# GEAR: Graph-enhanced Agent for Retrieval-augmented Generation

Anonymous ACL submission

## Abstract

Retrieval-augmented generation systems rely on effective document retrieval capabilities. By design, conventional sparse or dense retrievers face challenges in multi-hop retrieval scenarios. In this paper, we present GEAR, which advances RAG performance through two key innovations: (i) graph expansion, which enhances any conventional base retriever, such as BM25, and (ii) an agent framework that incorporates graph expansion. Our evaluation demonstrates GEAR's superior retrieval performance on three multi-hop question answering datasets. Additionally, our system achieves state-of-the-art results with improvements exceeding 10% on the challenging MuSiQue dataset, while requiring fewer tokens and iterations compared to other multi-step retrieval systems.

## 1 Introduction

Retrieval-augmented Generation (RAG) has further enhanced the remarkable success of Large Language Models (LLMs) (OpenAI, 2024) in Question Answering (QA) tasks (Lewis et al., 2020). Multi-hop QA usually requires reasoning capabilities across several passages or documents. A relevant example is displayed in Table 1 where reaching the appropriate answer requires building a 3-hop reasoning chain starting from the main entity in the question (i.e. "Stephen Curry").

| What year did the father of Stephen Curry joined the team from which he started his college basketball career? | | |
|---|---|---|
| Stephen Curry | son of → | Dell Curry |
| Dell Curry | college team → | Virginia Tech |
| Dell Curry | college start → | 1982 (answer) |

Table 1: Multi-hop question (top) involving a reasoning chain (bottom) extending across several entities.

More recently, existing methods sought to leverage graph representations of the retrieved passages in order to bridge the semantic gap introduced by multi-hop questions (Fang et al., 2024; Li et al., 2024; Edge et al., 2024; Gutierrez et al., 2024; Liang et al., 2024). Most of these approaches employ an LLM to traverse a graph involving the entities appearing in the corresponding textual passages. However, within the context of RAG, this typically leads to long and interleaved prompts that require multiple LLM iterations to arrive at answers involving distant reasoning hops (Trivedi et al., 2023). Several recent approaches build graphs associating passages with each other by extracting entities and atomic facts or semantic triples from passages in a separate offline step (Li et al., 2024; Fang et al., 2024; Gutierrez et al., 2024). Furthermore, GraphReader uses an LLM agent, with access to graph-navigating operations for exploring the resulting graph (Li et al., 2024). TRACE relies on an LLM to iteratively select triples to construct reasoning chains, which are then used for grounding the answer generation directly, or for filtering out irrelevant documents from an original set of retrieved results (Fang et al., 2024).

In this paper, we present GEAR, a Graph-enhanced Agent for Retrieval-augmented generation. During the offline stage, we *align* an index of passages with an index of triples extracted from these passages. With such alignment, passages are intermediately connected through graphs of triples. GEAR contains a graph-based passage retrieval component referred to as SyncGE. Differentiating from previous works that rely on expensive LLM calls for graph exploration, we leverage an LLM for locating initial nodes (triples) and employ a generic semantic model to expand the sub-graph of triples by exploring diverse beams of triples. Furthermore, GEAR utilises multi-hop contexts retrieved by SyncGE and constructs a memory that summarises information for multi-step retrieval.

Our work refines the neurobiology-inspired paradigm proposed by Gutierrez et al., by mod-

elling the communication between hippocampus and neocortex when forming an episodic memory. An array of *proximal triples*, in our design, functions as a gist of memory learnt through hippocampus within one or a few shots (iterations), which is projected back to neocortex for the later recall stages (Hanslmayr et al., 2016; Griffiths et al., 2019). We highlight the complementary potential of our graph retrieval approach and an LLM, which, within our system, assimilates the synergy between the hippocampus and neocortex, offering insights from a biomimetic perspective.

We evaluate the retrieval performance of GEAR on three multi-hop QA benchmarks: MuSiQue, HotpotQA, and 2WikiMultihopQA. GEAR pushes the state of the art, achieving significant improvements in both single- and multi-step retrieval settings, with gains exceeding $10\%$ on the most challenging MuSiQue dataset. Furthermore, we demonstrate that our framework can address multi-hop questions in fewer iterations with significantly fewer LLM tokens. Even in the case of a single iteration, GEAR offers a more efficient alternative to other iterative retrieval methods, such as HippoRAG w/ IRCoT. Our contributions can be summarised as follows:

- We introduce a novel graph-based retriever, SyncGE, which leverages an LLM for locating initial nodes for graph exploration and subsequently expands them by diversifying beams of triples that link multi-hop passages.
- We incorporate this graph retrieval method within an LLM-based agent framework, materialising GEAR, achieving state-of-the-art retrieval performance across three datasets.
- We conduct comprehensive experiments showcasing the synergetic effects between our proposed graph-based retriever and the LLM within the GEAR framework.

## 2 Related Work

Our work draws inspiration from two branches of research: (i) retrieval-augmented models for QA and (ii) multi-hop QA using combinations of LLMs with graphical structures.

### 2.1 Retrieval-augmented Models for QA

Since Lewis et al. showcased the benefits of augmenting the input context of language models with relevant passages, several solutions have been proposed for addressing different knowledge-intensive scenarios (Pan et al., 2023).

Recent works by Wang et al.; Shen et al. explore query expansion approaches, generating pseudo-documents from the LLM to expand the content of the original query. Subsequent frameworks, starting with IRCoT, looked into interleaving retrieval and prompting steps, allowing each step to guide and refine the other iteratively (Trivedi et al., 2023; Jiang et al., 2023; Su et al., 2024).

### 2.2 Multi-hop QA with LLMs and Graphs

In the recent years, several architectures introduce a separate, offline indexing phase during which they form a hierarchical summary of passages (Chen et al., 2023; Sarthi et al., 2024; Edge et al., 2024). However, the summarisation process must be repeated whenever new data is added. This can be computationally expensive and inefficient for updating the knowledge base.

More recently several approaches sought to leverage the benefits of incorporating structured knowledge for addressing multi-hop QA challenges with LLMs (Park et al., 2023; Fang et al., 2024; Li et al., 2024; Gutierrez et al., 2024; Liang et al., 2024; Wang et al., 2024). GraphReader, TRACE and HippoRAG propose offline methodologies for extracting entities and atomic facts or semantic triples from passages (Li et al., 2024; Fang et al., 2024; Gutierrez et al., 2024). TRACE relies on an LLM to iteratively select triples to construct reasoning chains, which are then used for grounding the answer generation directly or for filtering retrieved results. However, the search space is limited as an already filtered candidate list is provided for each query. Li et al. utilise an LLM agent capable of selecting from a set of predefined actions to traverse the nodes of a knowledge graph in real time given an input question. More recently, Liang et al. introduced further standardisation for the offline graph, such as instance-to-concept linking and semantic relation completion. Nonetheless, the approach relies heavily on associating triples with pre-defined concepts to facilitate logical form-based retrieval.

HippoRAG leverages an alignment of passages and extracted triples in order to retrieve passages based on the Personalised PageRank algorithm (Gutierrez et al., 2024). While achieving considerable improvements for single- and multi-step retrieval (i.e. when coupled with IRCoT (Trivedi et al., 2023)), it remains agnostic to the semantic

relationships of the extracted triples. In this paper, we leverage a similar alignment of passages and extracted triples; but, instead of fully relying on expensive LLM calls, we introduce a new graph-based retrieval framework that uses a small semantic model for exploring multi-hop relationships.

## 3 Preliminaries

Let $\mathbf{C} = \{c_1, c_2, \ldots, c_C\}$ be an index of passages and $\mathbf{T} = \{t_1, t_2, \ldots, t_T : t_j = (s_j, p_j, o_j)\}$ be another index representing a set of triples associated with the passages in $\mathbf{C}$ s.t. $\forall t_j \in \mathbf{T} \exists! \, c_i \in \mathbf{C}$, where $s_j$, $p_j$ and $o_j$ the respective subject, predicate and object of the $j$-th triple.

Given an input query $\mathbf{q}$ and an index of interest $\mathbf{R} = \{r_1, \ldots, r_R\}$, retrieving items from $\mathbf{R}$ relevant to $\mathbf{q}$ can be achieved by using a base retrieval function $h^k_{\text{base}}(\mathbf{q}, \mathbf{R}) \subseteq \mathbf{R}$ that returns a ranked list of $k$ items from $\mathbf{R}$ in descending order, according to a retrieval score. BM25 or a conventional dense retriever can serve as a base retrieval function, without requiring any multi-hop capabilities.

