

# Towards Global Retrieval Augmented Generation: A Benchmark for Corpus-Level Reasoning

Qi Luo<sup>1</sup>, Xiaonan Li<sup>1</sup>, Tingshuo Fan<sup>1,2</sup>, Xinchi Chen<sup>3,1</sup>, Xipeng Qiu<sup>1</sup>

<sup>1</sup>School of Computer Science, Fudan University

<sup>2</sup>School of Foreign Languages and Literature, Fudan University

<sup>3</sup>Institute of Trustworthy Embodied AI, Fudan University

qluo22@m.fudan.edu.cn

## Abstract

Retrieval-augmented generation (RAG) has emerged as a leading approach to reducing hallucinations in large language models (LLMs). Current RAG evaluation benchmarks primarily focus on what we call local RAG: retrieving relevant chunks from a small subset of documents to answer queries that require only localized understanding within specific text chunks. However, many real-world applications require a fundamentally different capability—global RAG—which involves aggregating and analyzing information across entire document collections to derive corpus-level insights (e.g. “What are the top 10 most cited papers in 2023?”). In this paper, we introduce GlobalQA—the first benchmark specifically designed to evaluate global RAG capabilities, covering four core task types: counting, extremum queries, sorting, and top-k extraction. Through systematic evaluation across different models and baselines, we find that existing RAG methods perform poorly on global tasks, with the strongest baseline achieving only 1.51 F1 score. To address these challenges, we propose GlobalRAG, a multi-tool collaborative framework that preserves structural coherence through chunk-level retrieval, incorporates LLM-driven intelligent filters to eliminate noisy documents, and integrates aggregation modules for precise symbolic computation. On the Qwen2.5-14B model, GlobalRAG achieves 6.63 F1 compared to the strongest baseline’s 1.51 F1, validating the effectiveness of our method.

## 1 Introduction

Retrieval-augmented generation (RAG) has been proposed as a solution to mitigate hallucination and knowledge limitations in large language models (LLMs) (Lewis et al., 2021; Li et al., 2025; Zhang et al., 2025; Rawte et al., 2023; Wang et al., 2025). Its core mechanism lies in retrieving relevant chunks from a corpus to inject factual knowl-

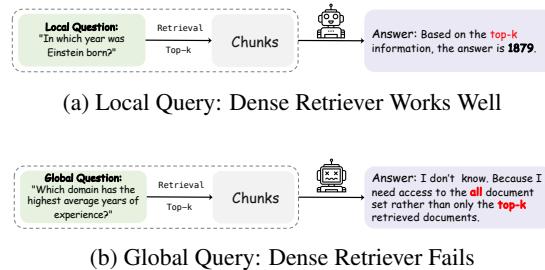


Figure 1: Why Dense Retriever Fails on Global Queries: (a) Local Query: The answer can be found in specific documents. Dense retriever ranks all documents and selects top-k, which contains the relevant information. (b) Global Query: The answer requires information from *all* documents. However, dense retriever only returns top-k ranked documents, missing critical information scattered across the entire corpus.

edge into the generation process, thereby improving the reliability of LLM outputs(Gao et al., 2023).

Current RAG systems operate at the **chunk level**, where the goal is to retrieve specific text segments from documents to answer queries. This approach are designed for answering questions that require **localized information**—for instance, determining when Einstein was born or identifying the capital of France. Current evaluation benchmarks exclusively focus on such local RAG tasks. Single-hop datasets like Natural Questions (Kwiatkowski et al., 2019) and MS MARCO (Bajaj et al., 2018) test the extraction of facts from individual documents, while multi-hop datasets such as HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020) evaluate reasoning across a small number of connected documents.

However, many real-world applications require **corpus-level** operations that aggregate information across entire document collections, as show in figure 1a. Systems for multi-document question answering (Wang et al., 2023b) demonstrate the need for analytical capabilities that extract and combine facts across large document corpora. Recent

OpenAI systems (OpenAI et al., 2024) have been developed to conduct multi-step research across hundreds of sources, highlighting growing practical demand for such capabilities. As shown in figure 1b, these **global RAG** tasks—such as “Which domain has the highest average years of experience?”—cannot be answered by retrieving individual documents. Instead, they require traversing thousands of documents, extracting comparable attributes, and performing corpus-wide analysis. GraphRAG (Edge et al., 2025) shows that baseline RAG performs poorly on such global queries, yet no systematic benchmark exists to evaluate these corpus-level reasoning capabilities.

To address this gap, we construct **GlobalQA**, the first benchmark explicitly specifically to evaluate **global RAG** abilities. Since global tasks often require integrating and analyzing a large number of documents, GlobalQA defines four task types: - Counting: computing the number of entities that satisfy given conditions; - Extremum queries: identifying entities with extreme attributes; - Sorting: ranking entities on a global scale; - Top- $k$  extraction: retrieving the top- $k$  entities after ranking.

Through systematic evaluation on GlobalQA, we find that existing RAG methods (Lewis et al., 2021; Trivedi et al., 2023; Luo et al., 2025) achieve at most 1.51 F1 score on global tasks. We manually analyzed the error cases and identified three fundamental limitations that contribute to the poor performance:

**Issue 1: Fixed-granularity chunking disrupts document integrity.** Mainstream RAG systems split documents into fixed-length chunks (e.g., 512 tokens) (Lewis et al., 2021), which mechanically breaks the inherent structure of documents (Luo et al., 2024). For example, after chunking a research paper, metadata (e.g., publication year, conference name) and citation counts may be separated into different segments, preventing the system from correctly linking attributes to values, often causing double-counting or omissions.

