HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation

Pei Liu[1,2], Xin Liu[2], Ruoyu Yao[2], Junming Liu[1], Siyuan Meng[1], Ding Wang[1*], Jun Ma[23*]
[1]Shanghai Artificial Intelligence Laboratory [2]The Hong Kong University of Science and Technology (Guangzhou)
[3]The Hong Kong University of Science and Technology
pliu061@connect.hkust-gz.edu.cn wangding@pjlab.org.cn jun.ma@ust.hk

ABSTRACT

While Retrieval-Augmented Generation (RAG) augments Large Language Models (LLMs) with external knowledge, conventional single-agent RAG remains fundamentally limited in resolving complex queries demanding coordinated reasoning across heterogeneous data ecosystems. We present HM-RAG, a novel Hierarchical Multi-agent Multimodal RAG framework that pioneers collaborative intelligence for dynamic knowledge synthesis across structured, unstructured, and graph-based data. The framework is composed of three-tiered architecture with specialized agents: a Decomposition Agent that dissects complex queries into contextually coherent sub-tasks via semantic-aware query rewriting and schema-guided context augmentation; Multi-source Retrieval Agents that carry out parallel, modality-specific retrieval using plug-and-play modules designed for vector, graph, and web-based databases; and a Decision Agent that uses consistency voting to integrate multi-source answers and resolve discrepancies in retrieval results through Expert Model Refinement. This architecture attains comprehensive query understanding by combining textual, graph-relational, and web-derived evidence, resulting in a remarkable 12.95% improvement in answer accuracy and a 3.56% boost in question classification accuracy over baseline RAG systems on the ScienceQA and CrisisMMD benchmarks. Notably, HM-RAG establishes state-of-the-art results in zero-shot settings on both datasets. Its modular architecture ensures seamless integration of new data modalities while maintaining strict data governance, marking a significant advancement in addressing the critical challenges of multimodal reasoning and knowledge synthesis in RAG systems. Code is available at https://github.com/ocean-luna/HMRAG.

KEYWORDS

Retrieval-Augmented Generation (RAG), Multimodal Representation, Multi-agent Systems, Multi-source RAG

1 INTRODUCTION

In an era defined by the rapid proliferation of data, the ability to efficiently retrieve relevant information from heterogeneous sources has emerged as a fundamental pillar of modern information systems [14]. Multimodal retrieval systems, which integrate text, images, vectorized data, and web-based content, are becoming indispensable across domains such as e-commerce, healthcare, and scientific research [59]. These systems enable the seamless navigation of diverse data types, empowering users to derive actionable insights across multiple modalities. However, despite remarkable progress in recent years, multimodal retrieval continues to present significant challenges. The complexity arises from the need to reconcile the diversity of query types, the heterogeneity of data formats, and the varying objectives of retrieval tasks, all of which demand sophisticated solutions to bridge the gap between data representation and user intent.

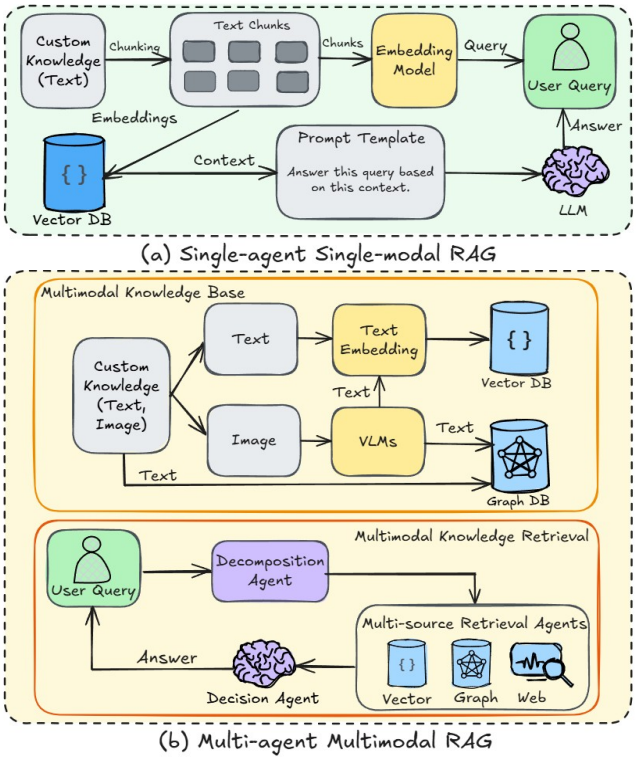


Figure 1: Comparison of (*a*) single-agent single-modal RAG and (*b*) multi-agent multimodal RAG. The multi-agent multimodal RAG processes multimodal data by converting them into vector and graph databases. It leverages multi-source retrieval across vector, graph, and web-based databases, enabling more comprehensive and efficient information retrieval. This advanced approach allows the multi-agent multimodal RAG to achieve superior performance in handling complex queries and diverse data types, setting it apart from the more limited single-agent single-modal RAG.

The evolution of retrieval technologies has historically centered on single-modal architectures, where queries and retrieval mechanisms operate within a single predefined modality [3, 33]. While text-based retrieval-augmented generation (RAG) systems have demonstrated robust performance in processing linguistic information [43], their inability to handle visual content has spurred

the development of image-based RAG approaches [7, 25, 41]. However, current multimodal implementations face a critical bottleneck: Although image-based RAG systems excel at visual content processing, they often fail to establish coherent cross-modal correlations between visual elements and textual context. This limitation is particularly acute in multimodal question answering, where systems must integrate visual perception with textual semantics to generate contextually relevant responses.

Recently, graph-based retrieval frameworks have been proposed to enhance the modeling of textual interdependencies based on the construction of knowledge graphs, represented by GraphRAG [12] and LightRAG [18]. These approaches are further extended to processing multimodal inputs [37], where graph structures are leveraged for the accurate capture of cross-modal relationships. Despite these advances, graph-based methods face an inherent trade-off: while they effectively capture high-level modality interactions, they often sacrifice fine-grained information fidelity. This becomes problematic in scenarios requiring precise textual segment retrieval, as the abstraction process inherent to graph modeling obscures granular textual details critical for nuanced analysis.

Meanwhile, another critical challenge has been noticed in reconciling the complementary strengths of different modalities [13, 15, 31]. Textual modalities excel at encoding granular semantic details and conceptual relationships, while visual modalities, by contrast, are capable of capturing spatial context and facilitating spatial relationship understanding. Current modality-specific systems [33, 54] exhibit critical limitations in cross-modal synthesis, producing retrieval outcomes that are either overspecialized in textual precision or confined to visual pattern recognition. This modality isolation creates systemic vulnerabilities in heterogeneous data environments, where the absence of cross-modal alignment protocols risks critical information loss during retrieval operations. For instance, visual queries in text-centric systems fail to map conceptual questions to illustrative elements, while text-intensive inquiries in vision-oriented frameworks lack mechanisms for lexical disambiguation. These architectural gaps highlight the urgent need for frameworks that can harmonize granular semantic detail with cross-modal contextual coherence.