Our goal is to retrieve relevant passages from $\mathbf{C}$ that enable a retrieval-augmented model to answer multi-hop queries (Lewis et al., 2020). To this end, we introduce GEAR, which is a graph-enhanced framework of retrieval agent (see Figure 1).

## 4 Retrieval with Graph Expansion

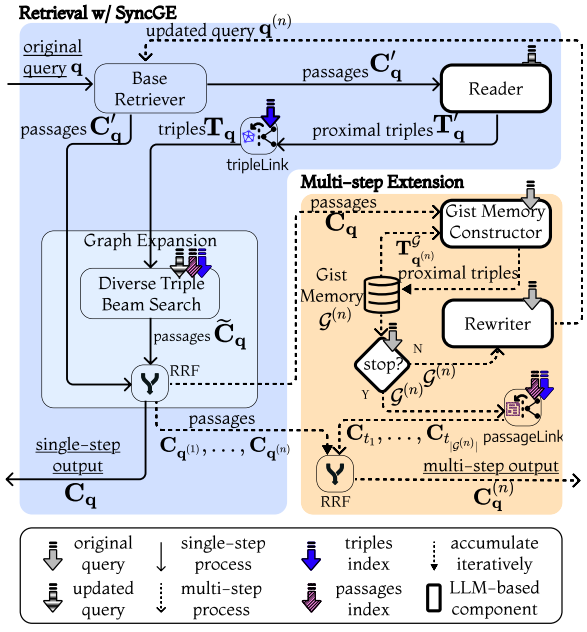


Figure 1: System Architecture

Given an input query $\mathbf{q}$, let $\mathbf{C}'_\mathbf{q} = h^k_{\text{base}}(\mathbf{q}, \mathbf{C})$ be a list of passages returned by the base retriever. Given this initially retrieved list of passages, $\mathbf{C}'_\mathbf{q}$, our goal is to derive relevant multi-hop contexts (passages) by retrieving a sub-graph of triples that interconnect their source passages. There are two challenges for materialising such sub-graph retrieval: (i) how to locate initial triples (i.e. starting nodes) $\mathbf{T}_\mathbf{q}$, and (ii) how to expand the graph based on initial triples while reducing the search space. The following sections address these challenges respectively, within GEAR.

### 4.1 Knowledge Synchronisation

We describe a knowledge **Sync**hronisation (**Sync**) process for locating initial nodes for graph expansion. We first employ an LLM to `read` $\mathbf{C}'_\mathbf{q}$ (see Appendix I.2) and summarise knowledge triples that can support answering the current query $\mathbf{q}$, as defined:

$$\mathbf{T}'_\mathbf{q} = \texttt{read}\left(\mathbf{C}'_\mathbf{q}, \mathbf{q}\right). \quad (1)$$

$\mathbf{T}'_\mathbf{q}$ is a collection of triples to which we refer as *proximal triples*. Initial nodes $\mathbf{T}_\mathbf{q}$ for graph expansion can then be identified by linking each triple in $\mathbf{T}'_\mathbf{q}$ to a triple in $\mathbf{T}$, using the `tripleLink` function:

$$\mathbf{T}_\mathbf{q} = \left\{t_i | t_i = \texttt{tripleLink}(t'_i) \; \forall t'_i \in \mathbf{T}'_\mathbf{q}\right\}. \quad (2)$$

The implementation of `tripleLink` can vary. However, in this paper we consider it to be simply retrieving the most similar triple from $\mathbf{T}$.

### 4.2 Diverse Triple Beam Search

We borrow the idea of constructing reasoning triple chains (Fang et al., 2024) for expanding the graph, and present a retrieval algorithm: *Diverse Triple Beam Search* (see Alg. 1).

We maintain top-$b$ sequences (beams) of triples and the scores at each step are determined by a scoring function. In this paper, we focus on leveraging a dense embedding model to compute the cosine similarity between embeddings of the query and a candidate sequence of triples, leaving other implementations of the scoring function for future work (see Section 9).

Considering all possible triple extensions at each step, in a Viterbi decoding fashion, would be intractable due to the size of $\mathbf{T}$. Consequently, we define the neighbourhood of a triple as the set of triples with shared head or tail entities (i.e. get_neighbours in Alg. 1). During each expansion step, we only consider neighbours of the last

**Algorithm 1** Diverse Triple Beam Search

**Input:** $\mathbf{q}$: query
$b$: beam size
$l$: maximum length
$\text{score}(\cdot,\cdot)$: scoring function
$\{t_1, t_2, \ldots, t_n\}$: initial triples
$\gamma$: hyperparameter for diversity

$B_0 \leftarrow [\,]$
**for** $t \in \{t_1, t_2, ..., t_n\}$ **do**
  $s \leftarrow \text{score}(\mathbf{q}, [t])$
  $B_0.\text{add}(\langle s, [t] \rangle)$
$B_0 \leftarrow \text{top}(B_0, b)$
**for** $i \in \{1, \ldots, l-1\}$ **do**
  $B \leftarrow [\,]$
  **for** $\langle s, T \rangle \in B_{i-1}$ **do**
    $V \leftarrow [\,]$
    **for** $t \in \text{get\_neighbours}(T.\text{last}())$ **do**
      **if** $\text{exists}(t, B_{i-1})$ **then**
        **continue**
      $s' \leftarrow s + \text{score}(\mathbf{q}, T \circ t)$ `# concat`
      $V.\text{add}(\langle s', T \circ t \rangle)$
    $\text{sort}(V, \text{descending})$
    **for** $n \in \{0, \ldots, V.\text{length}() - 1\}$ **do**
      $\langle s', T \circ t \rangle \leftarrow V[n]$
      $s' \leftarrow s' \times e^{-\frac{\min(n,\gamma)}{\gamma}}$
      $B.\text{add}(\langle s', T \circ t \rangle)$
  $B_i \leftarrow \text{top}(B, b)$
**return** $B_i$

triple in the sequence, and avoid selecting previously visited triples (i.e. exists in Alg. 1).

While regular beam search can reduce the search space, it is prone to producing high-likelihood sequences that differ only slightly from one another (Ippolito et al., 2019; Vijayakumar et al., 2018). Our algorithm increases the diversity across beams to improve the recall for retrieval. In detail, for each beam, we sort candidate sequences extended from that beam in descending order, and weight their scores based on their relative positions. Candidate sequences that are ranked lower, within a beam, will receive smaller weights. Consequently, the resulting top-$b$ beams at each step are less likely to share the same starting sequence.

The top-$b$ returned sequences are flattened in a breadth-first order. Each triple in the resulting list is then mapped to its source passage. This alignment between triples and passages is described in more detail in Section 3. Let $\widetilde{\mathbf{C}}_\mathbf{q}$ be the list of unique passages after alignment. The output of our graph expansion is then given by the Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) of $\widetilde{\mathbf{C}}_\mathbf{q}$ and the initial $\mathbf{C}'_\mathbf{q}$ list of passages :

$$\mathbf{C}_\mathbf{q} = \text{RRF}\left(\widetilde{\mathbf{C}}_\mathbf{q}, \mathbf{C}'_\mathbf{q}\right). \tag{3}$$

We refer to this graph-based method of retrieving relevant passages as **Sync**ronised **G**raph **E**xpansion (**SyncGE**).

## 5 Multi-step Extension

While SyncGE can enhance a base retriever with multi-hop context, some queries inherently require multiple steps to gather all necessary evidence. We materialise GEAR by incorporating an agent with multi-turn capabilities, capable of interacting with the graph-retriever described above. We focus on:

- maintaining a gist memory of proximal knowledge obtained throughout the different steps
- incorporating a similar synchronisation process that summarises retrieved passages in proximal triples to be stored in this multi-turn gist memory
- determining if additional steps are needed for answering the original input question

Within this multi-turn setting, the original input question $\mathbf{q}$ is iteratively decomposed into simpler queries: $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}$, where $\mathbf{q}^{(1)} = \mathbf{q}$ and $n \in \mathbb{N}$ represents the number of the current step. For each query $\mathbf{q}^{(n)}$, we use the graph retrieval method introduced in Section 4 in order to retrieve relevant passages $\mathbf{C}_{\mathbf{q}^{(n)}}$.