**Issue 2: Dense retrieval returns semantically relevant but factually irrelevant noise.** Dense retrievers frequently return many documents that are semantically related but do not contain the required answer. Prior work (Pu et al., 2024; Wang et al., 2023a) shows that such high-scoring yet irrelevant documents can harm LLM performance. For the above query, the retriever may return documents discussing “citation analysis methods” or “AI conference overviews,” but not the actual citation data.

These noisy documents consume the limited context window and distract the model from attending to the truly relevant document.

**Issue 3: Inherent limitations of LLMs in numerical computation.** Recent studies (Schwartz et al., 2024) demonstrate that LLMs generally perform poorly on numerical reasoning tasks. Even when provided with all necessary information, models often make mistakes in numerical comparisons, omit key items, or produce inconsistent results.

To tackle the above challenges, we propose three technical innovations: **(1) Document-level retrieval:** using entire documents as retrieval units to preserve structural integrity and semantic coherence, ensuring that all relevant attributes can be processed jointly. **(2) Filter:** designing a retrieval–reading–filtering pipeline, where an LLM-driven filter eliminates noisy documents before reasoning, ensuring that inference is based on high-quality documents. **(3) Data processing tools:** introducing auxiliary tools to assist the model in handling numerical and statistical problems, enabling hybrid reasoning that combines language understanding with symbolic computation.

The main contributions of this paper can be summarized as follows:

1. First, we introduce GlobalQA, establishing the first systematic benchmark for evaluating global RAG capabilities.
2. Second, through comprehensive experiments, we demonstrate that existing RAG approaches achieve only 1.51 F1 score on global tasks, revealing fundamental architectural limitations that cannot be addressed through incremental improvements.
3. Third, we present GlobalRAG, which achieves 6.63 F1 score on GlobalQA with Qwen2.5-14B—a 5 point improvement over the strongest baseline.

The dataset is publicly available at [GlobalQA](#).

<sup>1</sup>

## 2 Related Work

### 2.1 Retrieval-Augmented Generation Methods

The development of retrieval-augmented generation (RAG) can be broadly categorized into two

<sup>1</sup><https://huggingface.co/datasets/QiiLuo0/GlobalQA>

paradigms: unstructured retrieval and structured retrieval.

**Unstructured Retrieval Methods:** Unstructured retrieval directly indexes and searches over raw documents, and is the mainstream paradigm in current RAG systems. Dense document Retrieval (DPR) (Karpukhin et al., 2020) leverages a dual-encoder architecture to learn dense representations of queries and documents, achieving success in open-domain QA. Contriever (Izacard et al., 2022) further introduces an unsupervised contrastive learning framework to enhance the generalization ability of retrievers. RETRO (Borgeaud et al., 2022) integrates retrieval deeply into the Transformer architecture, achieving tighter coupling between retrieval and generation. Recent studies focus on optimizing retrieval strategies and reasoning processes. Self-RAG (Asai et al., 2023) incorporates a self-reflection mechanism, using special tokens to control retrieval timing and verify retrieved content. FLARE (Jiang et al., 2023) proposes an active retrieval strategy that dynamically triggers retrieval based on generation confidence. IRCoT (Trivedi et al., 2023) interleaves chain-of-thought reasoning with retrieval, demonstrating advantages in multi-hop reasoning tasks. Recent methods have begun exploring reinforcement learning to optimize multi-step retrieval processes (Chen et al., 2025; Jin et al., 2025). However, these methods are primarily designed for optimizing local information retrieval and face fundamental limitations when large-scale information aggregation is required.

**Structured Retrieval Methods:** Structured retrieval organizes information by constructing knowledge graphs or hierarchical indexes, aiming to capture semantic relationships between documents. KG-RAG (Sanmartin, 2024) transforms unstructured text into knowledge triples and performs multi-hop reasoning through graph traversal. GraphRAG (Edge et al., 2025) builds multi-level graph structures via community detection, marking the first attempt to address global query tasks. RAPTOR (Sarthi et al., 2024) constructs tree-structured indexes using recursive summarization, supporting multi-granularity retrieval. More recent graph-based approaches further enhance knowledge representation. HippoRAG (Gutiérrez et al., 2025) simulates hippocampal memory mechanisms and optimizes retrieval paths through personalized PageRank. HyperGraphRAG (Luo et al., 2025) extends

graph structures into hypergraphs, where hyperedges can connect multiple nodes to better model complex multi-relational structures. However, information loss during graph construction and the locality of graph traversal still limit their performance on global RAG tasks.

## 2.2 Evaluation Benchmarks for RAG

**Single-Hop Retrieval Datasets:** Single-hop retrieval datasets (Kwiatkowski et al., 2019; Bajaj et al., 2018; Joshi et al., 2017; Mallen et al., 2023) mainly evaluate the system’s ability to locate a single relevant piece of information. Natural Questions (NQ) (Kwiatkowski et al., 2019) contains real queries from Google Search, where answers typically lie in a single Wikipedia document. MS MARCO (Bajaj et al., 2018) provides 1M query-document pairs and has become the standard resource for training dense retrievers. TriviaQA (Joshi et al., 2017) consists of 95K QA pairs, emphasizing factual knowledge retrieval. A common limitation of these datasets is that answers are restricted to a single document span, making them incapable of evaluating information aggregation.

**Multi-Hop Reasoning Datasets:** Multi-hop datasets (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022) require systems to integrate information across multiple documents for reasoning. HotpotQA (Yang et al., 2018) includes 113K questions requiring two-hop reasoning, with annotated supporting facts. 2WikiMultihopQA (Ho et al., 2020) ensures unique reasoning paths through template-based generation. MuSiQue (Trivedi et al., 2022) extends to 4-hop reasoning and provides decomposed annotations from single-hop to 4-hop reasoning. However, existing multi-hop datasets still focus on *multi-hop chains*, i.e., connecting a small set of documents through a limited reasoning sequence. In contrast, global RAG tasks require *parallel aggregation* of information from a large number of documents. This qualitative difference prevents existing benchmarks from accurately reflecting a system’s ability to perform corpus-level reasoning. Table 1 provides a detailed comparison between GlobalQA and existing datasets.