To address these challenges, we introduce Hierarchical Multi-Agent Retrieval-Augmented Generation (HM-RAG), a novel framework that enhances multimodal retrieval through coordinated multi-agent collaboration. HM-RAG employs a three-tiered architecture with specialized agents operating in the RAG pipelines. The Decomposition Agent analyzes query intent and dynamically rewrites requests to ensure cross-modal compatibility. The Multi-Source Retrieval Agent conducts parallel knowledge acquisition via lightweight multimodal retrievals across diverse data sources, including vectors, graphs, and web-based databases. Finally, the Decision Agent synthesizes and refines candidate responses using domain-specific verification strategies to ensure accuracy and coherence. This hierarchical design systematically orchestrates text-image evidence integration through structured agent interactions, enabling layered reasoning. Unlike conventional approaches, HM-RAG combines query decomposition, parallelized information retrieval, and expert-guided answer refinement to achieve efficient and contextually relevant responses. Our contributions are summarized as follows:

- We propose a novel **Modularized Hierarchical Framework** that modularizes query processing into specialized agent-based components, and this facilitates scalable and efficient multimodal retrieval.
- We enable **Multi-source Plug-and-play Retrieval Integration**, which offers seamless connectivity across diverse data sources. By efficiently routing queries to vector, graph, and web-based retrieval agents, our approach ensures flexibility and efficiency in handling heterogeneous data environments, streamlining complex information retrieval processes.
- We employ **Expert-guided Refinement** processes to enhance response quality to ensure both operational efficiency and contextual precision through minimal expert oversight.
- We demonstrate the effectiveness of HM-RAG through extensive experiments on benchmark datasets, and the results attain **State-of-the-art Performance** on the ScienceQA and CrisisMMD benchmarks.

2 RELATED WORK

2.1 Retrieval-Augmented Generation

RAG systems have evolved significantly to enhance their multimodal reasoning capabilities [16, 20, 33, 47]. Initially, text-based RAG systems integrated Large Language Models (LLMs) with external textual knowledge, improving performance in question answering by retrieving relevant text fragments [4, 27, 57]. However, as visually rich documents became more prevalent, the limitations of text-only systems became evident, prompting the development of image-based RAG approaches [5, 6, 38, 46]. While these methods aimed to retrieve visual content for Large Vision-Language Models (VLMs), they faced challenges in effectively integrating text and image modalities, as the retrieval processes were largely independent, hindering a deep understanding of their interrelationships.

To address these challenges, graph-based RAG systems emerged, leveraging structured knowledge representations to capture both inter-modal and intra-modal semantic relationships [9, 18, 28, 44]. These systems utilize vector-space embeddings and topological relationships to model complex document structures, enabling the retrieval of semantically coherent contexts that go beyond simple text fragments [12, 42, 53]. Graph-based RAG systems are particularly effective in understanding relationships between text and images, as well as extracting relationships within the text itself [37]. However, current RAG implementations often rely on single-source retrieval, limiting their ability to handle complex queries that require simultaneous processing of vector, graph, and web-based databases [19]. This limitation is particularly significant in applications requiring private data retrieval and real-time updates, where the absence of integrated multi-source retrieval capabilities can lead to incomplete or outdated information. To fully leverage the strengths of each data modality and meet the demands of dynamic and heterogeneous data environments, RAG systems must evolve to support coordinated multi-source retrieval and synthesis.

2.2 Agents in RAG

RAG has become a key paradigm for knowledge-intensive tasks by integrating retrieval mechanisms with generative models, significantly enhancing language model capabilities. However, traditional RAG implementations often rely on static pipelines that struggle with multimodal query processing [8, 48]. Recent agent-based RAG architectures have addressed these limitations by improving system modularity and operational flexibility [11, 21, 29]. The agent-oriented approach breaks down query processing into specialized components like semantic parsing, cross-modal retrieval, and context-aware generation, allowing targeted optimization while maintaining overall adaptability. PaperQA [32] exemplifies this by leveraging academic literature to generate evidence-based responses, reducing hallucinations in scientific applications.

Building on this, Active RAG methodologies like FLARE [30] introduce temporal dynamism through anticipatory retrieval, enhancing performance in extended text generation. Despite these advances, challenges in multimodal integration persist. Emerging Dynamic RAG approaches [49, 50] propose entity-aware augmentation strategies to dynamically incorporate retrieved entity representations, addressing context window limitations while preserving semantic coherence. Our HM-RAG framework synthesizes these innovations through a hierarchical multi-agent architecture leveraging LLMs' semantic comprehension. This design enables dynamic query adaptation and multimodal retrieval, providing an optimized solution for complex information retrieval and generation tasks across diverse data modalities. By integrating these advancements, HM-RAG addresses key challenges in multimodal reasoning and knowledge synthesis, paving the way for more robust and adaptable RAG systems.

3 METHODOLOGY

We introduce HM-RAG, a novel framework tackling complex challenges in RAG systems. As depicted in Figure 2, HM-RAG features an innovative multi-agent, multimodal architecture with specialized agents for information extraction and multi-source retrieval. Given a natural language question q and a reference document $\mathcal{D}$, RAG retrieves semantically relevant content from $\mathcal{D}$, integrating it with generative language models to produce answers strictly grounded in $\mathcal{D}$. This approach advances multimodal question answering and multi-agent RAG capabilities. The subsequent sections provide a detailed exposition of HM-RAG's architectural design. Through this systematic description, we elucidate the framework's core mechanisms for effectively integrating and utilizing multimodal information and multi-source retrieval, ultimately leading to enhanced accuracy in RAG applications.

3.1 Multimodal Knowledge Pre-Processing

This section focuses on multimodal data processing, aiming to convert textual data and visual images into vector and graph database representations for enhanced retrieval operations. Our methodology employs VLMs to transcode visual information into textual representations, which are subsequently integrated with original text corpora to jointly construct vector and graph databases.