### 5.1 Gist Memory Constructor

To facilitate the multi-step capabilities of our agent, we introduce a *gist memory*, $\mathcal{G}^{(n)}$, which is used for storing knowledge as an array of proximal triples. At the beginning of the first iteration, the gist memory is empty. During the $n$-th iteration, similar to the knowledge synchronisation module explained in Section 4.1, we employ an LLM to read a collection of retrieved paragraphs $\mathbf{C}_{\mathbf{q}^{(n)}}$ and summarise their content with proximal triples:

$$\mathbf{T}^{\mathcal{G}}_{\mathbf{q}^{(n)}} = \begin{cases} \texttt{read}\left(\mathbf{C}_{\mathbf{q}^{(n)}}, \mathbf{q}\right), & \text{if } n = 1 \\ \texttt{read}\left(\mathbf{C}_{\mathbf{q}^{(n)}}, \mathbf{q}, \mathcal{G}^{(n-1)}\right), & \text{if } n \geq 2 \end{cases} \tag{4}$$

Apart from the first iteration where Eq. 1 and 4 are identical, the inclusion of the memory in the `read` operation differentiates the construction of proximal triples produced at the subsequent steps compared to the ones from Eq. 1. $\mathcal{G}^{(n)}$ maintains the aggregated content of proximal triples s.t.

$$\mathcal{G}^{(n)} = \left[\mathbf{T}^{\mathcal{G}}_{\mathbf{q}^{(1)}} \circ \cdots \circ \mathbf{T}^{\mathcal{G}}_{\mathbf{q}^{(n)}}\right], \tag{5}$$

where $\circ$ defines the concatenation operation. The triple memory serves as a concise representation of all the accumulated evidence, up to the $n$-th step.

We believe the process introduced by the `read` step along with the information storage paradigm served by the gist memory, aligns well with the communication between the hippocampus and neocortex. The combination of the two establishes the synergetic behaviour between our graph retriever and the LLM that we seek to achieve within GEAR.

### 5.2 Reasoning for Termination

After $\mathcal{G}^{(n)}$ is updated, we check the sufficiency of the accumulated evidence, within it, for answering the original question. This is achieved with the following LLM reasoning step:

$$\mathbf{a}^{(n)}, \mathbf{r}^{(n)} = \texttt{reason}(\mathcal{G}^{(n)}, \mathbf{q}), \tag{6}$$

where $\mathbf{a}^{(n)}$ denotes the query's answerability given the available evidence in $\mathcal{G}^{(n)}$, and $\mathbf{r}^{(n)}$ represents the reasoning behind this determination. When the query is deemed answerable, the system concludes its iterative process.

### 5.3 Query Re-writing

The query re-writing process leverages an LLM that incorporates three key inputs: the original query $\mathbf{q}$, the accumulated memory, and crucially, the reasoning output $\mathbf{r}^{(n)}$ from the previous step. This process can be formally expressed as:

$$\mathbf{q}^{(n+1)} = \texttt{rewrite}\left(\mathcal{G}^{(n)}, \mathbf{q}, \mathbf{r}^{(n)}\right), \tag{7}$$

where $\mathbf{q}^{(n+1)}$ represents the updated query, which serves as input for the retriever in the next iteration.

### 5.4 After Termination

GEAR aims to return a single ranked list of passages. Given the final gist memory $\mathcal{G}^{(n)}$ upon termination, we link each proximal triple in $\mathcal{G}^{(n)}$ to a list of passages as follows:

$$\mathbf{C}_{t_j} = \texttt{passageLink}\left(t_j, k\right), \tag{8}$$

where $j \in \{1, \ldots, |\mathcal{G}^{(n)}|\}$. Similar to `tripleLink`, `passageLink` is implemented by retrieving passages with a triple as the query (see Appendix C.2). The final list of passages returned by GEAR is the RRF of the resulting linked passages and passages retrieved across iterations:

$$\mathbf{C}^{(n)}_{\mathbf{q}} = \text{RRF}\big(\mathbf{C}_{t_1}, \ldots, \mathbf{C}_{t_{|\mathcal{G}^{(n)}|}}, \mathbf{C}_{\mathbf{q}^{(1)}}, \ldots, \mathbf{C}_{\mathbf{q}^{(n)}}\big). \tag{9}$$

All relevant prompts for the `read`, `reason` and `rewrite` steps are provided in Appendix I.2.

## 6 Experimental Setup

We evaluate our proposed framework on three multi-hop QA datasets in the open-domain setting: **MuSiQue** (answerable subset) (Trivedi et al., 2022), **HotpotQA** (Yang et al., 2018), and **2WikiMultiHopQA** (2Wiki) (Ho et al., 2020). For MuSiQue and 2Wiki, we use the data splits provided in IRCoT (Trivedi et al., 2023), while for HotpotQA we follow the same data setting as in HippoRAG (Gutierrez et al., 2024). Dataset-specific statistics can be found in Appendix B.

We measure both retrieval and QA performance, with our primary contributions focused on the retrieval component. For retrieval evaluation, we use Recall@$k$ (R@$k$) metrics for $k \in \{5, 10, 15\}$, showing the percentage of questions where the correct entries are found within the top-$k$ retrieved passages. We include an analysis about the selected recall ranks in Appendix B. Following standard practices, QA performance is evaluated with Exact Match (EM) and F1 scores (Trivedi et al., 2023).

### 6.1 Baselines

We evaluate GEAR against strong, multi-step baselines, including IRCoT (Trivedi et al., 2023) and a combination of HippoRAG w/ IRCoT (Gutierrez et al., 2024) which, similar to our framework, includes a graph-retrieval component and a multi-step agent. To showcase the benefits of our graph retriever (i.e. SyncGE), we evaluate it against several stand-alone, single-step retrievers: (i) BM25, (ii) Sentence-BERT (SBERT), (iii) a hybrid approach that combines BM25 and SBERT results through RRF and (iv) HippoRAG. Throughout the experiments, we refer to the single-step setup when an approach does not support several iterations and is not equipped with an LLM agent.

| | Retriever | MuSiQue | | | 2Wiki | | | HotpotQA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 |
| **Single-step Retrieval** | ColBERTv2 | 39.4 | 44.8 | 47.7 | 59.1 | 64.3 | 66.2 | 79.3 | 87.1 | 90.1 |
| | HippoRAG | 41.0 | 47.0 | 51.4 | **75.1** | **83.2** | **86.4** | 79.8 | 89.0 | 92.4 |
| | BM25 | 33.8 | 38.5 | 41.3 | 59.5 | 62.7 | 64.1 | 74.2 | 83.6 | 86.3 |
| | + NaiveGE | 37.5 | 45.5 | 48.4 | 65.0 | 70.7 | 71.8 | 79.1 | 89.1 | 91.9 |
| | + SyncGE | 44.7 | 52.6 | 57.4 | 70.5 | 76.1 | 79.3 | 87.4 | 93.0 | 94.0 |
| | SBERT | 31.1 | 37.9 | 41.6 | 41.2 | 48.1 | 51.5 | 72.1 | 79.3 | 84.0 |
| | + NaiveGE | 32.2 | 41.4 | 45.4 | 45.1 | 54.0 | 57.3 | 76.1 | 84.7 | 88.8 |
| | + SyncGE | 41.6 | 51.3 | 54.2 | 54.8 | 64.9 | 70.7 | 84.1 | 89.6 | 92.8 |
| | Hybrid | 39.9 | 46.3 | 49.1 | 60.0 | 65.8 | 66.6 | 77.8 | 85.8 | 89.7 |
| | + NaiveGE | 41.8 | 49.4 | 53.0 | 63.0 | 70.8 | 72.6 | 80.6 | 89.4 | 92.7 |
| | + SyncGE | **48.7** | **57.7** | **61.2** | 72.6 | 80.9 | 82.4 | **87.4** | **93.3** | **95.2** |
| **Multi-step Retrieval** | IRCoT (BM25) | 46.1 | 54.9 | 57.9 | 67.9 | 75.5 | 76.1 | 87.0 | 92.6 | 92.9 |
| | IRCoT (ColBERTv2) | 47.9 | 54.3 | 56.4 | 60.3 | 86.6 | 69.7 | 86.9 | 92.5 | 92.8 |
| | HippoRAG w/ IRCoT | 48.8 | 54.5 | 58.9 | 82.9 | 90.6 | 93.0 | 90.1 | 94.7 | 95.9 |
| | GEAR | **58.4** | **67.6** | **71.5** | **89.1** | **95.3** | **95.9** | **93.4** | **96.8** | **97.3** |

Table 2: Retrieval performance for single- and multi-step retrievers on MuSiQue, 2Wiki, and HotpotQA. Results are reported using Recall@$k$ (R@$k$) metrics for $k \in \{5, 10, 15\}$, showing the percentage of questions where the correct entries are found within the top-$k$ retrieved passages.