## 3 The GlobalQA Dataset

Existing retrieval-augmented QA datasets primarily focus on local, single-document retrieval tasks, making it difficult to assess a system’s ability to perform corpus-level information aggregation. To

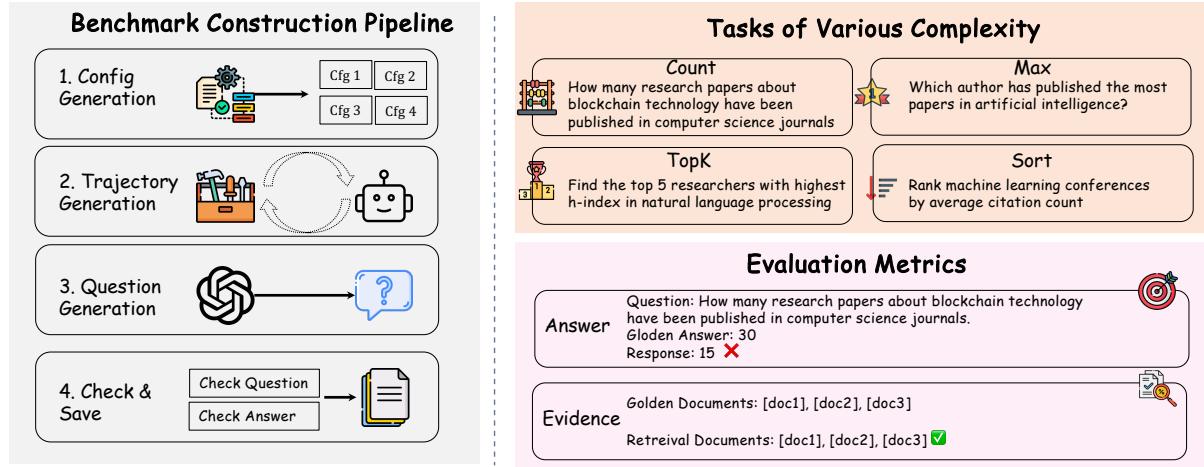


Figure 2: GlobalQA benchmark overview: construction pipeline (left), task examples with various complexity (top-right), and evaluation metrics (bottom-right).

fill this gap, we introduce **GlobalQA**—a benchmark specifically designed to evaluate the corpus-level aggregation capability of retrieval-augmented systems. In this section, we present the dataset overview and construction pipeline, followed by a comparison with existing datasets and a discussion of evaluation metrics.

### 3.1 Overview

Table 3 summarizes the dataset statistics. GlobalQA contains more than 13,000 question-answer pairs, constructed on a corpus of more than 2,000 real-world resumes across 23 professional domains. The dataset exhibits the following characteristics:

**Task diversity:** GlobalQA covers four core task types—**Counting** (16.7%), which computes the number of entities satisfying given conditions; **Extremum queries** (Max/Min, 33.9%), which identify entities with extreme attributes; **Sorting** (16.3%), which ranks entities globally; and **Top- $k$  extraction** (33.9%), which retrieves the top- $k$  entities after sorting. These tasks comprehensively evaluate the model’s ability in statistics, comparison, ranking, and filtering.

**Corpus-Level document Requirements:** Unlike existing benchmarks where answers typically reside in 1–4 documents, GlobalQA queries necessitate traversing significantly larger document sets. As shown in Figure 3 (middle), 42.6% of queries require documents from more than 20 documents, with the maximum reaching 50 documents per query. Specifically: 18.1% require 2–5 documents, 13.4% require 5–10, 25.9% require 10–20, and the remaining 42.6% demand over 20 docu-

ments. This distribution ensures that models cannot rely on heuristics designed for few-document scenarios and must develop genuine corpus-wide reasoning strategies.

**Keyword distribution:** Figure 3 (right) visualizes the query keyword distribution. Query keywords are concentrated on high-frequency concepts such as “experience”, “management”, “skills”, and “project” which are aligned with the structure of the resume corpus. The long-tail distribution of keywords covers both common scenarios and rare queries, allowing the dataset to evaluate model robustness in less frequent cases.

### 3.2 Construction Pipeline

Mainstream QA datasets(Yang et al., 2018; Ho et al., 2020) are typically constructed by starting from natural language questions and then manually annotating answers. This approach works for local retrieval tasks, where answers often reside in a small number of documents, making human verification feasible. However, global RAG tasks require analyzing and aggregating hundreds of documents for statistics, ranking, or set operations—making manual verification impractical and inconsistent.

To address this challenge, GlobalQA adopts a **reverse construction strategy**: As shown in Figure 2, we first programmatically design query trajectories, then allow an agent to execute these trajectories step by step to obtain deterministic answers, and finally generate natural language questions based on the completed trajectories. Since each trajectory execution is deterministic—guaranteed by the correctness of sub-trajectory inputs and outputs—the

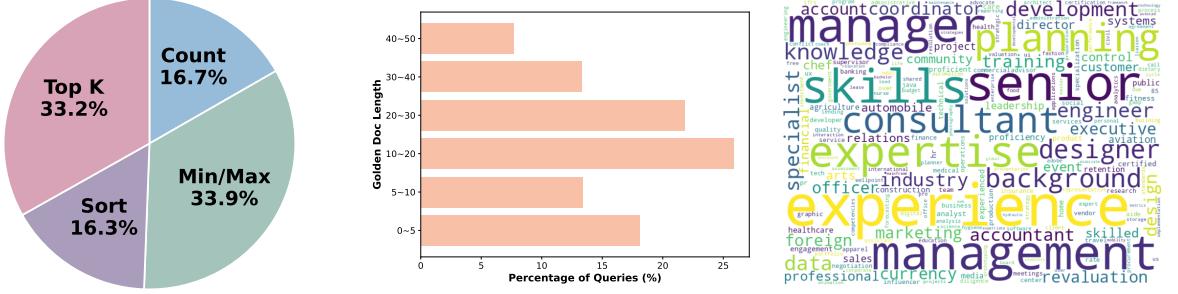


Figure 3: Statistical analysis of the GlobalQA dataset. **Left:** distribution of task types. **Middle:** distribution of the number of documents per query. **Right:** keyword distribution.

overall reasoning chain yields reliable answers, reducing the annotation requirements for large-scale corpus analysis.