3.1.1 Multimodal Textual Knowledge Generation. Conventional entity-centric approaches for multimodal knowledge extraction rely on predefined categorical boundaries, limiting their capacity to recognize novel visual concepts. We utilize the BLIP-2's framework [34] to harness the open vocabulary potential of pretrained VLMs. Building upon the generalized vision to language conversion paradigm:

$$T_v = \mathcal{D}_{blip2}(f_{align}(\mathcal{E}_{blip2}(I_v))) \tag{1}$$

where visual encoder $\mathcal{E}_{blip2}$ extracts features from input image I_v and cross-modal alignment module f_{align} bridges vision-language semantics. Our framework addresses the critical limitation of oversimplified machine-generated descriptions, particularly addressing BLIP-2's over-condensed outputs that lack visual specificity, through contextual refinement mechanisms leveraging original textual data.

This process is divided into three synergistic phases. **Hierarchical visual encoding** via established architectures [10, 22, 39] to generate patch embeddings $V_i \in \mathcal{R}^{d_v \times N_p}$. **Cross-modal interaction** where learnable queries $Q_i \in \mathcal{R}^{d_q \times L_q}$ attend to visual features through scaled dot product attention, dynamically weighting spatial semantic correlations. **Context-aware text generation** that fuses latent text features from prior descriptions $T_v^{i,t}$ with cross-modal representations for autoregressive decoding. Contextual refinement during this phase enhances semantic alignment, achieving measurable reductions in descriptive ambiguity and lexical sparsity for the final output T_v.

The resultant multimodal textual knowledge base is subsequently formed through the systematic integration of original textual inputs with generated textualizations.

$$T_m = Concate(T, T_v) \tag{2}$$

where T corresponds to the source textual corpus and T_m represents the multimodal textual aggregation formed through heterogeneous fusion processes.

3.1.2 Multimodal Knowledge Graphs Construction. We establish multimodal knowledge graphs (MMKGs) by synergizing VLM-enhanced descriptions with LLM-based structural reasoning. Building upon the refined visual descriptions T_v generated by VLMs, optionally fused with external textual knowledge T, we employ the LightRAG framework [18] for efficient multi-hop reasoning and dynamic knowledge integration:

$$G = LightRAG(T_v, T) \tag{3}$$

LightRAG processes multimodal inputs through a hybrid extraction strategy. **Entity-Relation Extraction**: a specialized function f decomposes inputs into entities $E = \{e_1, \cdots, e_n\}$ and relation triplets $R = \{(h_i, r_i, t_i)\}$, where $h, t \in E$ represent head/tail entities and $r \in R$ denotes relations. **Dual-level Reasoning Augmentation:** Dual-scale retrieval mechanisms $Retrieve_{global+local}$ dynamically fetch relevant triplets during inference; global retrieval identifies thematic clusters while local extraction focuses on entity-specific connections.

The constructed MMKG $G = (E, R)$ formalizes knowledge as triplets (h, r, t), where entities encompass both visual concepts

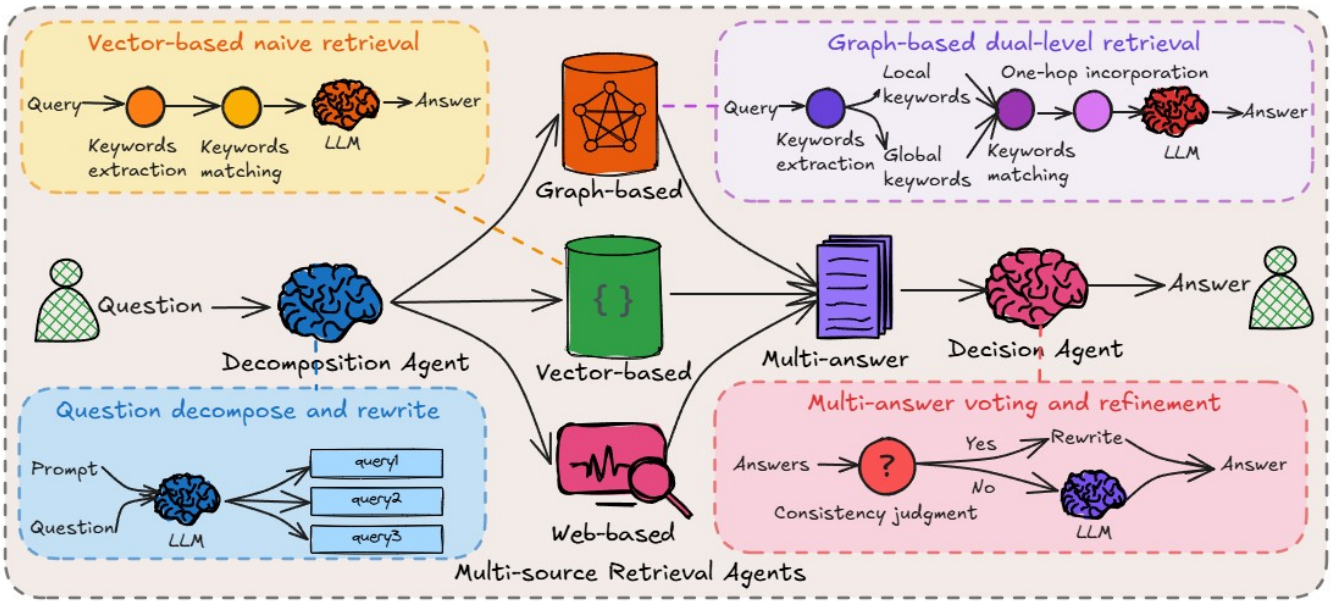


Figure 2: Overview of HM-RAG. A multi-agent multi-modal framework operates in three stages: First, the Decomposition Agent uses an LLM to rewrite and decompose the question into several sub-queries. Second, the Multi-source Retrieval Agent retrieves the top-k relevant documents from vector-, graph- and web-based sources as needed. Finally, the Decision Agent provides a voting mechanism and refinement process to generate the final answer.

from T_v and textual knowledge from T. Crucially, visual data storage locations are embedded during graph construction, enabling cross-modal grounding. This architecture establishes a bidirectional knowledge enhancement framework: language models achieve visual-contextualized reasoning through visual-semantic relationships embedded in G, and vision-language models dynamically update knowledge embeddings via continuous multimodal integration, effectively mitigating hallucination probabilities through representation consistency constraints.

3.2 Decomposition Agent for Multi-intent Queries

The Decomposition Agent is a pivotal component of the proposed framework, designed to break down complex, multi-intent user queries into coherent and executable sub-tasks. This agent addresses a critical limitation of traditional systems, which often struggle to process compound queries requiring joint reasoning across multiple data sources. By leveraging a hierarchical parsing mechanism, the Decomposition Agent identifies the underlying structure of user queries and decomposes them into atomic units, with each targeting a specific data modality or retrieval task.

