corpus.

C Annotation Guidelines

C.1 Annotating Perturbations

Given a question-answer pair sourced from TIMEQA or SITUATEDQA (*e.g.,* Q: "Arnolfini Portrait was owned by whom between Jul 1842 and Nov 1842?" A: "National Gallery"), annotators should ground the pair to facts in Wikipedia (*e.g.,* "The Arnolfini Wedding by Jan van Eyck, has been part of the National Gallery's collection in London since 1842."). Then they identify the key timestamps or durations of Wikipedia facts (*e.g.,* 1842). To create temporal perturbations, annotators are asked to come up with combinations of implicit conditions, temporal relations, and alternative dates to form complex temporal constraints. The implicit condition can be "`None`" or selected from a list of 4 types: "`first`", "`earliest`", "`last`", and "`latest`". The temporal relation should be selected from a list of 11 types: "`as of`", "`from to`", "`until`", "`before`", "`after`", "`around`", "`between`", "`by`", "`in`", "`on`", and "`since`". Finally, annotators rewrite questions naturally (*e.g.,* "Who is the last one owned Arnolfini Portrait after 1700?") by introducing perturbed temporal constraints (*e.g.,* "last ... after 1700") and ensure that the answers (*e.g.,* "National Gallery") remain unchanged. After different annotators create perturbed question-answer pairs, they exchange these pairs with each other to validate the correctness of the answers. Only the perturbed samples validated by two annotators are kept.

C.2 Annotating Gold Evidences

For gold evidence annotations, annotators are assigned different perturbed question-answer pairs. For each pair, 20 context passages are provided to annotators, which are retrieved by the leading retriever Contriever (Izacard et al., 2021), and the best reranker GEMMA (Li et al., 2023). Annotators are asked to identify up to two gold evidence passages from these passages. A passage is regarded as relevant and annotated as gold evidence if annotators can obtain the correct answer from this passage. If there is no relevant passage among these 20 retrieved ones, annotators should search Wikipedia pages related to the query entities to locate the gold evidence passages manually. Lastly, annotators exchange samples to validate gold evidence annotated by others. Only the gold evidence annotations validated by two annotators are kept.

D Sample Statistics

We gather a similar size of examples as previous temporal QA benchmarks (e.g., 3K for TimeQA (Chen et al., 2021) and 2K for MenatQA (Wei et al., 2023)), which is enough for an evaluation set (see examples in Appendix Q). As we manually annotate gold evidence passages in Wikipedia, it is time-consuming to scale up like other synthetic datasets (Dhingra et al., 2022; Gruber et al., 2024). To understand the difference between two subsets, we summarize the statistics in Appendix D. The average length of questions is measured by the GPT-2 tokenizer (Radford et al., 2019). We assess the popularity of key entities in questions using the average monthly page view counts of the corresponding Wikipedia page in 2024 (Pageviews, 2024). As we can see, the main difference lies in the question entity popularity. TEMPRAGEVAL-TIMEQA questions typically inquire about lesser-known individuals and are generally straightforward and clear. In contrast, TEMPRAGEVAL-SITUATEDQA questions commonly ask about a sports team and championship, which are more general and sometimes ambiguous. This difference may explain varying retrieval and QA performance across the two subsets.

	TimeQA	**SituatedQA**
# original questions	123	120
# perturbed questions	377	380
# total questions	500	500
Temporal complexity	hard	hard
Avg. question length	15.2	15.6
Avg. entity popularity	7,456	57,521

Table 5: Sample statistics for TEMPRAGEVAL-TIMEQA and TEMPRAGEVAL-SITUATEDQA.

E Retrieval Performance Degradation Due to Perturbations

As shown in Figure 7, we compare the retrieval performance between original queries and perturbed queries using the same baseline retrieval system (*i.e.,* Contriever retriever and GEMMA reranker). For both the TIMEQA and SITUATEDQA subsets, the perturbed questions significantly increase the

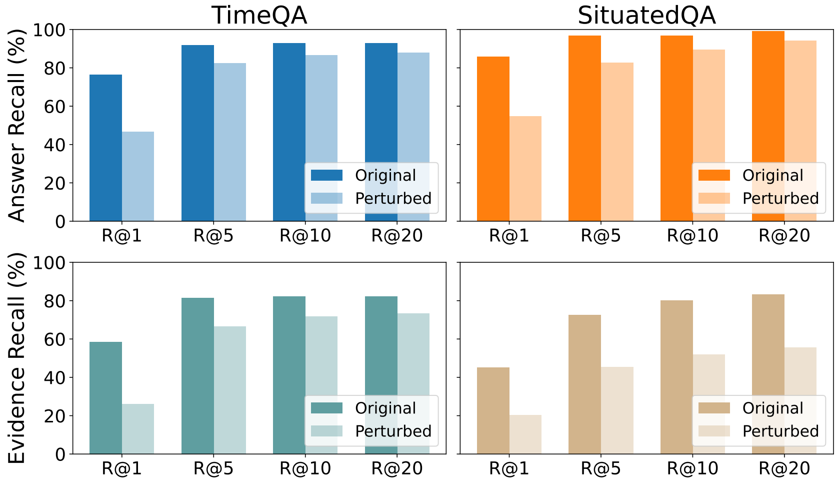


Figure 7: Retrieval performance difference between original queries and perturbed queries in TEMPRAGEVAL subsets for the baseline GEMMA retrieval.

difficulty of retrieving relevant documents, particularly when evaluating the top-1 and top-5 ranked documents. This suggests that the introduction of perturbations introduces greater complexity. The existing retrieval method has limited temporal reasoning capabilities and is not robust to such variations.

F Evaluation Experiment Implementation Details

We conduct empirical evaluations for MRAG and SOTA retrieving-and-reranking systems on TEMPRAGEVAL. In baselines, due to limited computing resources, we use LLM-based embedding models as a reranking model, such as GEMMA (Gemma et al., 2024) and NV-Embed (Lee et al., 2024a). MRAG consists of functional modules, which can be based on algorithms, models, or prompting methods. In implementation, algorithm based modules include question normalization, keyword ranking, time extraction, and semantic-temporal hybrid ranking. Model based modules are retrieving, semantic ranking, sentence tokenization. To ensure fair comparison, we use the same retriever model, i.e., Contriever (Izacard et al., 2021), as the first stage method for MRAG and two-stage systems. We use GEMMA embeddings (Li et al., 2023) as the main tool to measure semantic similarity for passages and sentences in MRAG. NLTK package is used for sentence tokenization (Bird and Loper, 2004). LLM prompting based modules are keyword extraction, query-focused summarization. As shown in Appendix I, we have tested Llama3.1-8B-Instruct and Llama3.1-70B-Instruct models (Dubey et al., 2024) for LLM prompting based modules. Detailed prompts are listed in Appendix R. The evaluation metrics are computed based on retrieved passages not sentences.