The construction pipeline consists of three key stages:

#### Stage 1: Automatic configuration generation

The system automatically generates query configurations, including: - randomly sampling 2–5 trajectory steps to ensure diverse complexity, - selecting query domains from a predefined knowledge base, - designing retrieval steps combining keyword- and semantics-based queries, - generating set operations (intersection, union) for document aggregation, - assigning task types (counting, sorting, top- $k$  extraction), - converting configurations into natural language queries via templates.

**Stage 2: Trajectory execution and data generation** *Document retrieval and verification:* Based on the generated configuration, the system retrieves relevant documents from the corpus. To ensure accuracy and completeness, we employ the DeepSeek model to traverse the entire corpus, validating and supplementing documents potentially missed by the retriever. This ensures each trajectory is grounded in a complete and accurate document set. *Trajectory data generation:* The system applies pre-defined set operations, computes relevance scores, executes counting or sorting tasks, assembles structured trajectory data, and filters out low-quality trajectories with abnormal document counts.

**Stage 3: Natural language generation** Structured trajectories are converted into natural language questions by parsing key information and logical relations. Standard answers are extracted from execution results, ensuring full consistency among questions, trajectories, and answers.

**Stage 4: Check and Save** To control query difficulty and ensure dataset quality, we filter out queries requiring more than 50 documents. Ad-

ditionally, we remove trajectories with abnormal document distributions or execution inconsistencies, ensuring all retained question-answer pairs have reliable ground truth and consistent reasoning paths.

### 3.3 Comparison with Existing Benchmarks

Table 1 highlights the differences between GlobalQA and existing RAG evaluation datasets. Existing datasets fall into two categories: *single-hop retrieval datasets* (e.g., NQ, MS MARCO, TriviaQA), which mainly test single-document retrieval and simple factoid queries; and *multi-hop datasets* (e.g., HotpotQA, 2WikiMultihop, MuSiQue), which involve multiple documents but remain limited to path-chaining or compositional reasoning over a small set of documents.

In contrast, GlobalQA focuses on queries requiring comprehensive corpus-level analysis and global pattern recognition. Each query averages 2–50 documents, involving both single-hop and multi-hop reasoning. Typical tasks, such as “*Which domain has the longest average experience?*”, demand statistical analysis over the entire corpus—capabilities that existing datasets cannot evaluate.

### 3.4 Evaluation Metrics

We adopt two metrics to evaluate performance:

**F1:** measures the quality of final answers, using the standard token-level F1 score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

**Document F1@k(D-F1@k):** evaluates the coverage of retrieved documents, defined as the F1 score between the retrieved document set and the gold document set:

$$\text{Precision}@k = \frac{|\mathcal{D}_{\text{ret}}^k \cap \mathcal{D}_{\text{gold}}|}{|\mathcal{D}_{\text{ret}}^k|} \quad (2)$$

Dataset	Info Requirement	Docs	Hops	Example
<i>Single-hop datasets</i>				
NQ	chunk-level	1	Single	“When was Einstein born?”
MS MARCO	chunk-level	1	Single	“What is photosynthesis?”
TriviaQA	chunk-level	1	Single	“Capital of France?”
<i>Multi-hop datasets</i>				
HotpotQA	chunk-level	2–10	Multi	“When was director X’s wife born?”
2WikiMultihop	chunk-level	2–4	Multi	“Where did A’s CEO graduate?”
MuSiQue	chunk-level	2–4	Multi	“How old was Y when X occurred?”
<i>Global aggregation dataset</i>				
<b>GlobalQA</b>	<b>Corpus-level</b>	<b>2–50</b>	<b>Both</b>	<b>“Which domain has longest avg experience?”</b>

Table 1: Core differences between GlobalQA and existing RAG evaluation datasets.

$$\text{Recall}@k = \frac{|\mathcal{D}_{\text{ret}}^k \cap \mathcal{D}_{\text{gold}}|}{|\mathcal{D}_{\text{gold}}|} \quad (3)$$

$$\text{F1}@k = \frac{2 \cdot \text{Precision}@k \cdot \text{Recall}@k}{\text{Precision}@k + \text{Recall}@k} \quad (4)$$

## 4 Method

Existing RAG methods face three key challenges in global RAG tasks: (i) document chunking disrupts structural integrity, (ii) noisy retrievals occupy the limited context window, and (iii) LLMs exhibit limited ability in numerical computation. To address these issues, we propose **GlobalRAG**, a multi-tool collaborative training-free framework

### 4.1 Framework Overview

GlobalRAG operates through a three-stage pipeline: *retrieval* → *filtering* → *aggregation*. Given a global query, the system first retrieves complete documents rather than fragmented chunks, then applies LLM-driven filters to remove irrelevant noise, and finally invokes task-specific aggregation tools to derive accurate answers.

### 4.2 Document-Level Retrieval

To preserve document integrity, GlobalRAG treats entire documents as atomic retrieval units rather than arbitrary text chunks.