The proposed framework operates in two stages, both driven by task-specific LLM-prompting strategies. **Decomposition Necessity Judgment.** The agent first determines whether the input question Q contains multiple intents using a binary decision prompt that instructs the LLM to classify it as single-intent or multi-intent. If the output is multi-intent, Q proceeds to decomposition. Otherwise, return question Q directly. **Intent Decomposition.** The LLM decomposes Q into candidate sub-questions $q = \{q_1, \cdots, q_n\}$ using a structured prompt: "Decompose the reasoning steps of the original question into 2 to 3 simply and logically connected sub-questions based on its intent while retaining keywords from the original question." inspired by [35].

3.3 Multi-source Plug-and-Play Retrieval Agents

We propose a modular multi-agent retrieval framework that dynamically composes heterogeneous multimodal search strategies through standardized interfaces. By decoupling retrieval functionalities into three specialized agents—vector-based retrieval agent, graph-based retrieval agent, and web-based retrieval agent—the system achieves domain-agnostic adaptability while ensuring interoperability across diverse search scenarios. Each agent adheres to unified communication protocols, enabling seamless integration of vector semantic search, graph topological exploration, and real-time web retrieval capabilities. This design allows each retrieval agent to function as a plug-and-play component, ensuring that they can be easily integrated or replaced without affecting the overall system performance. This modularity not only enhances flexibility but also maintains task-specific optimization objectives, making the framework highly adaptable to various applications and data modalities.

3.3.1 Vector-based Retrieval Agent for Fine-Grained Information. This agent leverages a naive retrieval architecture [18] to search unstructured textual corpora efficiently. Given the user query q, the system first computes its semantic embedding h_q using an encoder

$\mathcal{E}_{text}$:

$$h_q = \mathcal{E}_{text}(q) \tag{4}$$

where $h_q \in \mathbb{R}^d$ represents the query's embedding in a d-dimensional vector space.

Next, the system computes the semantic similarity between the query embedding h_q and all document embeddings h_j using cosine similarity:

$$s_j = \frac{h_q^T h_j}{||h_q||||h_j||}, \quad \forall j \in [1, M] \tag{5}$$

where $j \in [1, M]$, with M being the total number of documents. The similarity score s_j quantifies how closely each document aligns with the query, forming the basis for ranking retrieved documents.

Based on the similarity scores, the system retrieves the top-k most relevant documents:

$$\mathcal{R}_k = \{c_1, \cdots, c_k\} \quad s.t. \quad s_1 \geq s_2 \geq \cdots \geq s_k \tag{6}$$

where $\mathcal{R}_k$ denotes the set of top-k retrieved contexts, ensuring that only the most relevant information is used for subsequent processing.

Subsequently, the language model generates answers $\mathcal{A}_v$ conditioned on retrieved contexts through constrained decoding:

$$\mathcal{A}_v = \mathcal{P}(q, \mathcal{R}_k) = Concate(q, Context, \{c_1, \cdots, c_k\}) \tag{7}$$

where $\mathcal{P}$ represents the generation process, which concatenates the query q, retrieved contexts $\{c_1, \cdots, c_k\}$, and additional contextual information to produce the final answer.

Specifically, the conditional probability of generating a token sequence y given the query q and retrieved contexts $\mathcal{R}_k$ is modeled as:

$$p(y|q, \mathcal{R}_K) = \prod_{t=1}^{T} p_{lm}(y_t|y_{<t}, q, \mathcal{R}_K) \tag{8}$$

where p_{lm} denotes the conditional probability of a token in the auto-regressive generation process of a language model, ensuring that the generated answer is contextually coherent.

Furthermore, the attention mechanism explicitly incorporates retrieved content into the generation process:

$$Attention(Q, K, V) = softmax(\frac{Q[h_q; H_{\mathcal{R}}]^T}{\sqrt{d_k}})[h_q; H_{\mathcal{R}}] \tag{9}$$

where $H_{\mathcal{R}} \in \mathbb{R}^{K \times d}$ stacks the embeddings of retrieved chunks, and $[h_q; H_{\mathcal{R}}]$ concatenates the query embedding with the retrieved chunk embeddings, enhancing the model's ability to focus on relevant information. To ensure the reliability of the generated answers, constraints enforce top-$p = 1.0$ and a temperature of 0, ensuring deterministic decoding based on the highest probability tokens. This minimizes the risk of hallucination and ensures factual accuracy.

3.3.2 Graph-based Retrieval Agent for Relational Information. This agent leverages LightRAG's graph traversal capabilities to resolve multi-hop semantic queries over MMKGs [18]. Given an input query q, the agent constructs a context-aware subgraph $G_q \subseteq G$ by dynamically retrieving entities and relations through the joint attention mechanism of LightRAG. The subgraph is defined as:

$$G_q = \{(h, r, t)|LightRAG_{graph}(q, h, r, t) > \tau\} \tag{10}$$

where $LightRAG_{graph}$ computes relevance scores by aligning query embeddings with graph triplet representations through cross-modal attention, ensuring that only highly relevant triplets are included in the subgraph.

To efficiently address complex queries, the agent employs a hierarchical search strategy that balances efficiency and comprehensiveness. First, the agent prioritizes local 1-hop neighbors of query-relevant entities using relation-specific attention weights. This ensures that directly connected entities and relations are retrieved first, providing a foundation for further exploration. Next, the agent expands the search globally by identifying cross-modal paths through iterative message passing. This allows the agent to explore deeper semantic relationships beyond immediate neighbors, enhancing the richness of the retrieved information.

Furthermore, the framework is a dual-level retrieval framework that integrates graph-structured knowledge with vector representations through a three-phase retrieval process. First, the framework performs semantic decomposition of the input query q to derive local keywords q_l and global keywords q_g. This step captures both fine-grained and high-level semantic information. Second, the framework executes hybrid graph-vector matching. An optimized vector database aligns q_l with entity attributes while mapping q_g to relational patterns in the knowledge graph $G = (\mathcal{V}, \mathcal{E})$. This hybrid approach ensures that both explicit entity attributes and latent relational semantics are considered.

Finally, to enhance retrieval completeness, the framework performs higher-order context expansion. The retrieved subgraph is expanded to include one-hop neighbors of both retrieved nodes and edges:

$$\mathcal{A}_g = \{v_i \in \mathcal{V} \wedge (v_i \in \mathcal{N}_v \vee v_i \in \mathcal{N}_e)\} \tag{11}$$

where $\mathcal{N}_v$ and $\mathcal{N}_e$ denote the one-hop neighbors of retrieved nodes and edges, respectively. This step ensures that the retrieved subgraph retains structural integrity while capturing broader contextual relationships. The final answer $\mathcal{A}_g$ is generated using $\mathcal{A}_g = LLM(\mathcal{A}_g)$ with a lightweight LLM.