### 6.2 Implementation Details

To maintain consistency and validity in comparisons with the baselines on the splits used in this study, we conducted all experiments locally using their corresponding codebases.

In addition to our proposed single-step retriever, SyncGE, we evaluate a more *naive* implementation of GE (i.e. NaiveGE) in order to explore the generality of the method when in resource-constrained setting, where no LLM is involved. In NaiveGE, we use all triples that are associated with $\mathbf{C}'_{\mathbf{q}}$ (see Section 4) for diverse triple beam search.

For all models using an LLM, we employ GPT-4o mini (`gpt-4o-mini-2024-07-18`) as the backbone model with a temperature of 0, both for offline triple extraction (i.e. how the $\mathbf{T}$ index in Section 3 is formed) and online retrieval operations. Our triple extraction prompt (in Appendix I.1) is adapted[1] from the ones used by Gutierrez et al.. To ensure a fair comparison against Gutierrez et al., the closest work to ours, we run experiments with HippoRAG using our prompting setup[2] for triple extraction. For evaluating QA performance, we use the prompts provided in Appendix I.3. Further implementation details are provided in Appendix C.

[1]Our approach uses a modified version of HippoRAG's triple extraction prompt that combines entity and triple extraction into a single step, while incorporating an additional demonstration and updated in-context examples.

[2]For transparency, we also compare against HippoRAG's original triple extraction prompt in Appendix H, where we observe only minor differences across the two configurations.

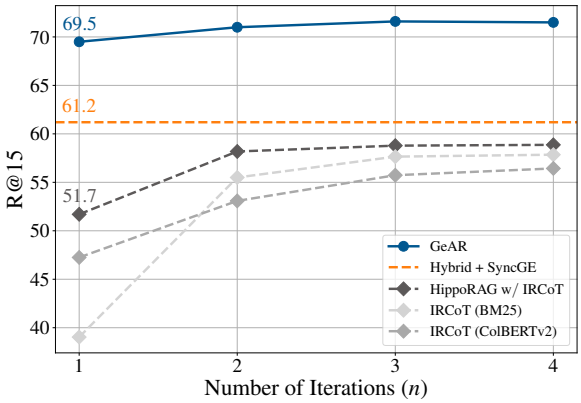


Figure 2: R@15 evolution over 4 iterations on MuSiQue. Recall is computed at each iteration using the cumulative set of retrieved documents, with prior recall values carried forward for questions that terminated in earlier iterations. The horizontal line indicates the single-step performance of Hybrid + SyncGE.

## 7 Results

**GEAR achieves state-of-the-art performance in multi-step retrieval** The multi-step section of Table 2 demonstrates that our agent setup for enabling multi-step retrieval is highly effective, achieving state-of-the-art performance across all datasets. While we observe significant improvements on saturated datasets like 2Wiki and HotpotQA, GEAR particularly excels on the MuSiQue dataset, delivering performance gains exceeding 10% over the competition.

**SyncGE contributes to state-of-the-art performance in single-step retrieval** As shown in the

single-step section of Table 2, our proposed Hybrid + SyncGE method achieves state-of-the-art single-step retrieval performance on both MuSiQue and HotpotQA datasets. We observe consistent improvements using NaiveGE and SyncGE, outperforming HippoRAG in many setups regardless of the base retriever (i.e. sparse, dense or hybrid). Most notably, Hybrid + SyncGE surpasses HippoRAG by up to 9.8% at R@15 on MuSiQue.

**Higher recall leads to higher QA performance** Aligning with findings from prior works, our analysis reveals a consistent correlation between recall and QA performance (Gutierrez et al., 2024). As shown in Table 3, GEAR achieves the highest EM and F1 scores. A closer examination of relative improvements yields interesting insights. Taking MuSiQue as an example, GEAR shows a 21% relative improvement in R@15 compared to HippoRAG w/ IRCoT, while achieving a 37% relative improvement in both EM and F1 scores. Mirroring the pattern observed in Table 2, its graph-based retriever (i.e. SyncGE) outperforms HippoRAG on both MuSiQue and HotpotQA.

| Retriever | MuSiQue | | 2Wiki | | HotpotQA | |
|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 |
| No Passages | 2.6 | 12.5 | 17.2 | 27.9 | 19.5 | 34.3 |
| Gold Passages | 36.6 | 59.2 | 54.4 | 70.3 | 55.0 | 75.9 |
| Hybrid + SyncGE | 14.0 | <u>27.1</u> | 38.0 | 50.2 | 45.0 | 63.4 |
| HippoRAG | 8.2 | 18.2 | 39.8 | 51.8 | 40.1 | 57.6 |
| IRCoT (BM25) | 7.6 | 15.9 | 28.8 | 38.5 | 34.3 | 50.8 |
| IRCoT (ColBERTv2) | 12.2 | 24.1 | 32.4 | 43.6 | 45.2 | 63.7 |
| HippoRAG w/ IRCoT | <u>14.2</u> | 25.9 | <u>45.6</u> | <u>59.0</u> | <u>49.2</u> | <u>67.9</u> |
| GEAR | **19.0** | **35.6** | **47.4** | **62.3** | **50.4** | **69.4** |

Table 3: End-to-end QA performance using the top-5 retrieved passages. The **best** model is in bold and <u>second best</u> is underlined. The top part shows the lower and upper bounds of QA performance, while the middle and bottom sections display scores for single-step and multi-step retrievers, respectively.

## 8 Discussion

### 8.1 What makes GEAR work?

**NaiveGE vs SyncGE** As shown in Table 2, both variants of graph expansion enhance the performance of every base retriever across all datasets. The case of SyncGE is particularly interesting since without any LLM agent, it is able to surpass the retrieval performance that HippoRAG w/ IRCoT can achieve after several LLM iterations, on the challenging MuSiQue dataset.

| Metric | Dataset | w/ Diversity | w/o Diversity |
|---|---|---|---|
| R@5 | MuSiQue | **48.7** | 47.0 |
| | 2Wiki | **72.6** | 68.2 |
| | HotpotQA | **87.4** | 85.0 |
| R@10 | MuSiQue | **57.7** | 53.9 |
| | 2Wiki | **80.9** | 76.0 |
| | HotpotQA | **93.3** | 92.2 |
| R@15 | MuSiQue | **61.2** | 58.4 |
| | 2Wiki | **82.4** | 77.4 |
| | HotpotQA | **95.2** | 94.3 |

Table 4: Effects of beam search diversity on Hybrid + SyncGE retrieval performances across MuSiQue, 2Wiki and HotpotQA.

**Diverse Triple Beam Search improves performance** As shown in Table 4, diverse beam search consistently outperforms standard beam search across all evaluated datasets and recall ranks. By incorporating diversity weights into beam search, we align a language modelling-oriented solution with information retrieval objectives that involve satisfying multiple information needs underlying multi-hop queries (Drosou and Pitoura, 2010).

**GEAR *mostly* nails it the first time** While GEAR supports multiple iterations, Figure 2 shows that on MuSiQue, GEAR can achieve strong retrieval performance within a single iteration. This differentiates it from the IRCoT-oriented setups that require at least 2 iterations to reach their maximum performance. This can be attributed to the fact that GEAR reads (Eq. 4) multi-hop contexts and associates the summarised proximal triples in the gist memory with passages, establishing a synergetic behaviour between our graph retriever and the LLM. We believe this mirrors the hippocampal process of forming and resolving sparse representations, where gist memories are learnt in a one or few-shot manner (Hanslmayr et al., 2016). The performance difference between Hybrid + SyncGE and GEAR at $n = 1$, approximately 10%, indicates that the involved LLM *reading* and linking processes can effectively approximate the role of hippocampus within our framework.