**Index construction** : For each document  $D_i$  in the corpus, we build a chunk-level index:

$$\mathcal{I} = \{(D_1, \text{emb}(D_1)), \dots, (D_n, \text{emb}(D_n))\} \quad (5)$$

**Retrieval strategy** : Given a query  $q$ , the retriever returns the top- $k$  most similar documents:

$$\mathcal{D}_{\text{ret}} = \text{TopK}(\text{sim}(\text{emb}(q), \text{emb}(D_i)), k) \quad (6)$$

This preserves document structure and metadata.

### 4.3 Document-Level Filter

Dense retrievers return semantically relevant but task-irrelevant documents, introducing substantial noise. To mitigate this, we design a two-stage filtering mechanism. A lightweight LLM is used to precisely determine document relevance:

$$r_i = \text{LLM}_{\text{filter}} \left( \begin{array}{l} \text{“Does document } D_i \\ \text{contain information to answer query } q? \end{array} \right) \quad (7)$$

This step discards irrelevant documents before reasoning, ensuring that inference is based on highly related documents.

### 4.4 Task-Level aggregation Tools

LLMs exhibit systematic deficiencies in large-scale numerical computation, statistical analysis, and precise ranking. When handling corpus-level queries that involve dozens or even hundreds of documents, models often suffer from counting errors, inconsistent numerical comparisons, or unstable sorting. To support numerical computation in global RAG, we design specialized computation modules that leverage symbolic precision to complement neural reasoning.

**Tool design**: We introduce four specialized tools, each tailored to specific global RAG needs. - The **Counting tool** systematically traverses documents, applies deduplication strategies, and accurately enumerates entities for corpus-level counting tasks. - The **Extremum tool** extracts numerical and ordinal attributes, performing exact comparisons without approximation errors to resolve global optimization queries. - The **Sorting tool** extracts comparable metrics from heterogeneous document formats, accounts for scale differences across documents, and guarantees consistent results through deterministic algorithms. - The **Top- $k$  extraction**

tool leverages efficient heap-based algorithms with pruning strategies to identify the top- $k$  entities, balancing global analysis with selective output.

These four tools can be combined to form a computation framework. Together, they provide a reliable foundation for numerical reasoning in GlobalRAG, ensuring both accuracy and consistency in global RAG tasks.

## 5 Experimental Setup

### 5.1 Baseline Methods

To comprehensively evaluate the performance of different retrieval-augmented paradigms on global RAG tasks, we select five representative baseline methods, covering three major technical directions: single-shot retrieval, iterative retrieval, and graph-structured retrieval.

**Single-Shot Retrieval Methods:** StandardRAG (Lewis et al., 2021): The standard retrieval-augmented generation approach, which performs a single round of dense retrieval to obtain the top- $k$  relevant documents as context.

**Iterative Retrieval Methods:** FLARE (Jiang et al., 2023): Dynamically triggers retrieval by monitoring uncertainty during generation. When the model produces tokens with low confidence, generation is paused and additional retrieval is performed. The confidence threshold is set to 0.5. IR-CoT (Trivedi et al., 2023): Integrates chain-of-thought reasoning with interleaved retrieval, where each reasoning step triggers a retrieval operation. While suitable for multi-step reasoning tasks, its rigid structure may limit performance in scenarios requiring parallel information aggregation.

**Graph-Structured Retrieval Methods:** HippoRAGv2 (Gutiérrez et al., 2025): Inspired by hippocampal memory mechanisms, this method constructs a knowledge graph and applies personalized PageRank to optimize retrieval paths. A “memory signal” mechanism is introduced to enhance adaptation to historical query patterns. HyperGraphRAG (Luo et al., 2025): Extends graph structures into hypergraph representations. It builds multi-granularity community structures through entity extraction, relation modeling, and hierarchical aggregation, combining local neighborhood expansion with global community matching for retrieval.

### 5.2 Implementation Details

During dataset construction, we employ DeepSeek-v3 to validate the completeness of retrieved documents for each query trajectory. For evaluation, we use Qwen2.5-Instruct models as the backbone LLM. The retriever is BGE-large-en, and the filter module uses Qwen3-4B. We set the maximum retrieval iterations to 10 and retrieve top-20 documents per query. Temperature is set to 0. All baseline methods use identical retriever and LLM configurations for fair comparison.

## 6 Main Experiments

Table 2 presents the performance comparison between our proposed **GlobalRAG** and multiple baseline methods across different Qwen2.5 model scales (3B, 7B, 14B) (Qwen et al., 2025) and four categories of global RAG tasks. The results show that GlobalRAG consistently outperforms all baselines in terms of average F1 score. On the 14B model, GlobalRAG achieves an average F1 of 6.63 and an average D-F1@20 of 12.01, surpassing the strongest baseline, StandardRAG, by 5.12 and 3.92 points, respectively.

**Comparison with Iterative Retrieval Methods.** On the 14B model, GlobalRAG’s average F1 score is 6.54 points higher than IR-CoT and 6.08 points higher than FLARE. This gap correlates with the predefined retrieval–reasoning workflows of iterative methods, which were originally designed for single-hop or multi-hop factoid QA. Such rigid pipelines constrain flexibility in global RAG scenarios, where the system must dynamically adjust its information acquisition strategy. In contrast, GlobalRAG empowers the model with tool invocation mechanisms, enabling it to autonomously select and combine external resources according to task requirements, thereby unleashing the full reasoning potential of LLMs. Notably, although IR-CoT achieves relatively high D-F1@20 scores (e.g., 10.38 on the 3B model), indicating that it can retrieve more relevant documents, its answer F1 remains much lower than GlobalRAG (1.00 point lower on 3B and 6.54 points lower on 14B). This highlights that merely retrieving more relevant information is insufficient for global RAG tasks—effective integration and processing of the retrieved document are equally crucial. Constrained by its iterative mode, IR-CoT struggles to perform global synthesis over dispersed information.