3.3.3 Web-based Retrieval Agent for Real-Time Information. The web retrieval component serves as a critical bridge between information retrieval and natural language generation, significantly enhancing the semantic fidelity and factual grounding of generated text. Our work utilizes the Google Serper API. The system acquires knowledge through parameterized API requests to Google's search engine. For an input query q, the retrieval process is formalized as:

$$\mathcal{R} = Google(q; \theta_{search}) \tag{12}$$

where θ_{search} specifies search configuration parameters. We adopt the setting that $\theta_{search} = \{num_{results} = k, language = en, type = web\}$. The API returns structured results $\mathcal{A}_w = \{a_i\}_{i=1}^k$, each containing a title, a snippet, a URL, and positional ranking metadata.

The Google Serper framework demonstrates particular efficacy in real-world deployment scenarios through three principal operational modalities, each addressing the critical requirements of modern knowledge-aware systems. First, the real-time fact verification module computes factual validity scores through neural memory interrogation. Second, the attribution-aware generation

protocol ensures traceability through dual-phase attention routing. Third, the adaptive query expansion mechanism addresses vocabulary mismatch through differential term weighting.

3.4 Decision Agent for Multi-answer Refinement

Consistency Voting. The framework evaluates the semantic agreement among answers $\{\mathcal{A}_v, \mathcal{A}_g, \mathcal{A}_w\}$ generated by vector-based, graph-based, and web-based retrieval systems using ROUGE-L and BLEU metrics. Summaries $\{\mathcal{S}_v, \mathcal{S}_g, \mathcal{S}_w\}$ are first generated for each answer. ROUGE-L measures the overlap of key information using the Longest Common Subsequence (LCS), defined as:

$$R_L = \frac{LCS(\mathcal{S}_i, \mathcal{S}_j)}{max(|\mathcal{S}_i|, |\mathcal{S}_j|)} \tag{13}$$

where the numerator represents the length of the LCS between summaries, while the denominator normalizes the score. This metric emphasizes consistency in retaining critical factual information.

BLEU evaluates the localized precision of n-gram matches between summaries, defined as:

$$BLEU = exp(\sum_{n=1}^{k} w_n log p_n) \cdot min(1, \frac{|\mathcal{S}_j|}{|\mathcal{S}_i|}) \tag{14}$$

where p_n represents n-gram precision, and w_n denotes weight coefficients. This metric excels in detecting precise matches of terminologies or numerical values.

A weighted fusion of R_L and $BLEU$ is then applied to balance macro-level semantic alignment with micro-level detail consistency, measuring the similarity between any two answers. If the pairwise similarity exceeds a predefined threshold, the result is refined using a Lightweight Language Model (LLM) to produce the final answer A. The framework proceeds to expert model refinement if the similarity is below the threshold.

Expert Model Refinement. For conflicting answers, the framework employs LLMs, Multimodal LLMs (MLLMs) or Cot-based language models (Cot-LMs) to synthesize a refined response by integrating multi-source evidence. The LLM or MLLM processes the original query q and the retrieved evidence to generate the final answer $\mathcal{A}$. This step serves as an expert-guidance,ensuring that the final response is both contextually coherent and factually accurate, even when initial answers exhibit discrepancies.

4 EXPERIMENTS

4.1 Experimental Setup

Dataset. We conduct experiments across two multimodal reasoning benchmarks spanning divergent modality configurations, including complex question answering (ScienceQA) and crisis event classification (CrisisMMD).

ScienceQA [40]. This dataset is the first large-scale multimodal benchmark for scientific question answering spanning 3 core disciplines (Natural Science, Social Science, and Formal Science). The dataset contains 21,208 carefully curated examples organized hierarchically across 26 topics, 127 categories, and 379 distinct reasoning skills. Each instance combines textual questions with optional visual contexts (diagrams, charts, or photographs), with a balanced split of 12,726 training, 4,214 validation, and 4,268 test samples. Following the evaluation protocol established in LLaVA [36], we report averaged accuracy across all test samples to assess model performance in multimodal understanding and multi-step scientific reasoning. Notably, 34.6% of test questions require simultaneous processing of both visual and textual information to derive correct answers.

CrisisMMD [2]. This dataset presents a challenging multimodal collection for disaster response applications, comprising approximately 35,000 social media posts containing both visual and textual content from real-world crisis events. It features a comprehensive annotation scheme with seven distinct disaster categories and four granular severity levels. Its unique value lies in capturing authentic user-generated content that preserves natural noise patterns and complex cross-modal relationships inherent in crisis communication. These characteristics make it particularly suitable for evaluating zero-shot adaptation models, as successful performance on this benchmark directly correlates with practical deployment capabilities in dynamic emergency scenarios where clean data and explicit modality alignments are typically unavailable.

Implementation Details. We utilize DeepSeek-R1-70B for dynamic graph construction and optimize LightRAG's hybrid retrieval mechanism through Qwen2.5-7B's parameter adaptation framework, which is consistent with VaLik [37]. During decision refinement, we employ GPT-4o for ScienceQA dataset processing and GPT-4 for CrisisMMD dataset analysis. All multimodal reasoning workflows operate on a single NVIDIA A800-80GB GPU, seamlessly supporting the concurrent execution of graph neural network computations and retrieval-augmented generation tasks through memory-optimized parallelization.

4.2 Main Results

In this section, we conduct a systematic evaluation of HM-RAG against state-of-the-art zero-shot LLMs, VLMs, and RAG-enhanced approaches across multiple benchmarks. The results are presented in Table 1 and Table 2, which demonstrate **the consistent superiority of HM-RAG over all comparative methods**.

4.2.1 Results on ScienceQA. Table 1 systematically quantifies the multimodal question-answering performance of HM-RAG and existing zero-shot approaches on the ScienceQA dataset. As shown in the table, HM-RAG establishes the state-of-the-art average accuracy of 93.73%, surpassing the previous best zero-shot VLM method LLaMA-SciTune and GPT-4o by 4.11% and 2.82%, respectively, and significantly outperforming the single-agent RAG variants. Compared to vector-based, graph-based, and web-based baselines, HM-RAG achieves 12.95%, 12.71%, and 12.13% absolute improvements, respectively. Notable gains are observed in the accuracy of Social Science (SOC) tasks, where the improvements over web-based and graph-based baselines reach 24.38% and 20.65%, respectively. The framework also exceeds human expert performance by 6.03%.