G Implementations for Semantic-Temporal Hybrid Scoring

The **Retrieval and Summarization** module in MRAG splits and summarizes top relevant passages into independent sentences for the downstream fine-grained reranking, which is inspired by Yang (2023). The **Semantic-Temporal Hybrid Scoring** module is designed to assess the semantic relevance and temporal relevance between the question and the evidence sentence. To quantify the semantic relevance, we apply an embedding model (*e.g.,* GEMMA) to the question main content and the sentence.

For temporal relevance, we employ a symbolic scoring approach, wherein the module automatically generates scoring functions for each question and computes temporal scores for individual sentences.

To generate scoring functions, we classify question temporal constraints into six categories and define a template for each. A scoring function is instantiated using the corresponding template and extracted timestamp(s) from the question. The six constraint types include: "`first - before`", "`first - after`", "`first - between`", "`last - before`", "`last - after`", and "`last - between`". Here, "`last`" denotes that the question seeks the most recent event, while "`before`" indicates that the event must precede a specified

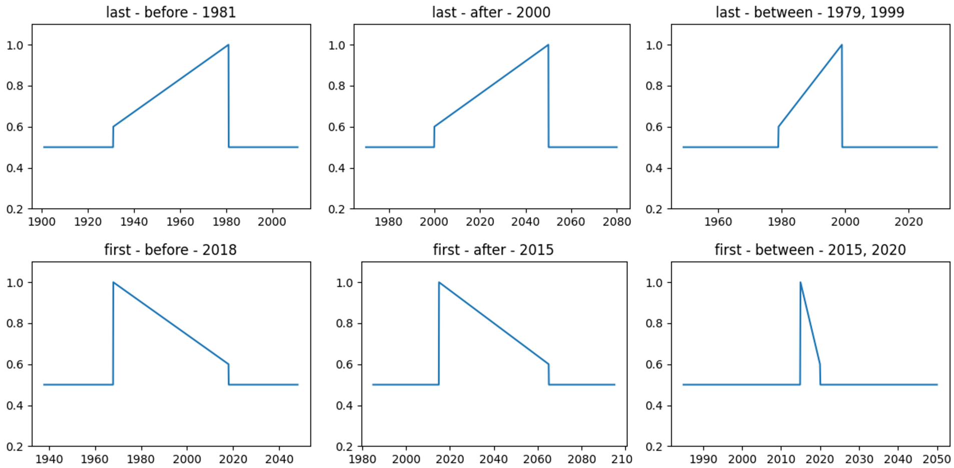


Figure 8: Pre-defined spline functions for temporal relevance scoring. The title of each subplot represents the type of query temporal constraint. The horizontal coordinate of each subplot is the date in the document sentence.

date. For instance, the question "Who won the latest game as of 1981?" corresponds to "`last - before - 1981`", as illustrated in the top-left subplot of Figure 8.

Afterwards, the module extracts the timestamp(s) from the evidence sentence. It computes the temporal score based on the extracted timestamp and the corresponding scoring function. If multiple timestamps are present, the highest temporal score is selected. For example, as depicted in Figure 8, for the constraint "`last - before - 1981`", an evidence sentence mentioning "`1970`" would receive a temporal score around 0.9.

The final score of each evidence sentence for each question is obtained by multiplying the temporal score and the semantic score. Finally, we select the passages that contains the highest-scoring sentences. The passages are fed into the later answer generation stage rather than the sentences. The passages provide better background information, which leads to higher generation quality for reader systems.

H LLM-based Summarization Case Study

LLM-based query-focused summarization enhances retrieval performance by distilling key information from passages while filtering out irrelevant context, as demonstrated in Table 6. In the second success case, the summarization effectively converts structured data into natural language, benefiting retrievers that are primarily trained on free-text retrieval tasks. However, LLM-generated summaries may introduce hallucinations and errors, though their occurrence is infrequent. As shown in Table 6, erroneous summaries can mislead the retriever with non-factual events or incorrect dates, resulting in irrelevant passages ranking higher. To balance retrieval improvements with the risk of errors, we summarize only the top-k passages per query, which also reduces computational overhead. Furthermore, to prevent error propagation, we provide the reader model with original passages rather than their summaries.

	Success Cases
Question	Who won the latest America's Next Top Model by May 8, 2021?
Answer	Kyla Coleman
Passage	America's Next Top Model (season 24) | The twenty-fourth cycle of America's Next Top Model premiered on January 9, 2018 ... The winner of the competition was 20 year-old Kyla Coleman from Lacey, Washington with Jeana Turner placing as the runner up.
Summarization	Kyla Coleman, a 20-year-old from Lacey, Washington, won the competition in 2018.
Question	When was the last time Kentucky won NCAA in basketball after 2010?
Answer	2012
Passage	Kentucky Wildcats | Men (8) ; Basketball (8): 1948, 1949, 1951, 1958, 1978, 1996, 1998, 2012 ; Women (2) ... List of NCAA schools with the most NCAA Division ... Kentucky has won 13 NCAA team national championships.
Summarization	The Kentucky Wildcats won the NCAA basketball championship in 1948, 1949, 1951, 1958, 1978, 1996, 1998, and 2012.
	Failure Cases
Question	When was the last time the Ducks won the Stanley Cup as of 2010?
Answer	2007
Passage	Anaheim Ducks | ... Despite the arenas being six hours away from each other, the teams have developed a strong rivalry, primarily from the 2009 and 2018 Stanley Cup playoffs. The Ducks won the series in 2009, but the Sharks came back in 2018.
Summarization	The Anaheim Ducks won the Stanley Cup in 2009.
True fact	The Anaheim Ducks won the Stanley Cup in 2007. That was their first and only championship so far.
Question	How many times has South Korea held the Winter Olympics as of 2018?
Answer	1 | one
Passage	2018 Winter Olympics | The 2018 Winter Olympics ... This marked the second time that South Korea had hosted the Olympic Games (having previously hosted the 1988 Summer Olympics in Seoul) ...
Summarization	South Korea held the Winter Olympics in 2018 and previously in 1988.
True fact	South Korea held the Winter Olympics in 2018 and Summer Olympics in 1988.