**Comparison with Graph-Structured Methods.** As a structured method that transforms documents into knowledge graphs, HyperGraphRAG performs poorly on global RAG tasks. Results show that its F1 score does not exceed 0.47 across all model scales, and its D-F1@20 score cannot even be computed (due to failure in returning valid document documents). By contrast, GlobalRAG achieves an average F1 that is 6.54 points higher than HyperGraphRAG on the 14B model. The performance gap correlates with information loss during graph construction. HyperGraphRAG decomposes each document into sets of nodes and edges, establishing relational networks between entities but destroying the structural integrity of documents as unified information units. In global RAG tasks, chunk-level metadata (e.g., document counts, attribute distributions) is crucial, but such information is lost during graph transformation. For instance, in counting tasks, systems must tally the number of documents satisfying certain conditions, yet HyperGraphRAG cannot trace original document boundaries, leading to failure. By contrast, GlobalRAG preserves document-level retrieval, ensuring structural integrity and allowing the model to directly access and manipulate complete document units, thereby achieving breakthrough performance in global RAG tasks.

**Cross-Scale Performance Trends.** As model size increases from 3B to 14B, GlobalRAG shows substantial improvement (average F1 rises from 2.52 to 6.63, a gain of 4.11 points), while other methods exhibit limited or unstable improvements. This demonstrates that GlobalRAG’s multi-tool collaborative framework better leverages the enhanced reasoning capacity of larger LLMs, enabling effective scalability in performance.

## 7 Analysis Experiments

### 7.1 Ablation Study

To systematically evaluate the contribution of each component, we conducted three ablation experiments: (1) replacing chunk-level retrieval with 512-token chunk-based retrieval, (2) removing the LLM-driven filter module while retaining chunk-level retrieval and aggregation tools, and (3) removing both filter and aggregation tool modules to create a chunk-level StandardRAG baseline. Table 3 presents the detailed results.

**Impact of retrieval granularity:** We first replaced the chunk-level retriever with a conventional chunk-based retriever. The results show that this replacement caused the F1 score to drop sharply from 7.27 to 0.4, a decrease of 94.5%. This performance degradation indicates the contribution of chunk-level retrieval in global RAG tasks. Notably, since chunk-based retrieval cannot provide chunk-level boundary information, we did not compute statistics in this setting.

**Progressive contribution of tool modules:** On top of chunk-level retrieval, we further analyzed the impact of each tool module. Removing the filter module led to drops of 0.53 and 1.72 in F1 and D-F1@20, corresponding to relative declines of 7.3% and 14.1%, respectively. This indicates that the filter removes redundant information and improves reasoning precision.

**Necessity of a complete toolchain:** When both the filter and aggregation tool modules were removed, the model degenerated into a standard RAG framework. The performance dropped substantially to F1=1.51 and D-F1@20=8.09, corresponding to relative declines of 79.2% and 33.6% compared to the full framework. This result demonstrates the contribution of our tool-based design, especially the critical synergy between the filter and aggregation tools in supporting complex global RAG.

### 7.2 Cross-Dataset Evaluation

We evaluate representative methods on 2WikiMulti-hopQA, HotpotQA, MusiQue and GlobalQA using Qwen2.5-7B-Instruct.

Table 4 shows a striking performance gap. While methods achieve 17-41 F1 on 2WikiMultihopQA, HotpotQA benchmarks, all collapse below 2 F1 on GlobalQA—a 95-99% relative decline. Notably, even HyperGraphRAG, designed for graph-structured reasoning, drops from 18.25 to 0.09 F1.

This reveals that multi-hop reasoning (chaining 2-4 documents) and global aggregation (parallel analysis of 20+ documents) are qualitatively different tasks. Existing retrieval paradigms—whether iterative (IRCoT, FLARE) or graph-based (HyperGraphRAG)—fail because they lack mechanisms for: (1) preserving document-level integrity during chunking, (2) filtering large-scale retrieval noise, and (3) performing corpus-wide statistical operations. This motivates our GlobalRAG framework.

Task	TopK		Count		Sort		MinMax		Avg		
	Metric	D-F1@20	F1	D-F1@20	F1	D-F1@20	F1	D-F1@20	F1	D-F1@20	F1
Qwen2.5-3B-Instruct											
StandardRAG	7.98	<b>2.14</b>		8.88	0.93	7.65	2.58	7.99	0.80	8.09	1.56
FLARE	0.01	0.18		0.00	<b>2.33</b>	0.03	0.10	0.00	0.53	0.01	0.68
IRCOT	<b>10.01</b>	1.67		10.30	1.09	11.03	1.56	<b>10.44</b>	<b>1.60</b>	<b>10.38</b>	1.52
HyperGraphRAG	-	0.06		-	0.39	-	0.04	-	0.00	-	0.09
GlobalRAG (ours)	2.14	1.32	<b>13.95</b>	0.02	<b>12.95</b>	<b>9.18</b>	3.35	1.51	6.62	<b>2.52</b>	
Qwen2.5-7B-Instruct											
StandardRAG	<b>7.98</b>	1.64		8.88	0.00	7.65	2.73	7.99	1.33	8.09	1.43
FLARE	0.33	0.00		0.00	2.33	0.30	0.11	0.53	0.53	0.33	0.62
IRCOT	7.76	1.30		8.68	0.04	7.68	0.96	<b>8.25</b>	<b>1.34</b>	<b>8.07</b>	1.02
HyperGraphRAG	-	0.40		-	0.46	-	0.23	-	0.00	-	0.25
GlobalRAG (ours)	2.65	<b>1.75</b>	<b>9.31</b>	<b>2.79</b>	<b>10.45</b>	<b>9.13</b>	1.50	0.80	4.89	<b>2.93</b>	
Qwen2.5-14B-Instruct											
StandardRAG	7.98	1.78		8.88	0.93	7.65	1.96	7.99	1.33	8.09	1.51
FLARE	0.04	0.00		0.00	<b>2.33</b>	0.05	0.17	0.10	0.27	0.05	0.55
IRCOT	8.86	0.13		8.52	0.01	9.56	0.25	8.38	0.00	8.77	0.09
HyperGraphRAG	-	0.01		-	0.47	-	0.04	-	0.00	-	0.09
GlobalRAG (ours)	<b>11.28</b>	<b>7.92</b>	<b>12.92</b>	0.01	<b>15.01</b>	<b>14.65</b>	<b>10.55</b>	<b>4.80</b>	<b>12.01</b>	<b>6.63</b>	