### 8.2 Where does GEAR demonstrate performance gains?

**GEAR excels at questions of low-to-moderate complexity** Figure 3 presents a detailed breakdown of retrieval performance across different hop

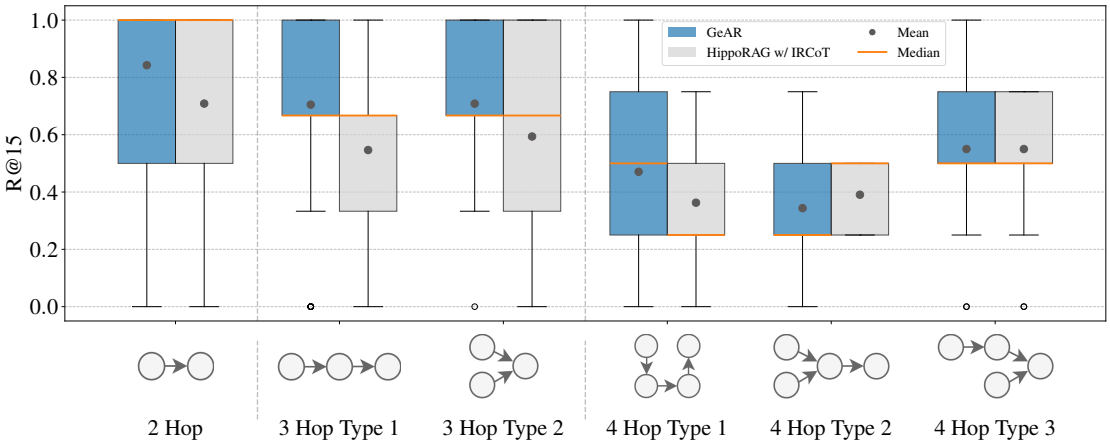


Figure 3: Analysis of R@15 performance divided by hop types on MuSiQue. The hop categorisation follows the MuSiQue documentation. Mean recall values are indicated by grey dots for each hop type.

types in MuSiQue. For 2-hop questions, while GEAR and HippoRAG w/ IRCoT achieve similar interquartile ranges, GEAR demonstrates a notably higher mean recall, indicating superior performance on low-complexity questions. This advantage becomes more pronounced with 3-hop questions, where GEAR's entire interquartile range exceeds HippoRAG w/ IRCoT's median performance across both types of hop subdivisions. This demonstrates GEAR's enhanced capability in handling questions of moderate complexity.

### 8.3 Is GEAR efficient?

**GEAR requires fewer iterations** As we observe in Figure 2, GEAR requires fewer iterations than the competition to reach its maximum recall performance. We attribute this to the fact that SyncGE enables GEAR to bridge passages across distant reasoning hops, resulting in fewer iterations.

**GeAR requires fewer LLM tokens** In Figure 4, we showcase that GEAR can act as a more efficient alternative with respect to LLM token utilisation[3]. We observe that even for a single iteration, GEAR uses fewer tokens than HippoRAG w/ IRCoT. In contrast to ours, this trend exacerbates for the competition as the number of iterations increases.

The findings from this figure also reiterate the value of SyncGE (see Section 8.1), which is able to outperform a significantly more LLM-heavy solution in MuSiQue, using almost 2.9 million fewer tokens. Even in the case that HippoRAG w/ IRCoT runs for a single iteration it would require more than 0.7 million tokens that Hybrid + SyncGE, with a substantially lower R@15 of 51.7.

[3]Tokenisation performed using OpenAI's `tiktoken` tool.

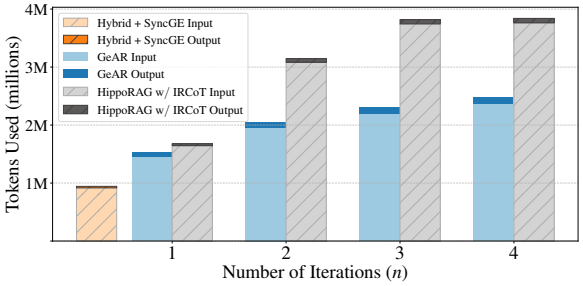


Figure 4: Progressive accumulation of input and output LLM tokens across agent iterations on MuSiQue.

## 9 Conclusion

We propose GEAR, a novel framework that incorporates a graph-based retriever within a multi-step retrieval agent to model the information-seeking process for multi-hop question answering.

We showcase the synergy between our proposed graph retriever (i.e. SyncGE) and the LLM within the GEAR framework. SyncGE leverages the LLM to synchronise information from passages with triples and expands the graph by exploring diverse beams of triples that link multi-hop contexts. Our experiments reveal that this strategy improves over more naive implementations, demonstrating the LLM's capability to guide the exploration of initial nodes for graph expansion. Furthermore, GEAR utilises multi-hop contexts returned by SyncGE and constructs a gist memory which is used for effectively summarising information across iterations. GEAR archives superior performance compared to other multi-step retrieval methods while requiring fewer iterations and LLM tokens.

## Limitations

The scope of this paper is limited to retrieval with the aid of a graph of triples that bridge corresponding passages. While we demonstrated the efficacy of our graph expansion approach and GEAR, we acknowledge that the implementation of the underlying graph is rather simple. Better graph construction that addresses challenges such as entity disambiguation (Dredze et al., 2010) and knowledge graph completion (Lin et al., 2015) can lead to further improvements.

We focused on employing a dense embedding model for our diverse triple beam search scoring function, though alternative functions could open up promising avenues for future research. For example, one can study the feasibility of formulating the scoring of neighbours as a natural language inference task (Wang et al., 2021), using a model that predicts how confidently a sequence of triples answers the given query.

Additionally, our approach relies on LLMs that can be better prompted to achieve superior performance on the relevant GEAR tasks. Nonetheless, we provide more experiments in Appendix D showcasing that GEAR can achieve equivalent performance with open-weight LLMs.

## References


Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *Preprint*, arXiv:2310.05029.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *Preprint*, arXiv:2402.13753.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Marina Drosou and Evaggelia Pitoura. 2010. Search result diversification. *SIGMOD Rec.*, 39(1):41–47.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *Preprint*, arXiv:2406.11460.

Benjamin J Griffiths, George Parish, Frederic Roux, Sebastian Michelmann, Mircea Van Der Plas, Luca D Kolibius, Ramesh Chelvarajah, David T Rollings, Vijay Sawlani, Hajo Hamer, et al. 2019. Directional coupling of slow and fast hippocampal gamma with neocortical alpha/beta oscillations in human episodic memory. *Proceedings of the National Academy of Sciences*, 116(43):21834–21842.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Simon Hanslmayr, Bernhard P Staresina, and Howard Bowman. 2016. Oscillations and episodic memory: addressing the synchronization/desynchronization conundrum. *Trends in neurosciences*, 39(1):16–25.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy. Association for Computational Linguistics.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng.

2024. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786, Miami, Florida, USA. Association for Computational Linguistics.

Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. KAG: Boosting llms in professional domains via knowledge augmented generation. *Preprint*, arXiv:2409.13731.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges. *ArXiv*, abs/2308.06374.

Jinyoung Park, Ameen Patel, Omar Zia Khan, Hyunwoo J. Kim, and Joo-Kyung Kim. 2023. Graph-guided reasoning for multi-hop question answering in large language models. *CoRR*, abs/2311.09762.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large language models are strong zero-shot retriever. *ArXiv*, abs/2304.14233.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

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

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *Preprint*, arXiv:1610.02424.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *ArXiv*, abs/2104.14690.

Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19206–19214.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

## A Hybrid Retrieval Strategy

A list of $k$ passages by merging returned passages from both a $\mathbf{C} = \{c_1, c_2, \ldots, c_C\}$ index of textual passages and a $\mathbf{T} = \{t_1, t_2, \ldots, t_T : t_j = (s_j, p_j, o_j)\}$ index representing a set of triples associated with the passages in $\mathbf{C}$, using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), can be obtained, as follows:

$$h^k_{\text{base}}(\mathbf{q}, \mathbf{C} \cup \mathbf{T}) = \text{RRF}\Big(h^k_{\text{base}}(\mathbf{q}, \mathbf{C}), h^k_{\text{base}}(\mathbf{q}, \mathbf{T})\Big), \quad (10)$$

where $h^k_{\text{base}}(\mathbf{q}, \mathbf{C})$ and $h^k_{\text{base}}(\mathbf{q}, \mathbf{T})$ are the passages retrieved from $\mathbf{C}$ and $\mathbf{T}$, after a base retrieval step on each index separately. In this case, each triple $\in h^k_{\text{base}}(\mathbf{q}, \mathbf{T})$ is mapped to its corresponding passage, ensuring that top-$k$ unique passages are returned after considering the triple scores in $\mathbf{T}$.