Table 6: Success and error cases of LLM-based query-focused summarization using Llama3.1-8B-Instruct.

I Complete Retrieval Evaluation Results

We evaluate MRAG on TEMPRAGEVAL with baseline retrieval methods, including ELECTRA[10], MiniLM[11], Jina[12], BGE[13], NV-Embed[14], and GEMMA[15]. Complete rsults are presented in Table 7 and Table 8.

J Human Evaluation of Retrieval

The Answer Recall (AR@k) represents the upper bound of the retrieval performance, while the Evidence Recall (ER@k) signifies the lower bound. As shown in Figure 9 and Figure 10, the gray areas are delineated by the AR@k and ER@k lines, within which the actual performance remains uncertain. To address this, we conduct a human evaluation of ranked document passages retrieved by MRAG and GEMMA (denoted as “Standard” in the figures) on a subset of 200 randomly selected examples from TEMPRAGEVAL.

The metric for the actual retrieval performance, termed **Ground Truth Recall (GR@k)**, is computed based on the annotations of the highest-ranking passages supporting the answers. As illustrated, the gray areas for MRAG are positioned higher in the plots than those for GEMMA. Furthermore, the actual performance curves (purple lines) for MRAG are consistently closer to the upper boundaries compared to those for GEMMA (green lines). These two observations demonstrate the superior performance of MRAG in temporal reasoning-intensive retrieval.

Figure 9: Human annotated retrieval performance on 100 examples from TEMPRAGEVAL-TIMEQA.

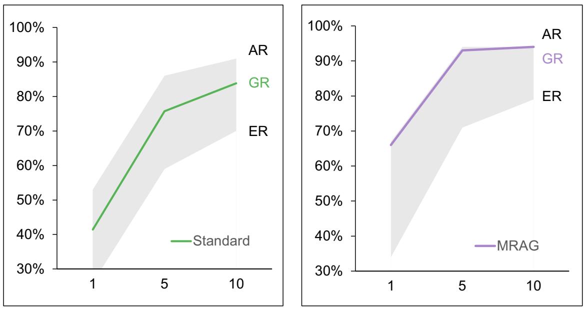


Figure 10: Human annotated retrieval performance on 100 examples from TEMPRAGEVAL-SITUATEDQA.

[10] cross-encoder/ms-marco-electra-base
[11] cross-encoder/ms-marco-MiniLM-L-12-v2
[12] jinaai/jina-reranker-v2-base-multilingual
[13] BAAI/bge-reranker-large
[14] nvidia/NV-Embed-v1
[15] BAAI/bge-reranker-v2-gemma

Method				Answer Recall @				Gold Evidence Recall @			
1st	**2nd**	**LLM**	**# QFS**	**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
BM25	-	-	-	17.5	39.0	49.1	59.0	4.2	14.1	22.6	33.7
Cont.	-	-	-	18.8	49.9	62.1	72.9	9.6	28.7	39.5	51.5
Hybrid	-	-	-	18.8	51.2	65.0	75.3	9.6	28.1	41.1	55.2
Cont.	ELECTRA	-	-	40.1	76.9	83.6	86.7	21.8	58.6	66.8	71.6
Cont.	MiniLM	-	-	34.0	76.1	84.4	87.0	16.2	57.3	68.2	72.4
Cont.	Jina	-	-	42.4	77.2	86.2	87.5	23.6	58.6	68.2	71.4
Cont.	BGE	-	-	40.3	80.9	85.7	87.0	23.3	61.3	68.7	72.2
Cont.	NV-Embed	-	-	49.9	81.2	85.7	87.5	33.4	62.9	70.6	72.7
Cont.	Gemma	-	-	46.7	82.5	86.5	87.8	26.0	66.6	71.6	73.2
Cont.	MRAG	Llama3.1	-	57.6	89.4	93.6	94.2	32.4	73.5	82.8	84.1
Cont.	MRAG	Llama3.1	5	58.6	90.0	93.4	94.2	37.1	74.3	82.5	84.1
Cont.	MRAG	Llama3.1	10	56.0	88.1	93.6	94.2	35.5	73.2	82.2	84.4
Cont.	MRAG	Llama3.1$^\flat$	-	57.6	89.4	93.6	94.2	32.1	73.5	82.5	84.1
Cont.	MRAG	Llama3.1$^\flat$	5	57.0	90.5	93.6	94.2	34.5	75.3	82.5	84.1
Cont.	MRAG	Llama3.1$^\flat$	10	53.3	90.7	93.6	94.2	33.2	74.8	82.8	84.1

Table 7: The answer recall (AR@k) and gold evidence recall (ER@k) performance of each retrieval system on perturbed temporal queries in TEMPRAGEVAL − TIMEQA subset. $^\flat$Meta-Llama-3.1-70B-Instruct.

Method				Answer Recall @				Gold Evidence Recall @			
1st	**2nd**	**LLM**	**# QFS**	**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
BM25	-	-	-	27.6	58.2	69.0	80.8	6.8	18.4	25.8	34.7
Cont.	-	-	-	22.6	51.1	65.5	79.5	6.8	17.1	22.9	30.5
Hybrid	-	-	-	22.6	55.8	71.8	81.6	6.8	19.7	26.6	35.0
Cont.	ELECTRA	-	-	35.5	71.3	82.4	88.4	15.3	37.1	45.0	52.9
Cont.	MiniLM	-	-	36.8	73.4	86.3	90.8	20.0	40.3	50.5	54.2
Cont.	Jina	-	-	47.9	78.4	87.6	93.2	19.5	41.1	48.2	54.2
Cont.	BGE	-	-	36.3	74.2	86.3	92.9	14.5	35.0	44.7	54.2
Cont.	NV-Embed	-	-	47.4	81.3	88.7	92.4	23.4	46.1	50.5	55.0
Cont.	Gemma	-	-	54.7	82.6	89.5	94.0	20.3	45.3	51.8	55.5
Cont.	MRAG	Llama3.1	-	61.1	88.2	92.1	93.7	27.4	56.3	64.0	68.7
Cont.	MRAG	Llama3.1	5	61.3	89.0	93.4	94.2	31.1	59.2	65.8	69.0
Cont.	MRAG	Llama3.1	10	62.1	87.9	92.6	94.2	30.8	57.9	66.1	70.3
Cont.	MRAG	Llama3.1$^\flat$	-	61.1	86.3	92.4	94.0	27.1	54.5	63.4	67.6
Cont.	MRAG	Llama3.1$^\flat$	5	63.2	86.1	92.6	93.7	29.7	56.6	64.0	67.1
Cont.	MRAG	Llama3.1$^\flat$	10	62.1	86.8	92.1	93.7	27.1	56.1	64.2	69.0

Table 8: The answer recall (AR@k) and gold evidence recall (ER@k) performance of each retrieval system on perturbed temporal queries in TEMPRAGEVAL − SITUATEDQA subset. $^\flat$Meta-Llama-3.1-70B-Instruct.