Table 2: Performance comparison across four aggregation tasks (TopK, Count, Sort, MinMax) and their average (Avg). We report Answer F1 (F1) and document F1 (D-F1@20) for different RAG methods using three Qwen2.5-Instruct model sizes.

Setting	F1	D-F1@20
GlobalRAG	7.27	12.18
w/ Chunk-based retrieval	0.40	-
w/o Filter	6.74	10.46
w/o aggregation module	1.51	8.09

Table 3: Ablation study results on different components.

### 7.3 Effect of Different Retrieval Steps

Multi-step retrieval is a core mechanism of our method, essentially enabling iterative information gathering and integration to construct a more comprehensive global view. The number of retrieval steps directly affects the system’s depth of understanding of complex information structures: too few steps may result in insufficient information collection and incomplete reasoning chains, while an appropriate increase allows our progressive reasoning framework to function more effectively. However, too many steps may run into the context window limitation of the model. Therefore, we systematically analyze the effect of varying retrieval steps on global RAG performance. Under the same experimental conditions, we gradually increased the maximum retrieval steps from 0 to 20 and evaluated GlobalRAG against the IRCOT baseline. All other hyperparameters were kept constant for a fair comparison.

As shown in Figure 4, both methods exhibited increasing F1 and D-F1@20 with more retrieval

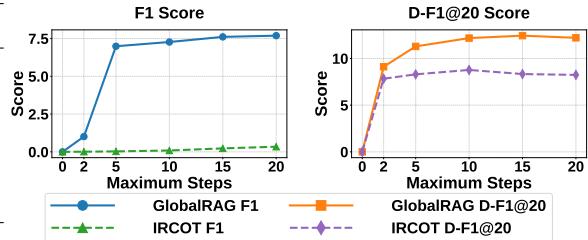


Figure 4: F1/D-F1@20 trends of GlobalRAG and IRCOT (baseline) under different retrieval steps.

steps, but GlobalRAG consistently maintained a clear performance advantage. Specifically, in the 0–5 step range, GlobalRAG’s F1 rose from nearly 0 to 7.0 and its D-F1@20 from 0 to 11.3, whereas IRCOT improved by less than 1 point. In the 5–20 step range, GlobalRAG continued to steadily improve, while IRCOT’s growth flattened, and the performance gap widened. These results indicate that multi-step retrieval indeed brings gains, and our progressive reasoning framework is better able to leverage additional retrieval steps to build a global knowledge graph.

### 7.4 Effect of Different Retrievers

The success of global RAG depends heavily on whether the retriever can accurately fetch complete and relevant external information. We therefore examined the effect of retrievers of different parameter scales. We compared four retrievers: the BGE baseline, Qwen3-Embedding-0.6B, Qwen3-Embedding-4B, and Qwen3-Embedding-8B. Re-

Method	2Wiki	HotpotQA	MuSiQue	GlobalQA
StandardRAG	12.75	16.58	4.53	1.51
HyperGraphRAG	21.10	37.50	20.4	0.09
FLARE	28.30	24.80	2.80	0.62
IRCoT	33.50	41.10	8.90	1.02

Table 4: Cross-dataset performance comparison using F1 score on Qwen2.5-7B-Instruct. Methods achieve reasonable performance on multi-hop QA benchmarks (2WikiMultihopQA, HotpotQA, MuSiQue) but experience catastrophic collapse on corpus-level reasoning tasks (GlobalQA), revealing a fundamental gap between local RAG and global RAG.

sults in Table 5 show that larger retrievers consistently provide higher-quality information and thus yield better global RAG accuracy. Using the BGE retriever as a baseline ( $F1=7.27$ ,  $D-F1@20=12.18$ ), Qwen3-0.6B brought improvements of 2.5% and 0.7%, while Qwen3-8B achieved gains of 18.7% and 6.6%. These results show that our framework maintains performance across different retriever sizes and its scalability: as retrievers improve, our method can further capitalize on the higher-quality documents to achieve larger performance gains. This indicates strong potential for future advances in global RAG with more powerful retrievers.

Retriever	F1	D-F1@20
bge-large-en	7.27	12.18
Qwen3-Embedding-0.6B	7.45	12.27
Qwen3-Embedding-4B	7.83	12.77
Qwen3-Embedding-8B	8.63	12.99

Table 5: Performance of different retrievers on the dataset.

### 7.5 Effect of Different Filters

A key limitation of existing RAG methods in global RAG is their vulnerability to noisy documents. Our GlobalRAG framework addresses this by incorporating an intelligent filter that accurately identifies and removes irrelevant documents. The scale of the filter directly determines its precision and thus the overall reasoning quality. We compared three filters of different parameter sizes: Qwen3-4B, Qwen3-8B, and Qwen3-14B. Each filter was tested with the same chunk-level retriever and aggregation module to validate the effectiveness of our retrieval–reading–filter pipeline.

Table 6 shows that larger filters consistently improved GlobalRAG performance: Qwen3-4B achieved  $F1=7.27$  and  $D-F1@20=12.18$ ; Qwen3-8B yielded gains of 2.2% and 0.9%; and Qwen3-

14B further increased scores to 7.79 and 12.46, representing relative improvements of 7.2% and 2.3%. Larger filters also reduced performance variance, improving stability. These findings confirm that LLM-driven filters effectively address noise interference, ensuring that only high-quality documents are passed to the aggregation execution module, thus enabling hybrid reasoning that combines language understanding with symbolic computation.

Filter	F1	D-F1@20
Qwen3-4B	7.27	12.18
Qwen3-8B	7.43	12.29
Qwen3-14B	7.79	12.46

Table 6: Performance of different filter sizes.

### 7.6 Effect of Different Retrieval Numbers

In global RAG tasks, increasing the number of retrieved documents can introduce noise, while decreasing it risks missing key documents. We therefore evaluated the effect of varying retrieval numbers from 5 to 30, comparing GlobalRAG with IRCOT under the same retriever setup. As shown in Figure 5, GlobalRAG’s  $F1$  and  $D-F1@k$  steadily increased with more retrieved documents, reaching peak performance at 30 documents. This demonstrates that GlobalRAG’s design effectively selects relevant documents from larger candidate sets without being distracted by noise. Its filtering module enables it to leverage expanded recall to boost coverage while discarding irrelevant content.

In contrast, IRCOT peaked at 10 retrieved documents but then declined, with performance at 30 documents falling below that at 5. This highlights its lack of noise filtering: as more documents are retrieved, semantically related but factually irrelevant noise is passed into the model, crowding out relevant document and misleading reasoning, thus degrading performance.

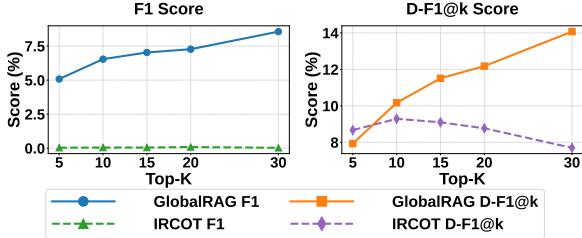


Figure 5: F1/D-F1@k trends of GlobalRAG and IRCOT (baseline) under different retrieval numbers (Top-K).

## 8 Conclusion

This paper investigates the challenges of global query processing in retrieval-augmented generation systems. We make the following contributions:

First, we introduce GlobalQA, a novel benchmark specifically designed to evaluate RAG systems on global aggregation tasks. Unlike existing benchmarks that focus on local information retrieval, GlobalQA requires systems to perform corpus-wide traversal and multi-document reasoning, better reflecting real-world knowledge-intensive applications. Our benchmark comprises four aggregation task types that systematically assess reasoning capabilities across different computational complexities.

Second, through extensive empirical evaluation, we identify and quantify three fundamental limitations of current RAG architectures when handling global queries: (1) information fragmentation across distributed documents, (2) retrieval noise amplification in large-scale searches, and (3) computational bottlenecks in exhaustive corpus processing. Our analysis reveals that state-of-the-art RAG systems achieve only 1.5% average F1 score on global tasks, highlighting a critical gap in current methodologies.

Third, we propose GlobalRAG, a hybrid architecture that demonstrates the necessity of multi-paradigm integration for complex reasoning tasks. By combining neural retrieval with programmatic execution, our approach achieves a 5% performance improvement over existing methods, validating that neither pure neural nor pure symbolic approaches alone are sufficient for global query processing.

Our work identifies global query processing as a challenge in RAG research, providing both evaluation tools and architectural insights for advancing the field.

## References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-rag: Learning to retrieve, generate, and critique through self-reflection](#). *Preprint*, arXiv:2310.11511.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. [Ms marco: A human generated machine reading comprehension dataset](#). *Preprint*, arXiv:1611.09268.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. [Improving language models by retrieving from trillions of tokens](#). *Preprint*, arXiv:2112.04426.
- Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. [Research: Learning to reason with search for llms via reinforcement learning](#). *Preprint*, arXiv:2503.19470.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global: A graph rag approach to query-focused summarization](#). *Preprint*, arXiv:2404.16130.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Fried. 2023. [Enabling large language models to generate text with citations](#). *arXiv preprint arXiv:2305.14627*.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. [From rag to memory: Non-parametric continual learning for large language models](#). *Preprint*, arXiv:2502.14802.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps](#). *Preprint*, arXiv:2011.01060.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Preprint*, arXiv:2112.09118.
- 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](#). *Preprint*, arXiv:2305.06983.
- 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. *Preprint*, arXiv:2503.09516.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.
- Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, Qinyuan Cheng, Bo Wang, Yining Zheng, Yuxin Wang, Zhangyue Yin, and Xipeng Qiu. 2025. R3-rag: Learning step-by-step reasoning and retrieval for llms via reinforcement learning. *Preprint*, arXiv:2505.23794.
- Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. *Preprint*, arXiv:2503.21322.
- Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3281, Bangkok, Thailand. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.
- Yuan Pu, Zhuolun He, Tairu Qiu, Haoyuan Wu, and Bei Yu. 2024. Customized retrieval augmented generation and benchmarking for eda tool documentation qa. *Preprint*, arXiv:2407.15353.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.
- Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.
- Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *Preprint*, arXiv:2405.12035.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *Preprint*, arXiv:2401.18059.
- Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. NumeroLogic: Number encoding for enhanced LLMs’ numerical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 206–212, Miami, Florida, USA. Association for Computational Linguistics.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.
- Alex Wang and 1 others. 2023a. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.
- Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. *Preprint*, arXiv:2412.13018.
- Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2023b. Knowledge graph prompting for multi-document question answering. *Preprint*, arXiv:2308.11730.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. [Siren’s song in the ai ocean: A survey on hallucination in large language models](#). *Preprint*, arXiv:2309.01219.