4.2.2 Results on CrisisMMD. Table 2 presents a comprehensive evaluation of multimodal understanding capabilities on the CrisisMMD benchmark. Our analysis reveals three key observations. First, multimodal enhanced LLMs consistently outperform both text-only LLMs and specialized VLMs across all tasks. The proposed method achieves state-of-the-art performance with an average accuracy of

Table 1: Top-1 retrieval performance comparison (Accuracy %) on the ScienceQA Dataset. #P denotes the number of trainable parameters. Categories include: NAT (Natural Science), SOC (Social Science), LAN (Language Science), TXT (Text Context), IMG (Image Context), NO (No Context), G1-6 (Grades 1-6), and G7-12 (Grades 7-12). The comparisons presented are based on the state-of-the-art zero-shot learning results obtained from the ScienceQA leaderboard[1].

Learning	Models	#P	Subject			Context Modality			Grade		Average
			NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Baseline	Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
Zero-shot LLMs	ChatGPT [56]	-	-	-	-	-	-	-	-	-	69.41
	GPT-3 (0-shot) [40]	173B	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87	74.04
	DDCoT (GPT-3) [58]	175B	78.60	73.90	80.45	77.27	69.96	82.93	80.65	73.50	78.09
	CoT GPT-3 + Doc [24]	173B	-	-	-	-	-	-	-	-	79.91
	DDCoT (ChatGPT) [58]	175B	80.15	76.72	82.82	78.89	72.53	85.02	82.86	75.21	80.15
Zero-shot VLMs	LaVIN-13B [56]	-	-	-	-	-	-	-	-	-	77.54
	LLaMA-SciTune [23]	7B	84.50	94.15	82.91	88.35	83.64	88.74	85.05	85.60	86.11
	LG-VQA (BLIP-2) [17]	-	-	-	-	-	-	-	-	-	86.32
	LG-VQA (CLIP) [17]	-	-	-	-	-	-	-	-	-	87.22
	LLaMA-SciTune [23]	13B	89.30	95.61	87.00	93.08	86.67	91.75	84.37	91.30	90.03
Zero-shot Single-agent RAG	Vector-based [37]	7B	84.54	74.24	86.91	82.74	72.53	90.03	84.51	80.28	82.98
	Graph-based [37]	7B	84.15	75.14	87.64	82.99	73.18	89.69	84.40	80.95	83.16
	Web-based	7B	83.79	72.89	91.82	81.09	70.55	94.01	85.98	79.30	83.59
	GPT-4o [26]	-	92.72	93.48	86.09	92.67	90.88	87.60	92.91	88.00	91.16
Zero-shot Multi-agent RAG	HM-RAG	-	**94.36**	**90.66**	**94.91**	**93.79**	**89.94**	**96.03**	**94.42**	**92.49**	**93.73**

Table 2: Top-1 retrieval performance comparison (Accuracy %) on the CrisisMMD Dataset. The -I indicates instruction-tuned variants. Bold denotes the highest value. Task 1 is a binary classification task, while Task 2 and Task 2 Merged are multi-classification tasks. The comparisons are sourced from [37], which represents the pioneering LLM-based work on the CrisisMMD Dataset.

Method	#P	Task 1	Task 2	Task 2 Merged	Average
Single-modal LLMs					
LLaMA-2 [51]	7B	62.32	18.32	21.45	34.03
	13B	63.80	21.82	33.15	39.59
	70B	63.15	28.87	36.89	42.97
Qwen2.5 [55]	7B	65.04	44.52	45.33	51.63
	32B	67.28	46.94	47.07	53.76
	72B	67.95	50.51	50.29	56.25
GPT-4 [1]	-	66.83	47.25	49.44	54.51
Multimodal VLMs					
Qwen2-VL [52]	2B-I	47.56	7.60	7.42	20.86
	7B-I	62.45	32.68	34.20	43.11
	72B-I	65.80	47.21	48.28	53.76
LLaVA [36]	7B	54.00	28.01	30.61	37.54
	13B	60.58	20.14	23.44	34.72
	34B	56.44	25.15	25.07	35.55
CLIP [45]	-	43.36	17.88	20.79	27.34
GPT-4o [26]	-	68.20	47.58	49.55	55.11
Single-agent RAG					
Vector-based [37]	7B	67.49	45.11	45.94	52.85
Graph-based [37]	7B	68.90	50.02	50.69	56.54
Multi-agent RAG					
HM-RAG	-	**72.06**	**51.50**	**52.09**	**58.55**

58.55%, representing 2.44% and 3.44% absolute improvements over the strongest baseline (GPT-4o) and text-only variant (Qwen2.5-72B), respectively, despite using only 7B parameters.

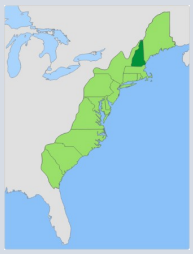
Second, the model scale exhibits a non-linear correlation with performance gains. While Qwen2.5-72B (text-only) achieves 56.25% average accuracy, our 7B multimodal enhanced variant attains an absolute improvement of 2.3%, demonstrating superior parameter efficiency. This trend holds across modalities, with Qwen2-VL-72B-I (VLM) underperforming our method by 4.79% despite equivalent parameter counts.

Third, multimodal integration significantly impacts task performance. Our method shows 5.7% and 2.01% improvements in average accuracy over its text-only and graph-only variants, respectively, which highlights the effectiveness of multi-source reasoning. Notably, the accuracy of 72.06% on Task 1 establishes a new benchmark, outperforming GPT-4o by 3.86% and demonstrating robust visual-textual alignment capabilities.

4.3 Qualitative Analysis

We provide a prediction example as shown in Figure 3 to demonstrate the effectiveness of our proposed model. This example was carefully chosen to showcase the model's ability to handle complex patterns and make accurate choices. For readers interested in additional cases, a more detailed set of examples is provided in Appendix A. In the given example, the multi-source retrieval agents all produce incorrect results since there is no relevant information recorded for this question in the database. To cope with the situation, the expert refinement in the decision agent is used to perform high-level thinking to derive the correct result. This manifests the

[1] https://scienceqa.github.io/leaderboard.html



Question: What is the name of the colony shown?

Choices: (A) Maryland
(B) New Hampshire √
(C) Rhode Island
(D) Vermont

Text-based

The document chunks … Since we do not have direct information from the provided chunks regarding…, and because none of the options are explicitly answered as correct, we must conclude that based on …, the answer is \"FAILED\".

Answer: The answer is (D).

Web-based

Using the map below, please match the colony name to the ... - Brainly\nTheir locations were: A New Hampshire. B Massachusetts colony. C Rhode Island colony. D Connecticut colony. E New York colony … https://brainly.com/question/13673860.
Answer: The answer is (C).

Graph-based

The answer is not provided in the options, but based on the context given by the Knowledge Base, it likely refers to one of the colonies listed. Since no additional information or context was provided, I will state that without sufficient details to choose among the given options.
Answer: The answer is (A).

HMAgent

The map highlights New Hampshire in a darker shade, indicating it as the selected colony.

Answer: The answer is (B).

Figure 3: Case Study: Comparison Between HM-RAG and the Baseline Methods (Vector-based, Graph-based, and Web-based Retrieval Agent).

Table 3: Performance comparison across different variants of HM-RAG on the ScienceQA Dataset. Components include: VA (Vector-based Retrieval Agent), GA (Graph-based Retrieval Agent), WA (Web-based Retrieval Agent), and DA (Decision Agent).

Agent Configuration				NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Average
VA	GA	WA	DA									
×	✓	✓	✓	90.72	88.08	94.09	89.30	84.58	95.68	92.47	88.46	91.04
✓	×	✓	✓	91.21	87.96	94.73	90.32	85.62	95.61	92.22	90.05	91.44
✓	✓	×	✓	88.99	84.81	90.27	88.17	83.09	91.78	89.46	86.62	88.45
✓	✓	✓	×	83.79	72.89	91.82	81.09	70.55	94.01	85.98	79.30	83.59
✓	✓	✓	✓	**94.36**	**90.66**	**94.91**	**93.79**	**89.94**	**96.03**	**94.42**	**92.49**	**93.73**

proficiency of our model in informed decision-making, which assures enhanced robustness compared to relying on a single type of retrieval mechanism.

4.4 Ablation Studies

Table 3 presents a systematic evaluation of individual agent components' contributions through controlled ablation studies on ScienceQA. Three key insights emerge regarding the framework's design. First, the decision agent (DA) establishes itself as the most critical element, with its removal triggering the most substantial performance decline at 10.82%. This component proves particularly vital for synthesizing multi-source decisions, as evidenced by significant accuracy reductions of 21.56% in image-based tasks and 19.60% in social reasoning tasks when DA is disabled. Second, the web-based retrieval agent (WA) demonstrates robust integration capabilities. Deactivating WA leads to an average performance decrease of 5.63%, with a more pronounced impact on grade 7-12 tasks, showing a 6.35% accuracy drop. Third, the fully integrated agent system achieves peak performance at 93.73%, surpassing the best ablated configuration by a notable margin of 2.44%. This optimal configuration delivers consistent enhancements across all task categories, particularly excelling in multimodal scenarios with 3.70% improvement in text-based tasks and 4.80% in image-based tasks compared to the baselines. The framework also shows superior handling of complex queries, attaining 2.64% higher accuracy for grade 7-12 problems. These empirical outcomes substantiate the architectural effectiveness in orchestrating specialized agents for holistic multimodal reasoning.

5 CONCLUSION

In this paper, we introduced HM-RAG, a novel Hierarchical Multi-Agent Multimodal Retrieval-Augmented Generation framework designed to address the challenges of complex multimodal query processing and knowledge synthesis. HM-RAG pioneers collaborative intelligence by integrating specialized agents for query decomposition, multi-source retrieval, and decision refinement, enabling dynamic knowledge synthesis across structured, unstructured, and graph-based data. Through extensive experiments on the ScienceQA and CrisisMMD benchmarks, HM-RAG demonstrated state-of-the-art performance in the accuracy of multimodal question answering and classification, with significant improvements over all categories of baseline methods. Our work advances RAG systems by effectively addressing critical challenges in multimodal reasoning and knowledge synthesis, paving the way for more robust and adaptable information retrieval and generation systems in diverse application domains.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[3] Abhijit Anand, Vinay Setty, Avishek Anand, et al. 2023. Context Aware Query Rewriting for Text Rankers using LLM. *arXiv preprint arXiv:2308.16753* (2023).

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511* (2023).

[5] Sukanya Bag, Ayushman Gupta, Rajat Kaushik, and Chirag Jain. 2024. RAG Beyond Text: Enhancing Image Retrieval in RAG Systems. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*. IEEE, 1–6.

[6] Mirco Bonomo and Simone Bianco. 2025. Visual RAG: Expanding MLLM Visual Knowledge without Fine-tuning. *arXiv preprint arXiv:2501.10834* (2025).

[7] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. *arXiv preprint arXiv:2407.21439* (2024).

[8] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. 2024. Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects. *arXiv preprint arXiv:2401.03428* (2024).

[9] Yuxin Dong, Shuo Wang, Hongye Zheng, Jiajing Chen, Zhenhong Zhang, and Chihang Wang. 2024. Advanced RAG Models with Graph Structures: Optimizing Complex Knowledge Reasoning and Text Generation. In *2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*. IEEE, 626–630.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Gustavo de Aquino e Aquino, Nádila da Silva de Azevedo, Leandro Youiti Silva Okimoto, Leonardo Yuto Suzuki Camelo, Hendrio Luis de Souza Bragança, Rubens Fernandes, Andre Printes, Fábio Cardoso, Raimundo Gomes, and Israel Gondres Torné. 2025. From RAG to Multi-Agent Systems: A Survey of Modern Approaches in LLM Development. (2025).

[12] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From Local to Global: A GraphRAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130* (2024).

[13] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.

[14] Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. 2023. Relational Deep Learning: Graph Representation Learning on Relational Databases. *arXiv preprint arXiv:2312.04615* (2023).

[15] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* 2 (2023).

[16] Jeanie Genesis and Frazier Keane. 2025. Integrating Knowledge Retrieval with Generation: A Comprehensive Survey of RAG Models in NLP. (2025).

[17] Deepanway Ghosal, Navonil Majumder, Roy Ka-Wei Lee, Rada Mihalcea, and Soujanya Poria. 2023. Language Guided Visual Question Answering: Elevate Your Multimodal Language Model Using Knowledge-Enriched Prompts. *arXiv preprint arXiv:2310.20159* (2023).

[18] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779* (2024).

[19] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv preprint arXiv:2410.12837* (2024).

[20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *International Conference on Machine Learning*. PMLR, 3929–3938.

[21] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. *arXiv preprint arXiv:2503.13964* (2025).

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[23] Sameera Horawalavithana, Sai Munikoti, Ian Stewart, and Henry Kvinge. 2023. SCITUNE: Aligning Large Language Models with Scientific Multimodal Instructions. *arXiv preprint arXiv:2307.01139* (2023).

[24] Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2023. Tool Documentation Enables Zero-Shot Tool-Usage with Large Language Models. *arXiv preprint arXiv:2308.00675* (2023).