## B Dataset Choices and Statistics

| | MuSiQue | 2Wiki | HotpotQA |
|---|---|---|---|
| Split Source | IRCoT | IRCoT | HippoRAG |
| # Hops | $2-4$ | 2 | 2 |
| # Documents | 139,416 | 430,225 | 9,221 |
| # Test Queries | 500 | 500 | 1,000 |
| # Chunks ($\mathbf{C}$) | 148,793 | 490,454 | 10,293 |
| # Triples ($\mathbf{T}$) | 1,521,136 | 4,993,637 | 122,492 |
| Av. # $\mathbf{T}/\mathbf{C}$ | 10.2 | 10.2 | 11.9 |

Table 5: Dataset characteristics and preprocessing statistics, where triples are extracted from chunks, and Av. # $\mathbf{T}/\mathbf{C}$ represents the average number of triples per chunk.

Table 5 serves as a summary of various facts and statistics related to the employed datasets and the chunking and triple extraction process introduced in Section 3.

**Reasoning behind dataset split choices** For MuSiQue and 2Wiki, we use the data provided by Trivedi et al., including the full corpus and subsampled test cases for each dataset. To limit the experimental cost for HotpotQA, we follow Gutierrez et al. setting where both the corpus and test split are smaller than IRCoT's counterpart.

**Reasoning behind retrieval metrics** Our evaluation employs recall at ranks 5, 10, and 15 (R@5, R@10, R@15). While previous work like HippoRAG evaluate R@2, we choose higher rank thresholds since many questions in MuSiQue require information from more than two documents. Additionally, given modern LLMs' expanding context length capabilities (Ding et al., 2024), examining recall beyond R@5 (HippoRAG's highest evaluated rank) provides valuable insights. Following IRCoT's approach, we measure up to R@15 and include R@10 as an intermediate point, offering a comprehensive view of model performance across retrieval depths.

## C More Implementation Details

### C.1 Baselines Details

We implement all proposed approaches using Elasticsearch[4]. For SBERT, we employ the `all-mpnet-base-v2` model with approximate k-nearest neighbours and cosine similarity for vector comparisons. In IRCoT experiments, we evaluate both ColBERTv2 and BM25 retrievers — ColBERTv2 for alignment with HippoRAG's baselines, and BM25 for consistency with the original IRCoT implementation.

For all multi-step approaches, including ours, we follow Gutierrez et al. with respect to the maximum number of retrieval iterations, which vary based on the hop requirements of each dataset. Thus, we use a maximum of 4 iterations for MuSiQue and 2 iterations for HotpotQA and 2Wiki.

### C.2 GEAR Details

GEAR involves several hyperparameters, such as the beam size inside graph expansion. We randomly sampled 500 questions from the MuSiQue development set, which we ensure not to overlap with the relevant test set. We select our hyperparameters based on this sample without performing a grid search across all possible configurations. Our goal is to demonstrate our method is ability to achieve state-of-the-art results without extensive parameter tuning. We acknowledge that a more thorough hyperparameter tuning may result in further improvements.

The initial retrieval phase utilises the chunks index $\mathbf{C}$ as the information source, while leaving the triple index $\mathbf{T}$ unused. Our graph expansion component implements beam search with length 2, width 10, and 100 neighbours per beam. The hyperparameter $\gamma$ employed in diverse triple beam search is set to twice the beam search width. For the scoring function, we use the cosine similarity score and the SBERT embedding model.

For the single-step configurations (i.e. any base retriever with NaiveGE or SyncGE), we set the base retriever's maximum number of returned chunks to match our evaluation recall threshold. With the multi-step setup, we maintain a consistent maximum of 10 retrieved chunks before knowledge synchronisation for the purpose of matching IRCoT's implementation. While this 10-chunk limitation applies to individual retrieval rounds, please note that the total number of accessible chunks can exceed this threshold through graph expansion and multiple GEAR iterations.

**`passageLink` Details** We use `passageLink` to link each triple $t_j \in \mathcal{G}^{(n)}$ to its corresponding passages in $\mathbf{C}$ by running a retrieval step as follows:

$$\mathbf{C}_{t_j} = h^k_{\text{base}}\left(t_j, \mathbf{C} \cup \mathbf{T}\right) \tag{11}$$

[4] https://www.elastic.co/

where $j \in \{1, \ldots, |\mathcal{G}^{(n)}|\}$ and $h^k_{\text{base}}(t_j, \mathbf{C} \cup \mathbf{T})$ is the RRF of passages returned by both $\mathbf{T}$ and $\mathbf{C}$ when queried with $t_j$ (as defined in Eq. 10).

## D Compatibility with Open-weight Models

**GEAR Results** As shown in Table 6, we evaluate GEAR using popular 7-8B parameter open-weight models, comparing them against a closed-source alternative. On HotpotQA, Llama-3.1-7B surpasses the closed-source alternative, achieving higher recall rates at R@10 and R@15. For MuSiQue and 2Wiki, while the closed-source model maintains a slight superior edge in performance, the margin is narrow. Importantly, all tested open-weight models consistently outperform the previous state-of-the-art, HippoRAG w/IRCoT. This decouples GEAR from the need to use closed-source models, suggesting that state-of-the-art multi-step retrieval can be achieved using more accessible models.

**Diverse Beam Search Results** Expanding upon Table 4, Table 7 demonstrates that diverse beam search consistently improves retrieval performance across both closed-source and open-weight models when using our proposed Hybrid + SyncGE setup. This further confirms the broader applicability of this approach.

## E Correlation between Question Hops and Agent Iterations

The left panel in Figure 5 demonstrates that the median stopping iteration remains consistently at 1 across all hop counts. Additionally, the upper quartile shows a clear upward trend as the number of hops increases. This suggests greater variability in processing time for more complex questions. The right panel illustrates two concurrent trends: as the question hop count increases, the number of questions in the dataset decreases, and the mean number of iterations GEAR requires to determine question answerability increases. This pattern indicates that higher-hop questions not only appear less frequently but also typically demand more computational effort to process.

## F Qualitative Analysis

Table 8 showcases some query instances where GEAR achieves perfect recall in a single iteration, while HippoRAG w/ IRCoT achieves lower recall and consumes all available iterations. The presented examples illustrate how GEAR's Gist Memory $\mathcal{G}^{(n)}$ precisely captures the essential information needed to answer MuSiQue's queries, maintaining the appropriate level of granularity without including superfluous details. In contrast, HippoRAG w/ IRCoT struggles to retrieve crucial information—whether due to limitations in its triple extraction step or retriever functionality—such as the exact population of Venice, which is necessary for accurate responses. Furthermore, the verbose nature of IRCoT's thought process component contrasts with GEAR's streamlined approach. The lack of such verbose component in our approach contributes to the fact that GEAR requires fewer LLM tokens than the competition, as explained in subsection 8.3.

## G Increasing Number of Agent Iterations

Figure 6 expands upon the analysis shown in Figure 2 by evaluating retrieval performance over 20 iterations, rather than the initial 4 iterations. The results demonstrate a consistent pattern across all methods: retrieval performance stabilises after approximately 4 iterations, with no substantial improvements or degradation in subsequent iterations. While some minor fluctuations occur beyond this point, they are negligible.

This performance plateau can be attributed to two key factors. First, the query re-writing mechanisms in all investigated approaches struggle to generate effective subsequent queries. Second, our analysis has identified several cases of unanswerable queries within MuSiQue's answerable subset. A representative example is provided in Table 9.

## H Comparison of Triple Extraction Prompting Strategies

HippoRAG employs a sequential approach to triple extraction: it first identifies named entities from a text chunk, and then uses these entities to guide triple extraction in a second step. In contrast, our method extracts both entities and triples simultaneously. Table 10 shows that both approaches achieve comparable retrieval performance across all datasets, with each method excelling in different scenarios. These results validate that joint entity and triple extraction can match the effectiveness of sequential extraction while reducing the number of required processing steps.

| | LLM | MuSiQue | | | 2Wiki | | | HotpotQA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 |
| **Closed-source** | GPT-4o mini | **58.4** | **67.6** | **71.5** | **89.1** | **95.3** | **95.9** | **93.4** | 96.8 | 97.3 |
| **Open-weight** | Llama-3.1-7B | 52.4 | 62.3 | 66.7 | 81.6 | 91.0 | 93.7 | 92.2 | **97.4** | **98.1** |
| | Qwen-2.5-8B | 53.7 | 63.7 | 66.7 | 85.9 | 91.6 | 93.0 | 91.7 | 96.2 | 96.9 |

Table 6: Retrieval performance of GEAR across different closed-source and open-weight models on MuSiQue, 2Wiki and HotpotQA. Results are reported using Recall@$k$ (R@$k$) metrics for $k \in \{5, 10, 15\}$, showing the percentage of questions for which the correct entries are found within the top-$k$ retrieved passages. The included open-weight models are Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, and the closed-source model is GPT-4o mini.

| | | MuSiQue | | | 2Wiki | | | HotpotQA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 |
| GPT-4o mini | w/ diversity | **48.7** | **57.7** | **61.2** | **72.6** | **80.9** | **82.4** | **87.4** | **93.3** | **95.2** |
| | w/o diversity | 47.0 | 53.9 | 58.4 | 68.2 | 76.0 | 77.4 | 85.0 | 92.2 | 94.3 |
| Llama-3.1-8B-Instruct | w/ diversity | **46.2** | **54.3** | **57.4** | **69.1** | **78.1** | **81.6** | **87.3** | **92.8** | **95.1** |
| | w/o diversity | 44.9 | 52.7 | 55.0 | 66.9 | 75.9 | 78.2 | 85.0 | 91.7 | 94.4 |

Table 7: Retrieval performance of the Hybrid + SyncGE method with different LLMs for the `read` step (see Eq. 1) w/ and w/o diversity for triple beam search. Results are reported using Recall@$k$ (R@$k$) metrics for $k \in \{5, 10, 15\}$, showing the percentage of questions for which the correct entries are found within the top-$k$ retrieved passages.