K Parametric Study on the Optimal Number of Passages for Concatenation

The number of concatenated passages and their order significantly impact the accuracy of reader QA tasks. This is largely due to the inherent primacy and recency biases exhibited by LLMs, where information presented earlier or later in the input sequence tends to be weighted more heavily during processing (Liu et al., 2024). Therefore, the retrieval performance is of great importance.

We evaluate the Llama reader accuracy with a varying number of concatenated documents retrieved by GEMMA and MRAG in Figure 4. The rapid accuracy improvement within the first five passages highlights the effectiveness of RAG in enhancing LLMs' performance by supplementing their knowledge with external information. In both TEMPRAGEVAL subsets, the reader demonstrates higher accuracy with MRAG-retrieved documents in most cases. Notably, MRAG achieves peak accuracy with only the top 5 retrieved documents, whereas GEMMA might require more, as illustrated in Figure 4b. For Llama3.1-8B, using 5 documents is optimal, as including more passages in the input may introduce noise and distractors, leading to errors made by the reader.

L RAG vs. Long-Document QA

TEMPRAGEVAL-TIMEQA is derived from TIMEQA, a dataset originally designed for long-document QA. TIMEQA questions are constructed using Wikipedia pages as evidence, ensuring answer presence in the source. The page name is explicitly included in each question (Chen et al., 2021). We compare retrieval-augmented generation (RAG) with GEMMA and MRAG retrievers against the long-document QA setup (without retrieval) using two Llama3.1 models (Table 9). In the long-document QA setup, the entire Wikipedia page is provided as context, leading to longer inputs with numerous distractors. Our results show that RAG, when using a high-quality retriever, outperforms long-document QA, validating our hypothesis that supplying full Wikipedia pages introduces noise that degrades performance. Although these LLMs have strong long-context reasoning capabilities (Dubey et al., 2024), they can still be misled by irrelevant passages. The RAG approach mitigates this issue by limiting input passages and excluding irrelevant passages, thereby improving QA accuracy.

M Computational Overhead Assessment

As MRAG involves multiple processing steps, including retrieval, re-ranking, summarization, and hybrid ranking, which could introduce computational overhead compared to standard RAG pipelines. To assess real-world scalability, we assess the average processing (inference) time in seconds per query in comparison to baseline retrieval methods (*i.e.,* MiniLM and GEMMA). Given retrieved passages by Contriever for each query, these methods rerank the top-100 passages. The processing time of MRAG is further broken down by each module in Table 10. The assessment is conducted on a machine with one NVIDIA L40S 46G GPU and one AMD EPYC 9554 64-Core CPU. MRAG is implemented using GEMMA for pure semantic scoring. The inference time can be significantly reduced by using MiniLM. Compared to GEMMA, MRAG incurs approximately twice the runtime overhead.

N Ablation Study on MRAG Modules

In the retrieval and summarization module, with retrieved passages, MRAG conducts two key steps: (1) *Passage Keyword Ranking* reduces the passages from 1,000 to 100 based on the keyword presence; (2) *Passage Semantic Ranking* reorders the top-100 passages and splits them into $\sim$500 sentences, including chunk summaries. Similarly, in the semantic-temporal hybrid ranking module, another two key steps are conducted: (1) *Sentence Keyword Ranking* further narrows the scope to 200 sentences; (2) *Hybrid Ranking* ranks these sentences by semantic-temporal hybrid scoring.

To validate the effectiveness of each step in each module, we conduct the ablation study. As shown in Table 4, keyword-based ranking steps effectively reduce ranking candidates without significant performance loss, while semantic and hybrid ranking steps markedly enhance retrieval performance. An efficiency-accuracy trade-off can be made by adjusting the number of targeted passages or sentences at each step.

Method	TEMPRAGEVAL-TimeQA		
	# Docs	EM	F1
Llama3.1-8B-Instruct			
Direct Prompt	-	16.0	23.9
Direct CoT	-	16.8	27.8
RAG-Concat	5	44.0	52.8
MRAG-Concat	5	**49.2**	**59.2**
Long-Doc QA	16.4$^\sharp$	<u>45.2</u>	<u>54.9</u>
Llama3.1-70B-Instruct			
Direct Prompt	-	31.0	
Direct CoT	-	33.2	45.8
RAG-Concat	5	<u>54.4</u>	<u>63.2</u>
MRAG-Concat	5	**58.0**	**68.4**
Long-Doc QA	16.4$^\sharp$	48.1	59.0

Table 9: End-to-end QA performance comparison for RAG and long-document QA setups. $^\sharp$The average number of passages in Wikipedia pages corresponding to TEMPRAGEVAL-TIMEQA questions.

	Question Processing	Retrieval & Summarization	Temporal-Semantic Hybrid Ranking	Total
MiniLM	-	-	-	0.14
GEMMA	-	-	-	1.03
MRAG	0.06	1.23	1.04	2.33

Table 10: Latency assessment (in seconds) for MRAG and baseline retrieval methods.

O Case Studies

O.1 Retrieval Failure Case Study

We conducted five case studies to qualitatively evaluate the advantages of MRAG over the GEMMA retriever as below. The results demonstrate MRAG's robustness to temporal perturbations and its ability to retrieve relevant context passages. For instance, in Case 1, the top-1 passage retrieved by GEMMA matches the query date "1988" but discusses a father-son record set in 2007. In contrast, the first passage retrieved by MRAG focuses on a teammate combination record in the same season, despite the date "1961" differing from the query date "1988". Since semantic relevance outweighs strict date matching in this situation, MRAG provides more contextually appropriate results for the time-sensitive question.

Question	**Gemma-based Retrieval**	**Modular Retrieval**
(1) Who had the most home runs by two teammates in a season as of 1988?	#7 is the top true evidence	#1 is the top true evidence
	#1 Bobby Bonds | ... until José Canseco of the Oakland Athletics in 1988. Barry and Bobby had 1,094 combined home runs through 2007 — a record for a father-son combination. #2 1987 in baseball | ... With teammate Howard Johnson already having joined, it marks the first time that two teammates achieve 30–30 seasons in the same year. #3 1988 Toronto Blue Jays season | April 4, 1988: George Bell set a major league record for the most home runs hit on Opening Day, with three ... #7 50 home run club | **M&M Boys**—are the only teammates to reach the 50 home run club in the same season, hitting a combined 115 home runs in 1961 and breaking the single-season record for home runs by a pair of teammates.	#1 50 home run club | **M&M Boys**—are the only teammates to reach the 50 home run club in the same season, hitting a combined 115 home runs in 1961 and breaking the single-season record for home runs by a pair of teammates. #2 1987 Major League Baseball season | Cal Ripken, Jr. is lifted from the lineup and replaced by Ron Washington ... it marks the first time that two teammates achieve 30–30 seasons in the same year. #3 1987 in baseball | Whitt connects on three of the home runs ... it marks the first time that two teammates achieve 30–30 seasons in the same year.

Question	Gemma-based Retrieval	Modular Retrieval
(2) Who had the most home runs by two teammates in a season by August 17, 1992?	No true evidence retrieved	#1 is the top true evidence
	#1 1992 in baseball | ... August 28 – The Milwaukee Brewers lash 31 hits in a 22-2 drubbing of the Toronto Blue Jays , setting a record for the most hits by a team in a single nine-inning game. #2 1997 in baseball | ... McGwire, who hit a major league-leading 52 homers for the Oakland Athletics last season, becomes the first player with back-to-back 50-homer seasons since Ruth did it ...	#1 50 home run club | **M&M Boys**—are the only teammates to reach the 50 home run club in the same season, hitting a combined 115 home runs in 1961 and breaking the single-season record for home runs by a pair of teammates. #2 List of career achievements by Babe Ruth | 1927 (Ruth 60, Lou Gehrig 47) ... Achieved by several other pairs of teammates since ... Two teammates with 40 or more home runs, season: Thrice Clubs with three consecutive home runs in inning ...
(3) Who won the latest America's Next Top Model as of 2021?	No true evidence retrieved	#3 is the top true evidence
	#1 America's Next Top Model (season 17) | the final season for Andre Leon Talley as a judge. The winner of the competition was 30-year-old Lisa D'Amato from Los Angeles, California, who originally placed sixth on Cycle 5 making her the oldest winner at the age of 30. Allison Harvard, who originally placed second on cycle 12 ... #2 Germany's Next Topmodel | that Soulin Omar who was the second runner up, should've won based on her performance throughout the season. German Magazine "OK!" and "Der Westen" stated ... #3 America's Next Top Model (season 21) | (Ages stated are at start of contest) Indicates that the contestant died after filming ended	#1 America's Next Top Model (season 23) | The twenty-third cycle of America's Next Top Model premiered on December 12, 2016 ... The winner of the competition was 20 year-old India Gants from Seattle ... #2 America's Next Top Model | ... five contestants were featured modeling Oscar gowns: ... On May 12, 2010, Angelea Preston, Jessica Serfaty, and Simone Lewis (all cycle 14) appeared on a Jay Walking ... On February 24, 2012, Brittany Brower (cycle 4), Bre Scullark (cycle 5) (both cycle 17), and Lisa D'Amato (cycle 5 and cycle 17 winner) appeared on a Jay #3 America's Next Top Model (season 24) | The twenty-fourth cycle of America's Next Top Model premiered on January 9, 2018 ... The winner of the competition was 20 year-old **Kyla Coleman** from Lacey, Washington ...

Question	Gemma-based Retrieval	Modular Retrieval
(4) When did Dwight Howard play for Los Angeles Lakers between 2000 and 2017?	#5 is the top true evidence	#1 is the top true evidence
	#1 List of career achievements by Dwight Howard | Defensive rebounds, 5-game series: 58, Orlando Magic vs. Los Angeles Lakers, 2009 #2 Dwight Howard | wanted". In a 2013 article titled "Is Dwight Howard the NBA's Worst Teammate?" ... When he was traded from the Atlanta Hawks to the Charlotte Hornets, some of his Hawks teammates reportedly cheered. After Charlotte traded Howard to the Washington Wizards, Charlotte player Brendan Haywood asserted ... #5 **2012-13** Los Angeles Lakers season | In a March 12, 2013 game against his former team, the Orlando Magic, Dwight Howard tied his own NBA record of 39 free throw attempts ...	#1 Dwight Howard | On August 10, **2012**, Howard was traded from Orlando to the Los Angeles Lakers in a deal that also involved the Philadelphia 76ers and the Denver Nuggets ... #2 Dwight Howard | ... In **2012**, after eight seasons with Orlando, Howard was traded to the Los Angeles Lakers ... Howard returned to the Lakers in 2019 and won his first NBA championship in 2020.

Question	Gemma-based Retrieval	Modular Retrieval
(5) Which political party did Clive Palmer belong to on Apr 20, 1976?	No true evidence retrieved	#3 is the top true evidence
	#1 Clive Palmer | ... On 25 April 2013, Palmer announced a "reformation" of the United Australia Party, which had been folded into the present-day Liberal Party in 1945, to stand candidates in the 2013 federal election, and had applied for its registration in Queensland ... #2 Clive Palmer | de-registering the party on 5 May 2017, Palmer revived his party as the United Australia Party, announcing that he would be running candidates for all 151 seats in the House of Representatives and later that he would run as a Queensland candidate for the Senate. In the 2019 federal election, despite extensive advertising ... #3 United Australia Party (2013) | Clive Palmer of bullying, swearing and yelling at people. Lazarus stated "I have a different view of team work. Given this, I felt it best that I resign from the party and pursue my senate role as an independent senator." ...	#1 Clive Palmer | Palmer deregistered the party's state branches in September 2016, initially intending to keep it active at the federal level. However, in April 2017, he announced that the party would be wound up. In February 2018, Palmer announced his intention to resurrect his party and return to federal politics. The party was revived in June under its original name, the United Australia Party ... #2 Clive Palmer | ... Palmer resigned his life membership of the Liberal National Party. His membership of the party had been suspended on 9 November 2012, following his comments on the actions of state government ministers. He was re-instated to the party on 22 November, but resigned the same day ... #3 Clive Palmer | Palmer was instrumental in the split of the South Australian conservatives in the 1970s, and was active in the Liberal Movement headed by former Premier of South Australia, Steele Hall. Palmer joined the **Queensland division of the Nationals** in 1974 ...

O.2 Downstream QA Failure Case Study

To identify the most error-prone component (retrieval or generation), we manually analyze 50 random failure cases for GEMMA and another 50 for MRAG from TEMPRAGEVAL-TIMEQA. The same analysis is applied to TEMPRAGEVAL-SITUATEDQA, focusing on the RAG-Concat and MRAG-Concat methods using the Llama3.1-8B-Instruct model. We categorize each failure by root cause: retrieval, format, or reader. Failures are attributed to the retriever when it fails to find at least one relevant passage within the top-5 retrieved results. In cases where the reader model receives relevant passages, errors are classified as format if the generated answer is correct but in a different format, or as reader errors if the model fails to perform temporal reasoning correctly, despite having access to relevant knowledge. Our analysis, shown in Figure 11, reveals that the majority of errors stem from the reader, indicating that both retrievers perform well. Compared to GEMMA, MRAG exhibits a lower percentage of retrieval errors, *e.g.*, Figure 11(b) vs. Figure 11(a), demonstrating the effectiveness of our proposed retrieval approach.

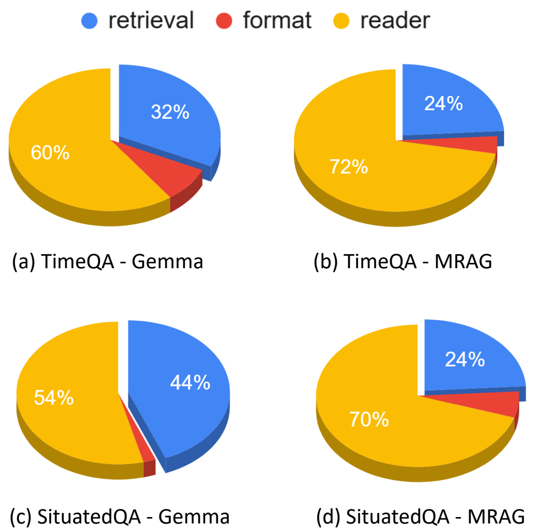


Figure 11: Percentage distribution of error case root causes on TEMPRAGEVAL-TIMEQA and TEMPRAGEVAL-SITUATEDQA. Gemma refers to RAG-Concat and MRAG refers to MRAG-Concat.

P Knowledge Conflicts Between Parametric Knowledge and External Passages

RAG systems commonly confront knowledge conflicts either between LLM internal knowledge and external passage knowledge or across different passages (Jin et al., 2024). We illustrate two categories of examples where the LLM internal knowledge conflicts with the retrieved passage, using the GEMMA retriever and the Llama3.1-8B-Instruct model. RAG systems typically prioritize external knowledge in the retrieved context. Therefore, we observe a significant amount of errors in parametric knowledge are avoided by providing relevant passages as examples in the top of Table 11. Besides, when only irrelevant passages are retrieved, the correct parametric knowledge can be misled by the the distracting context as examples in the bottom of Table 11. Thus, the retriever performance is of great importance.

External passages may have conflicting knowledge, requiring LLMs to make nuanced judgments in such cases (Pham et al., 2024). In our experiments on time-sensitive question answering using the Wikipedia corpus, we rarely observe conflicting passages within Wikipedia. Different answers typically correspond to different temporal constraints. To gain a deeper understanding of conflicting passages, one approach would be to introduce counterfactual documents. However, as our research focuses on temporal reasoning in retrieval, we leave this direction for future work.

Wrong parametric predictions with relevant passages.			
Query	**Relevant passage**	**CoT pred.**	**RAG pred.**
Who owned the Newton D. Baker House in Washington DC from 1978 to 1982?	Newton D. Baker House | … Straight and his wife lived in the home from until 1976. In 1976, Yolande Bebeze Fox, the former Miss America 1951, bought the home from Straight. Fox lived in the home until her death in February 2016.	American Enterprise Institute	Yolande Bebeze Fox
What was the last position of Homer Thornberry between 1941 to 1943?	Homer Thornberry | Thornberry was born in Austin, Texas … He was district attorney of Travis County, Texas from 1941 to 1942. He was a United States Navy Lieutenant Commander from 1942 to 1946 …	United States Senator from Texas	United States Navy Lieutenant Commander
Who was the chair of National Council of French Women in Dec 1951?	National Council of French Women | … Marguerite Pichon-Landry (1878–1972) chaired the Legislation section of the CNFF from 1914 to 1927, and was secretary-general from 1929 to 1932. She was president from 1932 to 1952 …	Éliane Brault	Marguerite Pichon-Landry
Warlugulong was owned by whom in 1997?	Warlugulong | … the work was sold by art dealer Hank Ebes on 24 July 2007, setting a record price for a contemporary Indigenous Australian art work bought at auction when it was purchased by the National Gallery of Australia for A$2.4 million.	the Pritzker family	Hank Ebes
Correct parametric predictions with irrelevant (distracting) passages.			
Query	**Irrelevant passage**	**CoT pred.**	**RAG pred.**
What was the first U-boat unit Erich Topp commanded between 5 October 1937 and December 1941?	Erich Topp | World War II commenced following the German invasion of Poland on 1 September 1939. U-46, under the command of Sohler, had already been at sea since 19 August, returning to port on 15 September.	1st U-boat Flotilla	U-46
Who is the first one owned Arnolfini Portrait after 1900?	Arnolfini Portrait | The Arnolfini Portrait (or The Arnolfini Wedding, The Arnolfini Marriage, the Portrait of Giovanni Arnolfini and his Wife, or other titles) is a 1434 oil painting on oak panel by the Early Netherlandish painter Jan van Eyck …	The National Gallery	There is no information about who owned the Arnolfini Portrait after 1900 in the given context.

Table 11: Examples of LLM parametric knowledge and retrieved passages.

Q TEMPRAGEVAL Examples

	TEMPRAGEVAL-SITUATEDQA
Question	When did Dwight Howard play for Los Angeles Lakers between 2000 and 2017?
Answer	2012 | 2013 | 2012-2013
Gold evidence	Dwight Howard | ... On August 10, 2012, Howard was traded from Orlando to the Los Angeles Lakers in a deal that also involved the Philadelphia 76ers and the Denver Nuggets ...
Question	When was the earliest time Dwight Howard play for the Lakers after August 10, 2014?
Answer	2019 | 2020 | 2019-2020
Gold evidence	Dwight Howard | ... On August 26, 2019, Howard signed a $2.6 million veteran's minimum contract with the Los Angeles Lakers, reuniting him with his former team ...
Question	When did the last season on The 100 come out between 2018 and 2021?
Answer	May 20, 2020 | 2020
Gold evidence (1)	The 100 (TV series) | ... The CW renewed the series for a seventh season, that would consist of 16 episodes and premiered on May 20, 2020 ...
Gold evidence (2)	The 100 season 7 | ... On March 4, 2020, it was revealed that the last season of The 100 would premiere on The CW on May 20, 2020 ...
Question	Who was the leader of the Ontario PC Party after 2020?
Answer	Doug Ford
Gold evidence (1)	Progressive Conservative Party of Ontario | ... On March 10, 2018, Doug Ford, former Toronto city councillor ... was elected as leader of the PC Party ...
Gold evidence (2)	New Blue Party of Ontario | ... on March 10, 2018, Doug Ford was elected as leader of the Progressive Conservative Party of Ontario ...
	TEMPRAGEVAL-TIMEQA
Question	Oliver Bulleid was an employee for whom as of Oct 1905?
Answer	Great Northern Railway
Gold evidence	Oliver Bulleid | ... In 1901 ... he joined the Great Northern Railway (GNR) at Doncaster at the age of 18, as an apprentice under H. A. Ivatt ...
Question	Fred Hoiberg was the coach of which team between 2016 and 2017?
Answer	Chicago Bulls | Bulls
Gold evidence	Fred Hoiberg | On June 2, 2015, the Chicago Bulls hired Hoiberg as head coach ... On December 3, 2018, the Bulls fired Hoiberg ...
Question	Who was the first spouse of Merle Oberon since May 7, 1948?
Answer	Lucien Ballard
Gold evidence	Merle Oberon | ... She divorced him in 1945, to marry cinematographer Lucien Ballard ...
Question	When was the last airplane crashing for All Nippon Airways as of 1970?
Answer	13 November 1966 | 1966
Gold evidence	All Nippon Airways | ... On 13 November 1966, Flight 533 operated by a NAMC YS-11, crashed in the Seto Inland Sea off ...

R Prompts List

Keyword Extraction Prompting
Your task is to extract keywords from the question. Response by a list of keyword strings. Do not include pronouns, prepositions, articles.

There are some examples for you to refer to:
<Question>
When was the last time the United States hosted the Olympics?
</Question>
<Keywords>
["United States", "hosted", "Olympics"]
</Keywords>

<Question>
Who sang 1 national anthem for Super Bowl last year?
</Question>
<Keywords>
["sang", "1", "national anthem", "Super Bowl"]
</Keywords>

<Question>
Who runs the fastest 40-yard dash in the NFL?
</Question>
<Keywords>
["runs", "fastest", "40-yard", "dash", "NFL"]
</Keywords>

<Question>
When did Khalid write Young Dumb and Broke?
</Question>
<Keywords>
["Khalid", "write", "Young Dumb and Broke"]
</Keywords>

Now your question is
<Question>
{normalized question}
</Question>
<Keywords>

Table 12: Detailed prompts for Keyword Extraction.

Query-Focused Summarization Prompting

You are given a context paragraph and a specific question. Your goal is to summarize the context paragraph in one standalone sentence by answering the given question. If dates are mentioned in the paragraph, include them in your answer. If the question cannot be answered based on the paragraph, respond with "None". Ensure that the response is relevant, complete, concise and directly addressing the question.

There are some examples for you to refer to:
<Context>
Houston Rockets | The Houston Rockets have won the NBA championship twice in their history. Their first win came in 1994, when they defeated the New York Knicks in a seven-game series. The following year, in 1995, they claimed their second title by sweeping the Orlando Magic. Despite several playoff appearances in the 2000s and 2010s, the Rockets have not reached the NBA Finals since their last championship victory in 1995.
</Context>
<Question>
When did the Houston Rockets win the NBA championship?
</Question>
<Summarization>
The Houston Rockets have won the NBA championship in 1994 and 1995.
</Summarization>

<Context>
2019 Grand National | The 2019 Grand National (officially known as the Randox Health 2019 Grand National for sponsorship reasons) was the 172nd annual running of the Grand National horse race at Aintree Racecourse near Liverpool, England. The showpiece steeplechase is the pinnacle of a three-day festival which began on 4 April, followed by Ladies' Day on 5 April.
</Context>
<Question>
Who won the Grand National?
</Question>
<Summarization>
None
</Summarization>

Now your question and paragraph are
<Context>
{title} | {text}
</Context>
<Question>
{normalized question}
</Question>
<Summarization>

Table 13: Detailed prompts for Query-Focused Summarization.

Reader Direct Prompting
As an assistant, your task is to answer the question directly after <Question>. Your answer should be after <Answer>.

There are some examples for you to refer to:
<Question>
When did England last get to the semi final of a World Cup before 2019?
</Question>
<Answer>
2018
</Answer>

<Question>
Who sang the national anthem in the last Super Bowl as of 2021?
</Question>
<Answer>
Eric Church and Jazmine Sullivan
</Answer>

<Question>
What's the name of the latest Pirates of the Caribbean by 2011?
</Question>
<Answer>
On Stranger Tides
</Answer>

<Question>
What was the last time France won World Cup between 2016 and 2019?
</Question>
<Answer>
Priscilla Joan Torres
</Answer>

<Question>
Which school did Marshall Sahlins go to from 1951 to 1952?
</Question>
<Answer>
Columbia University
</Answer>

Now your Question is
<Question>
{question}
</Question>
<Answer>

Table 14: Detailed prompts for Reader Direct Question Answering.

Reader Chain-of-Thought Prompting
As an assistant, your task is to answer the question after <Question>. You should first think step by step about the question and give your thought and then answer the <Question> in the short form. Your thought should be after <Thought>. The direct answer should be after <Answer>.

There are some examples for you to refer to:
<Question>
When did England last get to the semi final of a World Cup before 2019?
</Question>
<Thought>
England has reached the semi-finals of FIFA World Cup in 1966, 1990, 2018. The latest year before 2019 is 2018. So the answer is 2018.
</Thought>
<Answer>
2018
</Answer>

<Question>
Who sang the national anthem in the last Super Bowl as of 2021?
</Question>
<Thought>
The last Super Bowl as of 2021 is Super Bowl LV, which took place in February 2021. In Super Bowl LV, the national anthem was performed by Eric Church and Jazmine Sullivan. So the answer is Eric Church and Jazmine Sullivan.
</Thought>
<Answer>
Eric Church and Jazmine Sullivan
</Answer>

<Question>
Where was the last Rugby World Cup held between 2007 and 2016?
</Question>
<Thought>
The Rugby World Cup was held in 1987, 1991, 1995, 1999, 2003, 2007, 2011, 2015, 2019. The last Rugby World Cup held between 2007 and 2016 is in 2015. The IRB 2015 Rugby World Cup was hosted by England. So the answer is England.
</Thought>
<Answer>
England
</Answer>

Now your Question is
<Question>
{question}
</Question>
<Thought>

Table 15: Detailed prompts for Reader Chain-of-Thought Question Answering.

Retrieval-Augmented Reader Prompting

As an assistant, your task is to answer the question based on the given knowledge. Your answer should be after <Answer>. The given knowledge will be after the <Context> tage. You can refer to the knowledge to answer the question. If the context knowledge does not contain the answer, answer the question directly.

There are some examples for you to refer to:
<Context>
Sport in the United Kingdom Field | hockey is the second most popular team recreational sport in the United Kingdom. The Great Britain men's hockey team won the hockey tournament at the 1988 Olympics, while the women's hockey team repeated the success in the 2016 Games.

Three Lions (song) | The song reached number one on the UK Singles Chart again in 2018 following England reaching the semi-finals of the 2018 FIFA World Cup, with the line "it's coming home" featuring heavily on social media.

England national football team | They have qualified for the World Cup sixteen times, with fourth-place finishes in the 1990 and 2018 editions.
</Context>
<Question>
When did England last get to the semi final of a World Cup before 2019?
</Question>
<Answer>
2018
</Answer>

<Context>
Bowl LV | For Super Bowl LV, which took place in February 2021, the national anthem was performed by Eric Church and Jazmine Sullivan. They sang the anthem together as a duet.

Super Bowl LVI | For Super Bowl LVI, which took place in February 2022, the national anthem was performed by Mickey Guyton. She delivered a powerful rendition of the anthem.
</Context>
<Question>
Who sang the national anthem in the last Super Bowl as of 2021?
</Question>
<Answer>
Eric Church and Jazmine Sullivan
</Answer>

Now your question and context knowledge are
<Context>
{texts}
</Context>
<Question>
{question}
</Question>
<Answer>

Table 16: Detailed prompts for Retrieval-Augmented Question Answering.

Relevance Checking Prompting

You will be given a context paragraph and a question. Your task is to decide whether the context is relevant and contains the answer to the question. Requirements are as follows:

- First, read the paragraph after <Context> and the question after <Question> carefully.
- Then you should think step by step and give your thought after <Thought>.
- Finally, write the response as "Yes" or "No" after <Response>.

There are some examples for you to refer to:

<Context>

Petronas Towers | From 1996 to 2004, they were officially designated as the tallest buildings in the world until they were surpassed by the completion of Taipei 101. The Petronas Towers remain the world's tallest twin skyscrapers, surpassing the World Trade Center towers in New York City, and were the tallest buildings in Malaysia until 2019, when they were surpassed by The Exchange 106.

</Context>

<Question>

Tallest building in the world?

</Question>

<Thought>

The question asks what the tallest building in the world is. The context paragraph talks about the Petronas Towers. The context paragraph states that Petronas Towers were officially designated as the tallest buildings in the world from 1996 to 2004. And the Taipei 101 became the tallest building in the world after 2004. This context paragraph contains two answers to the question. Therefore, the response is "Yes".

</Thought>

<Response>

Yes

</Response>

Now your context paragraph and question are:

<Context>

{context}

</Context>

<Question>

{normalized question}

</Question>

<Thought>

Table 17: Detailed prompts for relevance checking.

Independent Reading Prompting

You are a summarizer summarizing a retrieved document about a user question. Keep the key dates in the summarization. Write "None" if the document has no relevant content about the question.

There are some examples for you to refer to:

<Document>

David Beckham | As the summer 2003 transfer window approached, Manchester United appeared keen to sell Beckham to Barcelona and the two clubs even announced that they reached a deal for Beckham's transfer, but instead he joined reigning Spanish champions Real Madrid for €37 million on a four-year contract. Beckham made his Galaxy debut, coming on for Alan Gordon in the 78th minute of a 0–1 friendly loss to Chelsea as part of the World Series of Soccer on 21 July 2007.

</Document>

<Question>

David Beckham played for which team?

</Question>

<Summarization>

David Beckham played for Real Madrid from 2003 to 2007 and for LA Galaxy from July 21, 2007.

</Summarization>

<Document>

Houston Rockets | The Houston Rockets have won the NBA championship twice in their history. Their first win came in 1994, when they defeated the New York Knicks in a seven-game series. The following year, in 1995, they claimed their second title by sweeping the Orlando Magic. Despite several playoff appearances in the 2000s and 2010s, the Rockets have not reached the NBA Finals since their last championship victory in 1995.

</Document>

<Question>

When did the Houston Rockets win the NBA championship?

</Question>

<Summarization>

The Houston Rockets won the NBA championship twice in 1994 and 1995.

</Summarization>

Now your document and question are:

<Document>

{document}

</Document>

<Question>

{normalized question}?

</Question>

<Summarization>

Table 18: Detailed prompts for Independent Reading.

Combined Reading Prompting

As an assistant, your task is to answer the question based on the given knowledge. Answer the given question; you can refer to the document provided. Your answer should follow the <Answer> tag. The given knowledge will be after the <Context> tag. You can refer to the knowledge to answer the question. Answer only the name for 'Who' questions. If the knowledge does not contain the answer, answer the question directly.

There are some examples for you to refer to:

<Context>

In 1977, Trump married Czech model Ivana Zelníčková. The couple divorced in 1990, following his affair with actress Marla Maples.

Trump and Maples married in 1993 and divorced in 1999.

In 2005, Donald Trump married Slovenian model Melania Knauss. They have one son, Barron (born 2006).

</Context>

<Question>

Who was the spouse of Donald Trump between 2010 and 2014?

</Question>

<Thought>

According to the context, Donald Trump married Melania Knauss in 2005. The period between 2010 and 2014 is after 2005. Therefore, the answer is Melania Knauss.

</Thought>

<Answer>

Melania Knauss

</Answer>

Now your question and context knowledge are:

<Context>

{generations}

</Context>

<Question>

{question}

</Question>

<Thought>

Table 19: Detailed prompts for Combined Reading.