[25] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. *arXiv preprint arXiv:2403.12895* (2024).

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276* (2024).

[27] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models. *arXiv preprint arXiv:2208.03299* 1, 2 (2022), 4.

[28] Cheonsu Jeong. 2024. A Graph-Agent-Based Approach to Enhancing Knowledge-Based QA with Advanced RAG. *Knowledge Management Research* 25, 3 (2024), 99–119.

[29] Cheonsu Jeong. 2024. A Study on the Implementation Method of an Agent-Based Advanced RAG System Using Graph. *arXiv preprint arXiv:2407.19994* (2024).

[30] 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* (2023).

[31] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 39–48.

[32] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research. *arXiv preprint arXiv:2312.07559* (2023).

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*. PMLR, 19730–19742.

[35] Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2025. Topology-of-Question-Decomposition: Enhancing Large Language Models with Information Retrieval for Knowledge-Intensive Tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*. 2814–2833.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems* 36 (2023), 34892–34916.

[37] Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Botian Shi, and Ding Wang. 2025. Aligning Vision to Language: Text-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning. *arXiv preprint arXiv:2503.12972* (2025).

[38] Jiawang Liu, Ye Tao, Fei Wang, Hui Li, and Xiugong Qin. 2025. SiQA: A Large Multi-Modal Question Answering Model for Structured Images Based on RAG. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[40] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[41] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15630–15640.

[42] Costas Mavromatis and George Karypis. 2024. GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning. *arXiv preprint arXiv:2405.20139* (2024).

[43] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A Comprehensive Overview of Large Language Models. *arXiv preprint arXiv:2307.06435* (2023).

[44] Tyler Thomas Procko and Omar Ochoa. 2024. Graph Retrieval-Augmented Generation for Large Language Models: A Survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*. IEEE, 166–169.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. PmLR, 8748–8763.

[46] Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. *arXiv preprint arXiv:2410.21943* (2024).

[47] Tolga Şakar and Hakan Emekci. 2025. Maximizing RAG efficiency: A comparative analysis of RAG methods. *Natural Language Processing* 31, 1 (2025), 1–25.

[48] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2023), 68539–68551.

[49] 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. *arXiv preprint arXiv:2403.10081* (2024).

[50] Sabrina Toro, Anna V Anagnostopoulos, Susan M Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion M Dooley, William D Duncan, Petra Fey, et al. 2024. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *Journal of Biomedical Semantics* 15, 1 (2024), 19.

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191* (2024).

[53] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation. *arXiv preprint arXiv:2408.04187* (2024).

[54] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. *arXiv preprint arXiv:2410.13085* (2024).

[55] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[56] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023. MM-BigBench: Evaluating Multimodal Models on Multimodal Content Comprehension Tasks. *arXiv preprint arXiv:2310.09036* (2023).

[57] Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. RAFT: Adapting Language Model to Domain Specific RAG. In *First Conference on Language Modeling*.

[58] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models. *Advances in Neural Information Processing Systems* 36 (2023), 5168–5191.

[59] Zhiqiang Zhong and Davide Mottin. 2023. Knowledge-augmented Graph Machine Learning for Drug Discovery: From Precision to Interpretability. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5841–5842.

A PREDICTED EXAMPLES

We present additional predicted examples in Figure 3. Furthermore, we include two representative questions with or without image context to assess the models' language reasoning capabilities. HM-RAG consistently produces accurate answers.

Figure 4: Comparison on single-modal question answering.

Question: Which better describes the Daintree rain forest ecosystem?

Choices: (A) It has year-round rain. It also has soil that is poor in nutrients. √
(B) It has cold winters. It also has many different types of organisms.

Text-based

According to the provided information and context from the Document Chunks, the Daintree rain forest ecosystem in northeastern Australia better fits the description \"It has year-round rain. It also has soil that is poor in nutrients.\" This aligns with typical characteristics of a tropical rainforest, as opposed to an option describing cold winters and limited biodiversity.\n\nReferences:\n- [DC] The Daintree rain forest is a tropical rain forest ecosystem in northeastern Australia (File: #daintree_rain_forest_context).

Answer: The answer is (A).

Web-based

Daintree Rainforest Wikipedia\u3011\n The Daintree Rainforest, also known as the Daintree, is a region on the northeastern coast of Queensland, Australia, about 105 km (65 mi), by road, north of ...uff1a https://en.wikipedia.org/wiki/Daintree_Rainforest.

Answer: The answer is (A).

Graph-based

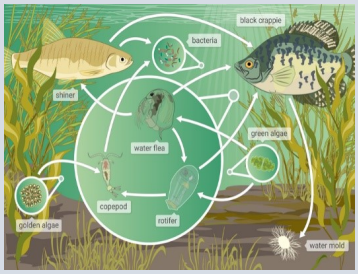
Based on the provided context, I will now generate a comprehensive summary about … [KG] Genes are fundamental units of heredity that carry biological instructions for traits and characteristics, passed from parents to offspring through generations.

Answer: The answer is (A).

HMAgent

It has year-round rain. It also has soil that is poor in nutrients.

Answer: The answer is (A).



Question: Which of the following organisms is the primary consumer in this food web?

Choices: (A) copepod √
(B) black crappie
(C) bacteria

Text-based

In a food web, primary consumers are organisms that eat producers (such as plants) or other primary consumers. From the given information and the context of Little Rock Lake, the black crappie (option B) is listed as a consumer in the ecosystem model. This indicates it feeds on smaller organisms within the lake, making it a primary consumer. n\nReferences:\n[DC] Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. (File: Document_Chunk_3.

Answer: The answer is (B). ✗

Web-based

Below is a food web from Little Rock Lake, a freshwater lake ecosystem in Wisconsin. black crappie bacteria shiner green algae water flea ...\n\u94fe\u63a5\uff1ahttps://www.gauthmath.com/solution/1794627532113926.

Answer: The answer is (A).

Graph-based

Based on the provided food web, black crappie likely feed on smaller organisms like copepods or zooplankton, making it a primary consumer in this ecosystem.\n\n**Answer: B. black crappie**\nReferences:\n1. [KG] Food Web Analysis in Ecosystems (File: KG_005).

Answer: The answer is (B). ✗

HMAgent

Bacteria crappie has anarrows pointing to it from the water flea, the rotifer and and the shiner. None of these organisms is a producer, so the black crappie is not a primary consumer.\nThe bacteria have anarrows pointing to them from the waterpod and the shiner. Neither the waterpod nor the shiner is a produce, so the are not primary consumers.

Answer: The answer is (A).

Figure 5: Comparison on multimodal question answering.