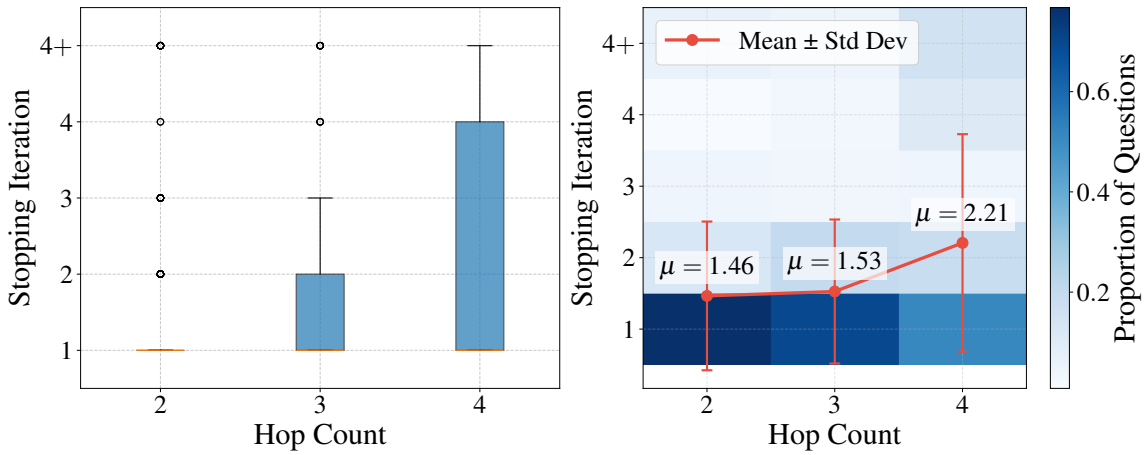


Figure 5: Analysis of the relationship between the number of hops in questions and the required number of agent iterations on the MuSiQue dataset. For each hop count, we analyse the number of iterations required by GEAR to determine question answerability. The maximum iteration limit was set to 4, with "4+" indicating cases where the agent could not determine answerability within this limit. The visualization presents two complementary perspectives on the same data: the left panel shows a box plot emphasizing the median and distribution of stopping iterations, while the right panel focuses on the mean number of iterations across different hop counts.

| Query | GEAR | HippoRAG w/ IRCoT |
|---|---|---|
| What team is the player who led the league in sacks in 2017 a member of? | ("Chandler Jones", "team", "Arizona Cardinals")<br>("Chandler Jones", "led the league in sacks in", "2017") | To answer the question about the player who led the league in sacks in 2017, I need to identify the player and their team from external sources, as the provided documents do not contain that specific information |
| Who is the CEO of FleetBoston Financial's purchaser? | ("FleetBoston Financial", "purchaser", "Bank of America")<br>("Bank of America", "CEO", "Brian Moynihan") | The purchaser of FleetBoston Financial is Bank of America. To find the CEO of Bank of America, I need to look for information about the current leadership of the bank. Next, I will check for the current CEO of Bank of America to answer the question accurately. |
| What UK label was purchase by the company that, along with the distributor of The Cube, and ABC, is the other major New York-based broadcaster? | ("CBS", "purchased", "Oriole Records")<br>("CBS", "is a major broadcaster based in", "New York")<br>("NBC", "is a major broadcaster based in", "New York")<br>("The Cube", "was distributed by", "NBC") | The original network of Undercovers is NBC, which is one of the major broadcasters based in New York. Now, I need to find out which UK label was bought by NBC ... |
| What was the 2018 population of the Italian city that's underwater? | ("Venice", "population in 2018", "260,897") | The Italian city that is underwater is Krag, British Columbia, which is a ghost town... |

Table 8: Comparison of MuSiQue queries where GEAR achieves 100% recall at R@15 in a single iteration, while HippoRAG w/ IRCoT shows lower performance despite using all four available iterations. Cell colors indicate recall performance: green for 100% recall, red for 0% recall, and yellow for any intermediate value. Cell values in GEAR represent the proximal triples stored in the Gist Triple Memory. Cell values in HippoRAG w/ IRCoT represent IRCoT's thought process.

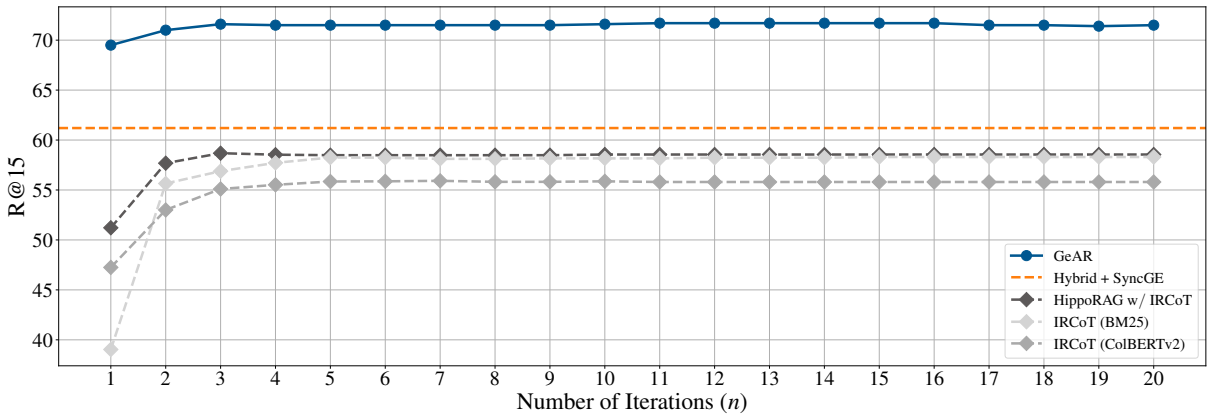


Figure 6: Evolution of R@15 over 20 iterations on MuSiQue. Recall is computed at each iteration using the cumulative set of retrieved documents, with prior recall values carried forward for questions that terminated in earlier iterations. The horizontal line indicates the single-step performance of Hybrid + SyncGE.

| | |
|---|---|
| **Question** | Who did the producer of Big Jim McLain play in True Grit? |
| **Gold Passages** | 1. Big Jim McLain: Big Jim McLain is a 1952 political thriller film starring John Wayne and James Arness as HUAC investigators. |
| | 2. True Grit is a 1969 American western film. It is the first film adaptation of Charles Portis' 1968 novel of the same name. The screenplay was written by Marguerite Roberts. The film was directed by Henry Hathaway and starred Kim Darby as Mattie Ross and John Wayne as U.S. Marshal Rooster Cogburn. Wayne won his only Academy Award for his performance in this film and reprised his role for the 1975 sequel Rooster Cogburn. |
| **Comment** | No information about who was the producer of Big Jim McLain is provided in the gold passages |

Table 9: Example of a query from MuSiQue that in not answerable solely based on the provided gold passages.

| | | **MuSiQue** | | | **2Wiki** | | | **HotpotQA** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 | R@5 | R@10 | R@15 |
| **HippoRAG** | original prompt | **41.9** | 46.9 | 51.1 | **75.4** | **83.5** | **86.9** | 79.7 | 88.4 | 91.4 |
| | our prompt | 41.0 | **47.0** | **51.4** | 75.1 | 83.2 | 86.4 | **79.8** | **89.0** | **92.4** |
| **HippoRAG w/ IRCoT** | original prompt | **49.9** | **56.4** | **59.3** | 81.5 | 90.2 | 92.3 | **90.2** | **94.7** | 95.8 |
| | our prompt | 48.8 | 54.5 | 58.9 | **82.9** | **90.6** | **93.0** | 90.1 | **94.7** | **95.9** |

Table 10: Retrieval performance comparison between HippoRAG's sequential triple extraction method and our joint extraction approach across three datasets.

# I Prompts

## I.1 Offline Prompts

Reader

**# Instruction**

Your task is to construct an RDF (Resource Description Framework) graph from the given passages and named entity lists.
Respond with a JSON list of triples, with each triple representing a relationship in the RDF graph.
Pay attention to the following requirements:
- Each triple should contain at least one, but preferably two, of the named entities in the list for each passage.
- Clearly resolve pronouns to their specific names to maintain clarity.

Convert the paragraph into a JSON dict containing a named entity list and a triple list.

**# Demonstration #1**

Paragraph:
‘‘‘
Magic Johnson

After winning a national championship with Michigan State in 1979, Johnson was selected first overall in the 1979 NBA draft by the Lakers, leading the team to five NBA championships during their "Showtime" era.
‘‘‘
{{"named_entities": ["Michigan State", "national championship", "1979", "Magic Johnson",
"National Basketball Association", "Los Angeles Lakers", "NBA Championship"]}}
{{
"triples": [
("Magic Johnson", "member of sports team", "Michigan State"),
("Michigan State", "award", "national championship"),
("Michigan State", "award date", "1979"),
("Magic Johnson", "draft pick number", "1"),
("Magic Johnson", "drafted in", "1979"),
("Magic Johnson", "drafted by", "Los Angeles Lakers"),
("Magic Johnson", "member of sports team", "Los Angeles Lakers"),
("Magic Johnson", "league", "National Basketball Association"),
("Los Angeles Lakers", "league", "National Basketball Association"),
("Los Angeles Lakers", "award received", "NBA Championship"),
]
}}
‘‘‘

**# Demonstration #2**

Paragraph:
‘‘‘
Elden Ring

Elden Ring is a 2022 action role-playing game developed by FromSoftware. It was directed by Hidetaka Miyazaki with worldbuilding provided by American fantasy writer George R. R. Martin.
‘‘‘
{{"named_entities": ["Elden Ring", "2022", "Role-playing video game", "FromSoftware", "Hidetaka Miyazaki", "United States of America", "fantasy", "George R. R. Martin"]}}
{{
"triples": [
("Elden Ring", "publication", "2022"),
("Elden Ring", "genre", "action role-playing game"),
("Elden Ring", "publisher", "FromSoftware"),
("Elden Ring", "director", "Hidetaka Miyazaki"),
("Elden Ring", "screenwriter", "George R. R. Martin"),
("George R. R. Martin", "country of citizenship", "United States of America"),
("George R. R. Martin", "genre", "fantasy"),
]
}}

**# Input**

Convert the paragraph into a JSON dict, it has a named entity list and a triple list.

Paragraph:
‘‘‘
{**wiki_title**}

{**passage**}

### I.2 Online Retrieval Prompts

The blue-highlighted portions of the Reader prompt below indicate additional text that is only required when the Gist Memory $\mathcal{G}^{(n)}$ is active. When Gist Memory is inactive, these blue sections should be omitted, and the {triples} parameter should be left empty.

**Reader with and without Gist Memory**

Your task is to find facts that help answer an input question.

You should present these facts as knowlege triples, which are structured as ("subject", "predicate", "object").
Example:
Question: When was Neville A. Stanton’s employer founded?
Facts: ("Neville A. Stanton", "employer", "University of Southampton"), ("University of Southampton", "founded in", "1862")

Now you are given some documents:
{**docs**}

Based on these documents and some preliminary facts provided below,
find additional supporting fact(s) that may help answer the following question.

Note: if the information you are given is insufficient, output only the relevant facts you can find.

Question: {**query**}
Facts: {**triples**}

**Reasoning for Termination**

**# Task Description:**
You are given an input question and a set of known facts:
Question: {**query**}
Facts: {**triples**}

Your reply must follow the required format:
1. If the provided facts contain the answer to the question, your should reply as follows:
Answerable: Yes
Answer: ...

2. If not, you should explain why and reply as follows:
Answerable: No
Why: ...

**# Your reply:**

**Query Re-writing**

**# Task Description:**
You will be presented with an input question and a set of known facts.
These facts might be insufficient for answering the question for some reason.
Your task is to analyze the question given the provided facts and determine what else information is needed for the next step.

**# Example:**

Question: What region of the state where Guy Shepherdson was born, contains SMA Negeri 68?
Facts: ("Guy Shepherdson", "born in", "Jakarta")
Reason: The provided facts only indicate that Guy Shepherdson was born in Jakarta, but they do not provide any information about the region of the state that contains SMA Negeri 68.
Next Question: What region of Jakarta contains SMA Negeri 68?

**# Your Task:**
Question: {**query**}
Facts: {**triples**}
Reason: {**reason**}

Next Question:

### I.3 Online Question Answering Prompts

The following prompt with retrieved passages combines the QA generation prompts from Gutierrez et al. and Wang et al.. For the variation without retrieved passages, we omit the preamble and only include the instruction, highlighted in purple .

**Retrieved Passages with In-context Example**

As an advanced reading comprehension assistant, your task is to analyze text passages and corresponding questions meticulously, with the aim of providing the correct answer.
==========
For example:
==========
Wikipedia Title: Edward L. Cahn
Edward L. Cahn (February 12, 1899 – August 25, 1963) was an American film director.

Wikipedia Title: Laughter in Hell
Laughter in Hell is a 1933 American Pre-Code drama film directed by Edward L. Cahn and starring Pat O'Brien. The film's title was typical of the sensationalistic titles of many Pre-Code films. Adapted from the 1932 novel of the same name buy Jim Tully, the film was inspired in part by "I Am a Fugitive from a Chain Gang" and was part of a series of films depicting men in chain gangs following the success of that film. O'Brien plays a railroad engineer who kills his wife and her lover in a jealous rage and is sent to prison. The movie received a mixed review in "The New York Times" upon its release. Although long considered lost, the film was recently preserved and was screened at the American Cinematheque in Hollywood, CA in October 2012. The dead man's brother ends up being the warden of the prison and subjects O'Brien's character to significant abuse. O'Brien and several other characters revolt, killing the warden and escaping from the prison. The film drew controversy for its lynching scene where several black men were hanged. Contrary to reports, only blacks were hung in this scene, though the actual executions occurred off-camera (we see instead reaction shots of the guards and other prisoners). The "New Age" (an African American weekly newspaper) film critic praised the scene for being courageous enough to depict the atrocities that were occurring in some southern states.

Wikipedia Title: Theodred II (Bishop of Elmham)
Theodred II was a medieval Bishop of Elmham. The date of Theodred's consecration unknown, but the date of his death was sometime between 995 and 997.

Wikipedia Title: Etan Boritzer
Etan Boritzer( born 1950) is an American writer of children 's literature who is best known for his book" What is God?" first published in 1989. His best selling" What is?" illustrated children's book series on character education and difficult subjects for children is a popular teaching guide for parents, teachers and child- life professionals. Boritzer gained national critical acclaim after" What is God?" was published in 1989 although the book has caused controversy from religious fundamentalists for its universalist views. The other current books in the" What is?" series include What is Love?, What is Death?, What is Beautiful?, What is Funny?, What is Right?, What is Peace?, What is Money?, What is Dreaming?, What is a Friend?, What is True?, What is a Family?, What is a Feeling?" The series is now also translated into 15 languages. Boritzer was first published in 1963 at the age of 13 when he wrote an essay in his English class at Wade Junior High School in the Bronx, New York on the assassination of John F. Kennedy. His essay was included in a special anthology by New York City public school children compiled and published by the New York City Department of Education.

Wikipedia Title: Peter Levin
Peter Levin is an American director of film, television and theatre.

Question: When did the director of film Laughter In Hell die?
Answer: August 25, 1963.
==========
Given the following text passages and questions, please present a concise, definitive answer, devoid of additional elaborations, and of maximum length of 6 words.
==========

Wikipedia Title : {**title**} {**text**} `for each retrieved passage` ...
Question: {**question**}

Answer:

### No Retrieved Passages

Given the following question, please present a concise, definitive answer, devoid of additional elaborations, and of maximum length of 6 words.

Question: {**question**}

Answer: