**BIORAG: A RAG-LLM Framework for Biological Question Reasoning Chengrui Wang**1,2**, Qingqing Long**1,2**, Meng Xiao**1,2**, Xunxin Cai**1,2**, Chengjun Wu**1,2**, Zhen Meng**1,2, **Xuezhi Wang**1,2, **Yuanchun Zhou**1,2\*   
1Computer Network Information Center, Chinese Academy of Sciences.

2University of the Chinese Academy of Sciences.

{crwang,qqlong,shaow,xxcai,cwu,zhenm99,wxz,zyc}cnic.cn

|  |  |  |  |
| --- | --- | --- | --- |
| arXiv:2408.01107v2 [cs.CL] 14 Aug 2024 | **Abstract** | |  |
| The question-answering system for Life sci- | |
| ence research, which is characterized by the | |
| rapid pace of discovery, evolving insights, and | |
| complex interactions among knowledge enti- | |
| ties, presents unique challenges in maintain- | |
| ing a comprehensive knowledge warehouse | |
| and accurate information retrieval. To address | |
| these issues, we introduce **BIORAG**, a novel | |
| Retrieval-Augmented Generation (RAG) with | |
| the Large Language Models (LLMs) frame- | |
| work. Our approach starts with parsing, index- | |
| ing, and segmenting an extensive collection of | |
| 22 million scientific papers as the basic knowl- | |
| edge, followed by training a specialized embed- | | Figure 1: An illustration of the difference between three |
| ding model tailored to this domain. Addition- | |
| ally, we enhance the vector retrieval process | | paradigms: (a) fine-tuned language model embedded |
| by incorporating a domain-specific knowledge | | domain knowledge into deep space; (b) RAG-based |
| hierarchy, which aids in modeling the intricate | | method retrieve supplementary information from con- |
| interrelationships among each query and con- | | structed knowledge base; (c) BIORAG adaptively select |
| text. For queries requiring the most current in- | | knowledge source and domain-specific tools to advance |
| formation, BIORAG deconstructs the question | | the biology question-reasoning task. |
| and employs an iterative retrieval process incor- | | thesis. To bridge the gap and facilitate multidisci- |
| porated with the search engine for step-by-step | |
| reasoning. Rigorous experiments have demon- | |
| strated that our model outperforms fine-tuned | | pline cooperation, automated question-reasoning |
| LLM, LLM with search engines, and other sci- | | systems (Auer et al., 2023) play a pivotal role in |
| entific RAG frameworks across multiple life | |
| enabling experts from diverse fields to effectively |
| science question-answering tasks. | |
| navigate and integrate this burgeoning and com- |
| **1** | **Introduction** | plex body of biological knowledge (Yang et al., |
| 2023). However, this ever-changing landscape and |
| Research and trends in the *Biology* have shown a | |
| the complex interplay between different knowledge |
| continuously evolving, marked by rapid discover- | |
| components present obstacles (Lee et al., 2023; |
| ies and the increasing complexity of its knowledge | |
| Castro Nascimento and Pimentel, 2023; Lecler |
| domains (Bertoline et al., 2023; Long et al., 2021b). | |
| et al., 2023; Song et al., 2020) in creating efficient |
| In addition, the growing trend for interdisciplinary | |
| domain-specific question-reasoning systems. |
| research between Biology and other fields (Lepore | |
| The prior literature partially addresses question- |
| et al., 2023; Xiao et al., 2023; Xiao et al.), such as | |
| reasoning in the biology domain and can be |
| artificial intelligence (Holzinger et al., 2023; Long | |
| grouped into two mainstream (Nguyen et al., 2024) |
| et al., 2021a), material science (Atkins et al., 2023), | |
| (as shown in Figure 1 (a-b)). **Fine-tuned Lan-** |
| and environmental science (Cole et al., 2021), fur- | |
| **guage Model** (Gu et al., 2021) includes models |
| ther amplifies the complexity of knowledge syn- | |
| like bioBERT (Lee et al., 2020), sciBERT (Beltagy |

\*Yuanchun Zhou is the corresponding author.

et al., 2019), and large language models tailored for

specific domains, such as PMC-Llama (Wu et al., 2024) and Llava-med (Li et al., 2024). These mod-els are trained on domain-specific corpora, thereby embedding deep domain knowledge within their architectures. *However, that embedded knowledge could be incomplete and computationally expen-sive to update.* **Retrieval-Agumented Generation** methods follow the information indexing and re-trieval, information augmentation, and answer gen-eration paradigm. For instance, PGRA (Guo et al., 2023) adopts a retriever to search and re-ranking the context, then generate the answer. Later re-search has aimed to improve these systems by ei-ther optimizing the retrieval processes using prior answers (Wang et al., 2023), enhancing model func-tionality through iterative feedback cycles (Liu

pora. BIORAG then addresses the complexity of biological knowledge systems by combining a pre-built research hierarchy with an embedding model for accurate context retrieval. To cope with emerg-ing biology knowledge, BIORAG can adaptively select knowledge sources from search engines, ex-isting domain-specific tools, or indexed research articles. Once the framework determines that it has gathered sufficient information, it will generate the answer based on the reasoned material.

We illustrate the question-reasoning power of BIORAG on 6 popularly used biology QA datasets and compare it against 6 baseline methods. Ex-tensive case studies show the great potential to apply this framework to general science question-reasoning scenarios.

|  |  |  |
| --- | --- | --- |
| et al., 2024), or expanding the knowledge base | **2** | **Biological Retrieval-Augmented Generation LLM Framework** |
| with search engines to incorporate the latest infor- |
| mation (O’Donnell, 2023). Although RAG-based |

methods address the issue of updating information, *they often oversee the intricate complexities inher-ent in the domain knowledge of biology.*

Based on the aforementioned discussion, we summarize three challenges in building efficient biology question-reasoning systems: **(C1) The scarcity of high-quality domain-specific corpora.** While biological research publications are abun-dant, there remains a significant void in the avail-ability of extensive, high-quality datasets to build robust information indexing models. **(C2) The inherent complexity of biological knowledge sys-tems.** This complexity is compounded by the inter-disciplinary nature of modern biological research. Consequently, automated question-reasoning sys-tems must be able to understand and process multi-faceted and often ambiguous biological query. **(C3) The continual updating of knowledge.** Biology is a dynamic field where discoveries are frequently made, and existing theories are regularly revised or replaced. This fluidity necessitates that question-reasoning systems adeptly select the knowledge source from databases or contemporary search en-gines to reflect the correct scientific understanding.

**Our Perspective and Contributions:** To solve the above challenges, we proposed BIORAG, a novel Retrieval-Augmented Generation framework integrated with Large Language Models for bi-ological question-reasoning. To obtain a robust domain-specific information indexing embedding model, we start by parsing, indexing, and segment-ing extensive research articles from the biology domain and constructing high-quality training cor-

In this paper, we propose the ***Bio****logical* ***R****etrieval-****A****ugmented* ***G****eneration LLM Framework*, namely BIORAG (as shown in Figure 2). In the following sections, we first introduce the preliminary step of constructing a high-quality local information source and training the biological domain-specific information indexing embedding model. For ques-tions that require the most current or other domain-related data, we introduce external information sources. Then, we demonstrate the knowledge hierarchy-based query pre-processing, retriever ex-ecution component, and how the model iteratively collects sufficient information. Finally, the large language model will generate the answer based on the information obtained. The details of cus-tomized prompts are given in Section 2.4.

**2.1 Internal Biological Information Source**

High-quality domain-specific corpora are crucial for enriching the information source and enhancing the embedding model in the context of biological question-reasoning systems. To achieve this goal, we extract research papers from the global biomedi-cal article database maintained by the National Cen-ter for Biotechnology Information1(NCBI) (Schoch et al., 2020). This extensive repository aggregates over 37 million scientific citations and abstracts spanning from the 1950s to the present, encompass-ing a broad array of biomedical fields, including clinical medicine, molecular biology, etc. For the purposes of this study, we utilize the abstracts from

1<https://www.ncbi.nlm.nih.gov/>

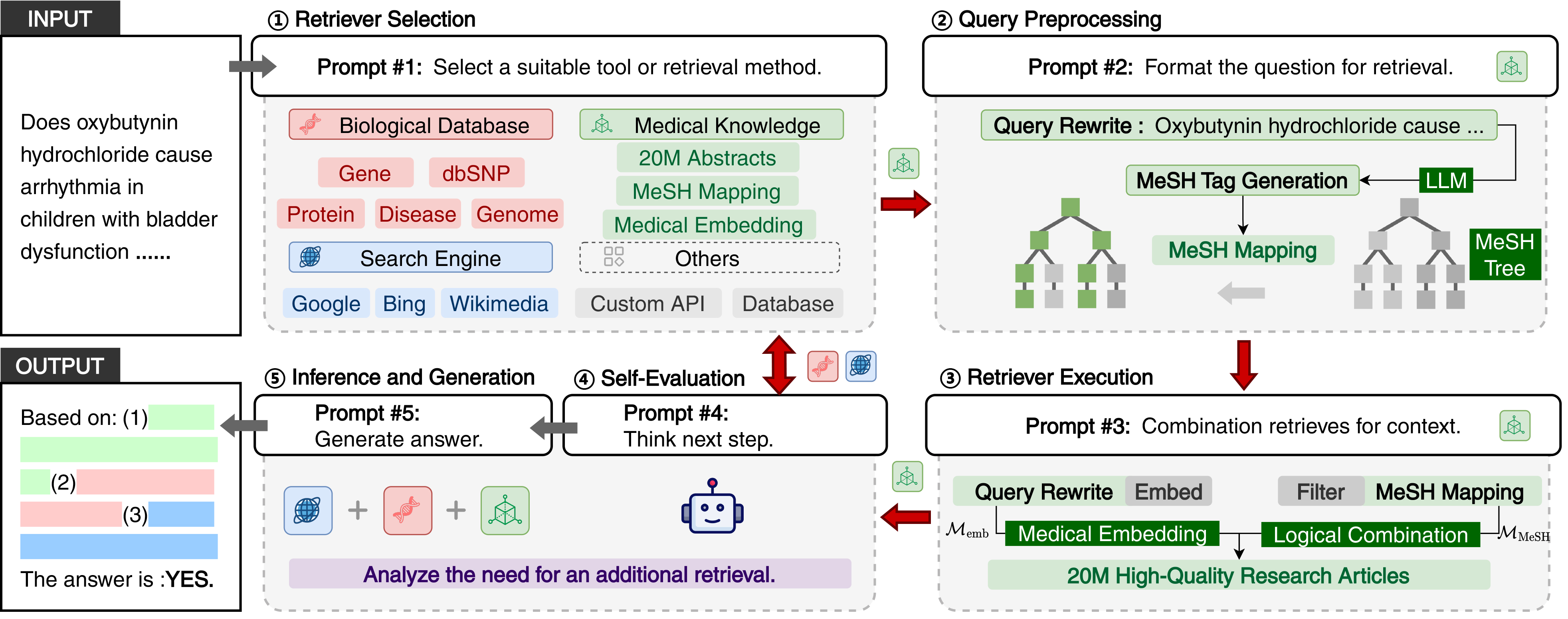


Figure 2: The architecture of our proposed BIORAG framework. The pipeline consists of five iterative components designed to enhance the process of biological question-reasoning: ①**Retriever Selection** aims to choose the most ideal information source; ②**Query Pre-processing** aims to rewrite the query and find closed topic tag from pre-defined knowledge hierarchy; ③**Retriever Execution** aims to combination retrieve the correlated context from knowledge base; ④**Self-Evaluation** assess the adequacy of the retrieved information and decides whether to cycle through additional retrieval tools or to move to the next phase; ⑤**Inference and Generation** uses the information gathered to generate an informed and accurate answer to the biological query.

these PubMed papers as the supporting corpus for the BIORAG framework.

*Local Data Preparation:* Specifically, we initially downloaded over 37 million original papers from which we subsequently filtered out 14 million en-tries deemed to be of low quality. The preprocess-ing of these texts was conducted using the *Unstruc-tured* tool2, specifically designed to ingest and pre-process unstructured textual data effectively. Our filtration process involved the removal of gibberish using regular expression techniques, as well as the exclusion of non-semantic content such as hyper-links, charts, tables, and other embedded tags. This meticulous process yielded a corpus of 22,371,343 high-quality, processed PubMed abstracts.

*Information Indexing:* To further refine the re-trieval performance of abstracts tailored to spe-cific biological questions, we developed a spe-cialized biological embedding model within the BIORAG framework. This model employs Pub-MedBERT (Gu et al., 2021) as the foundational model. We enhanced this model using the CLIP (Contrastive Language-Image Pretraining) tech-nique (Li et al., 2021; Nussbaum et al., 2024), al-lowing us to fine-tune the model, denoted as *M*emb. Based on this, we constructed a local, high-quality biological vector database (Xian et al., 2024) to sup-port efficient and effective query processing and re-

2<https://github.com/Unstructured-IO>

trieval operations. This database serves as a critical resource in facilitating rapid and accurate access to relevant biomedical information, significantly ad-vancing the capabilities of our BIORAGframework in handling complex biological questions.

**2.2 External Information Sources**

External biology knowledge is crucial to biological reasoning due to the rapidly evolving nature of biological research, which continuously integrates new discoveries. To address this challenge, we introduce two external information sources.

*Biological Data Hub:* In BIORAG, we harness several specialized biological Hubs to ensure the accuracy of experimental data and to pro-vide detailed biological insights. Specifically, BIORAGintegrates the following databases, each serving a unique purpose in the broader context of biological analyses: (1) Gene Database3: This resource provides comprehensive information on the functions, structures, and expressions of spe-cific genes. It is invaluable for addressing queries related to gene mechanisms, gene actions, and gene expressions, facilitating a deeper understanding of gene-related phenomena. (2) dbSNP Database4: This database houses a vast repository of single nucleotide polymorphisms (SNPs), offering critical insights into genetic variants and their potential as-

3<https://www.ncbi.nlm.nih.gov/gene/>4[https://www.ncbi.nlm.nih.gov/snp/](https://www.ncbi.nlm.nih.gov/gene/)

sociations with various diseases. It is instrumental for studies exploring the genetic basis of disease and trait inheritance. (3) Genome Database5: Pro-viding complete genome sequences, this database is essential for studying the structure, function, and evolution of genomes across different organisms. It supports comprehensive genomic analyses and comparative studies, enhancing our understanding of genomic architecture and its functional implica-tions. (4) Protein Database6: This resource offers detailed information about the sequences, struc-tures, and functions of proteins. It is crucial for exploring protein-related biological processes, un-derstanding molecular functions, and investigating the complex interactions within the proteome.

*Search Engine:* To ensure access to the most cur-rent discussions and developments, BIORAG in-corporates a variety of search engines, including Google, Bing, arXiv, Wikimedia, and Crossref. Each platform contributes uniquely to the aggrega-tion of information: (1) Google and Bing: These search engines scour the web for a diverse range of content, including news articles, blogs, and forums, providing insights into public discussions and con-cerns related to scientific topics. This breadth of information is crucial for understanding the so-cietal impact and general discourse surrounding scientific issues. (2) arXiv: As a repository for preprint papers, arXiv offers access to the latest re-search reports and scholarly articles across multiple scientific disciplines before they undergo peer re-view. This source is invaluable for staying abreast of the newest scientific theories and experiments. (3) Wikimedia: Known for its user-friendly content, Wikimedia offers easily digestible explanations of complex scientific concepts and principles. This re-source helps simplify advanced topics for broader public understanding and educational purposes. (4) Crossref: This service acts as a comprehensive ag-gregator of academic citation data, providing links to peer-reviewed scholarly publications and their ci-tation networks. Crossref is essential for accessing high-quality research outputs and understanding their impact on the academic community.

|  |  |
| --- | --- |
| **2.3** | **Self-evaluated Information Retriever** |

external information source, BIORAG is firstly tasked with comprehending the complex disci-Following the construction of the internal and

5<https://www.ncbi.nlm.nih.gov/genome/>6<https://www.ncbi.nlm.nih.gov/protein/>

|  |
| --- |
| Based on the **QUESTION**, analyze the related MeSH terms to format them properly.  **QUESTION**: [.....]  **MeSH**: [*κ*1, *κ*2, ...] |

Figure 3: Training Template for *M*MeSH.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | **Input Question** |   What are the differences between innate immunity and adaptive immunity?  **Predicted MeSH by** *M***MeSH**  [Adaptive Immunity, Animals, ...]   |  | | --- | | **Generated SQL** |   "**filtered by**":[eq("MeSH", "Adaptive Immunity") or eq("MeSH", "Animals") or ...],  "**ordered by**": embedding similarity |

Figure 4: An example of MeSH filtering SQLs Genera-tion.

plinary framework of the life sciences to retrieve the most relevant information accurately. More-over, BIORAG integrates a self-evaluation mecha-nism to continuously assess the adequacy and rele-vance of the information it has collected.

*Internal Information Retrieve:* To effectively nav-igate the inherent complexity of biological knowl-edge systems, BIORAG leverages an integrated approach, combining a well-defined hierarchical structure with indexed information to conduct a comprehensive internal information retrieval. The Medical Subject Headings7(MeSH) thesaurus is popularly used for indexing, cataloging, and search-ing for biomedical-related information and research papers. Specifically, we first train a model *M*MeSH to predict MeSH of the input questions. We then use the templates in Figure 3 for fine-tuning a Llama3-8B model to classify given questions. Af-ter that, we construct MeSH filtering SQLs (as shown in Figure 4) to generate the scalar condition retrieval. A candidate result is considered relevant to the given question because it has one consistent MeSH with the question. Then, the vector retrieval process is adopted to sort the relative results based on the cosine similarity of the sentence embedding between the input questions and the filtered results.

*Self-evaluation Strategy:* In order to ensure the ac-curacy and contemporary of the retrieved informa-tion, BIORAG incorporates a self-evaluation strat-egy that assesses the adequacy of data collected from the internal knowledge base. In detail, this

7<https://www.nlm.nih.gov/mesh/meshhome.html>

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **LLM** |  | **BioLLM** | | **SciRAG** | | **BioRAG** |
| GPT3.5 | Llama3-8B | Llama-70B | PMC-Llama | BioMistral | GeneGPT | NewBing |  |

**Nomenclature**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene alias | 7 | 0 | 0 | 0 | 8 | 84 | 68 | **98** |
| Gene name conversion | 0 | 0 | 0 | 0 | 0 | **100** | **100** | **100** |

**Genomic location**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene SNP association | 0 | 0 | 0 | 0 | 0 | **100** | 0 | **100** |
| Gene location | 9 | 20 | 28 | 14 | 12 | 66 | 70 | **86** |
| SNP location | 5 | 48 | 94 | 0 | 0 | 98 | **100** | **100** |

**Functional analysis**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene disease association | 31 | 0 | 0 | 0 | 8 | 66 | 64 | **71** |
| Protein-coding genes | 54 | 6 | 12 | 40 | 80 | **100** | **100** | **100** |

Table 1: Performance of BioRAG compared to other RAG-LLMs on the GeneTuring QA dataset.The scores represent accuracy.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **LLM** |  | **BioLLM** | | **SciRAG** | | **BioRAG** |
|  | GPT3.5 | Llama3-8B | Llama-70B | PMC-Llama | BioMistral | GeneGPT | NewBing |  |
| MedMCQA | 54 | 51 | 71 | 56 | 49 | 0 | 55 | **73** |
| Medical Genetics | 74 | 51 | 67 | 28 | 67 | 0 | **88** | **88** |
| College Biology | 73 | 75 | 88 | 30 | 67 | 0 | 71 | **90** |
| College Medicine | 65 | 61 | 70 | 23 | 51 | 0 | **78** | **78** |

Table 2: Performance of BioRAG compared to other RAG-LLMs on the biological-related QA benchmarks.The scores represent accuracy. **Bold** and underlined results denote the highest and second-highest performance, respec-tively.

critical evaluation is driven by the backend large language model which aims to determine whether the information retrieved internally is sufficient to address the posed question substantively. If the internal content is insufficient, the model will loop back to pertinent external knowledge sources. Ad-ditionally, when the initial assessment indicates that the scientific questions require broader searches or retrieval of entity-specific data, the model tends to deploy external tools. This methodology supports the framework’s goal of providing precise, up-to-date, comprehensive answers, facilitating more in-formed decision-making, and advancing research and applications in the life sciences.

**2.4**  **Customized Prompts Detail**

To maximize the effect of the retrieved corpus and knowledge, we design customized prompts in BIORAG. The prompts in Figure. 2 is detailed defined as follows,

• **Prompt # 1**: To provide the most helpful and accurate response to the following Question: *{Question}*. You have been given descrip- tions of several RETRIEVAL METHODS: *{Retrieval}*. Please select the RETRIEVAL

METHODS you consider the most appropri-ate for addressing this question.

• **Prompt # 2**: Based on the RETRIEVAL METHODS you selected, and considering the *Question* and the *Input Requirements* of the retrieval method, please REWRITE the search query accordingly.

• **Prompt # 3**: Now, using the rewritten QUERY and the retrieval FILTER methods, perform a logical combination to execute the search effectively.

• **Prompt # 4**: Based on the RETRIEVAL RE-SULTS from the above steps, please evaluate whether the RESULTS support answering the original *Question*. If they do not support it, output "**NO**". If they do support it, output "**YES**".

• **Prompt # 5**: Based on the RETRIEVAL RE-SULTS, perform a comprehensive reasoning and provide an answer to the *Question*.

Furthermore, we designed instruction manuals for specialized biological tools and databases, aim

at exploiting their potentialities. These instructions are shown as follows,

• **Manual #Gene**: The Gene database search engine is a valuable tool for retrieving com-prehensive information about genes, including gene structure, function, and related genetic events. It is particularly useful for answer-ing detailed questions regarding gene-related research and findings. To utilize this search

• **Manual #PubMed**: The PubMed local vector database search engine is an advanced tool de-signed for retrieving biomedical literature and research articles using vector-based search techniques. It is particularly useful for answer-ing detailed questions about medical research, clinical studies, and scientific discoveries. To utilize this search engine effectively, the input should be a specific query or topic of interest.

|  |  |  |
| --- | --- | --- |
| engine effectively, the input must be a specific | **3** | **Results & Analysis** |
| gene name. |

• **Manual #dbSNP**: The dbSNP database search engine is an essential tool for retrieving detailed information about single nucleotide polymorphisms (SNPs) and other genetic vari-ations. It is particularly useful for answering questions related to genetic diversity, allele frequency, and related genetic studies. To uti-lize this search engine effectively, the input must be a specific SNP identifier or genetic variant name.

• **Manual #Genome**: The Genome database search engine is an indispensable tool for ac-cessing comprehensive information about en-tire genomes, including their sequences, an-notations, and functional elements. It is par-ticularly useful for answering complex ques-tions about genomic structures, variations, and comparative genomics. To use this search en-gine effectively, the input must be a specific genome name or identifier.

• **Manual #Protein**: The Protein database search engine is a crucial resource for obtain-ing detailed information about proteins, in-cluding their sequences, structures, functions, and interactions. It is particularly useful for answering questions related to protein biol-ogy, biochemical properties, and molecular function. To use this search engine effectively, the input must be a specific protein name or identifier.

• **Manual #Web Search**: The Web Search En-gine is a powerful tool designed to help you find information about current events quickly and efficiently. It is especially useful for ob-

**3.1 Datasets**

We conduct experiments on 6 popularly used biological-related QA datasets to evaluate our pro-posed BIORAG, i.e., GeneTuring (Hou and Ji, 2023), MedMCQA (Pal et al., 2022), Medical Genetics (Hendrycks et al., 2020), College Biol-ogy (Hendrycks et al., 2020), College Medicine (Hendrycks et al., 2020). Note that the GeneTuring dataset contains more specialized biological ques-tions. It contains 12 tasks, and each task has 50 question-answer pairs. We use 7 GeneTuring tasks that are related to NCBI resources to evaluate the proposed BIORAG. The chosen tasks are classified into three modules and briefly described as follows,

• **Nomenclature:** This is about gene names. The objectives of the gene alias task and name conversion task are finding the official gene symbols for their non-official synonyms.

• **Genomics location:** The tasks are about the locations of genes, single-nucleotide polymor-phism (SNP), and their relations. We include the gene location, SNP location, and gene SNP association tasks. The first two tasks ask for the chromosome locations of a gene or an SNP, and the last one asks for related genes for a given SNP.

• **Functional analysis** asks for gene functions. We use the gene-disease association task where the goal is to return related genes for a given disease, and the protein-coding genes task which asks whether a gene is a protein-coding gene or not.

|  |  |  |
| --- | --- | --- |
| taining the latest news, updates, and devel- | **3.2** | **Baslines** |
| opments on a wide range of topics. To use |

this search engine effectively, simply enter a relevant search query.

We compare BIORAGwith various baselines, which can be classified into three categories,

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Data** |  | **Component** | | | **Base Model** | |
| D1 | D2 | D3 | C1 | C2 | C3 | M1 | M2 |

**Nomenclature**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene\_location | 74 | 94 | 96 | 98 | 90 | 91 | 88 | 98 |
| SNP\_location | 50 | 100 | 100 | 100 | 100 | 100 | 92 | 100 |

**Genomic location**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene\_SNP\_association | 6 | 100 | 98 | 100 | 6 | 100 | 94 | 100 |
| Gene\_disease\_association | 82 | 60 | 84 | 84 | 84 | 36 | 70 | 86 |
| Protein\_coding\_genes | 20 | 100 | 100 | 100 | 18 | 100 | 100 | 100 |

**Functional analysis**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene\_name\_conversion | 64 | 70 | 62 | 70 | 66 | 32 | 64 | 71 |
| Gene\_alias | 100 | 100 | 100 | 100 | 98 | 80 | 92 | 100 |

Table 3: Ablation study on the GeneTuring dataset.The scores represent accuracy.

• **LLM** (General LLMs): We select GPT-3.5-Turbo (175B), Llama3-8B (8B), Llama-70B (70B) as representative baselines.

• **BioLLM** (Biological LLMs): PMC-Llama (13B) (Wu et al., 2024) and BioMistral (7B) (Labrak et al., 2024) are two medical LLMs. They are pre-trained on open-source biomedi-cal texts.

• **SciRAG** (Scientific RAG-LLM framework): GeneGPT (175B) (Jin et al., 2024) is a bio-logical RAG-LLM framework that integrates the NCBI databases, i.e., Gene, dbSNP, OIMI, and Blast. NewBing8(>400B) is a retrieval-augmented LLM that has access to relevant web pages retrieved by Bing.

**3.3**  **Experimental Settings**

We take the Llama3-70B as the basic language model of BIORAG. For our embedding model *M*emb, we take AdamW as the optimizer and fine-tune 2 epochs. The number of retrieved results by biological databases, search engines, and local PubMed databases are set to 10, 10, and 4, respec-tively. The max iteration of self-evaluation is set to 15. If the model does not output the final an-swer within 15 times, BIORAG stops the iteration and outputs the current wrong answer. We use the *accuracy* to verify the overall performance. For the GeneTuring dataset, we only consider *exact* matches between model predictions and the ground truth as correct predictions for all nomenclature and genomics location tasks. For the gene-disease association task, we measure the recall as in the

8<https://www.bing.com/chat>

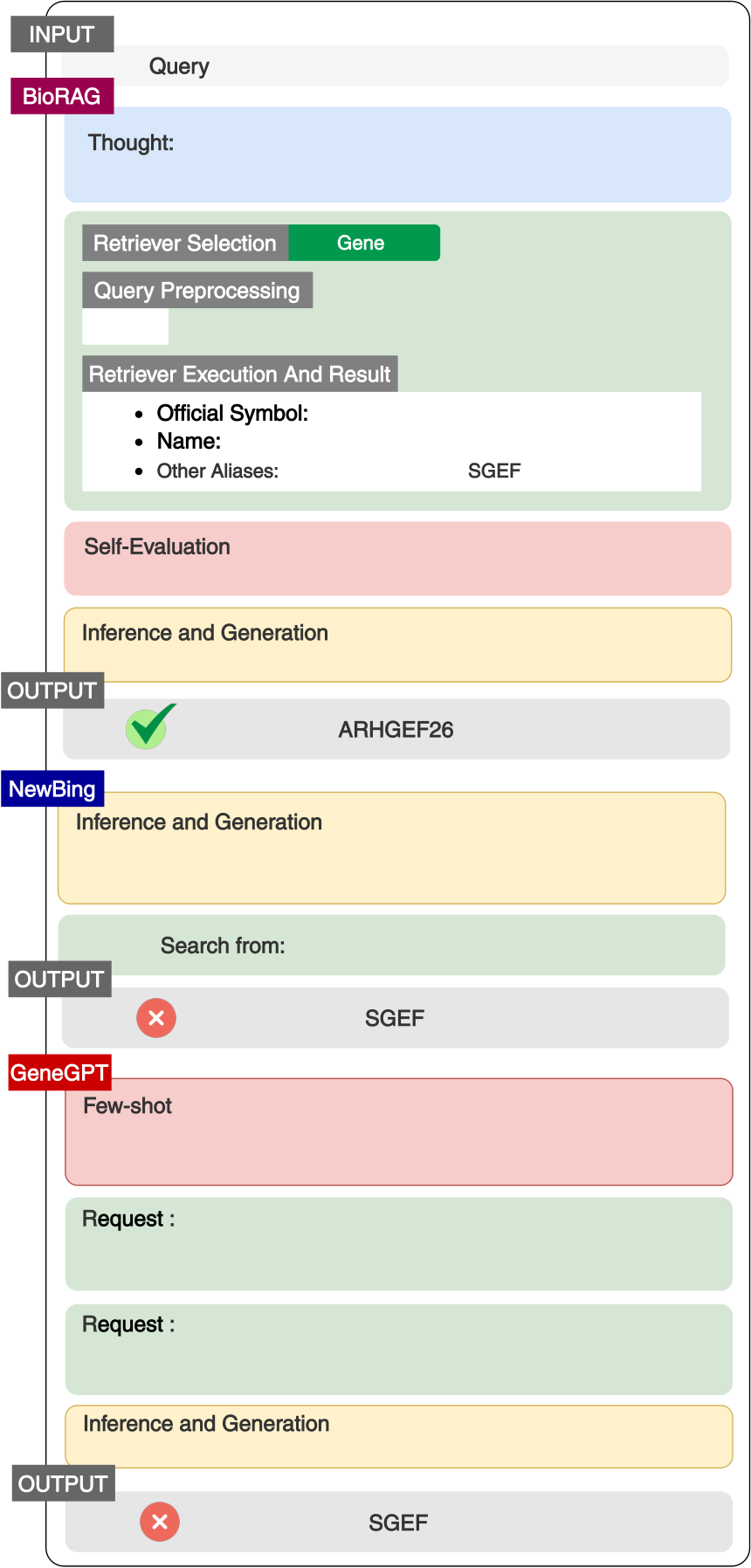
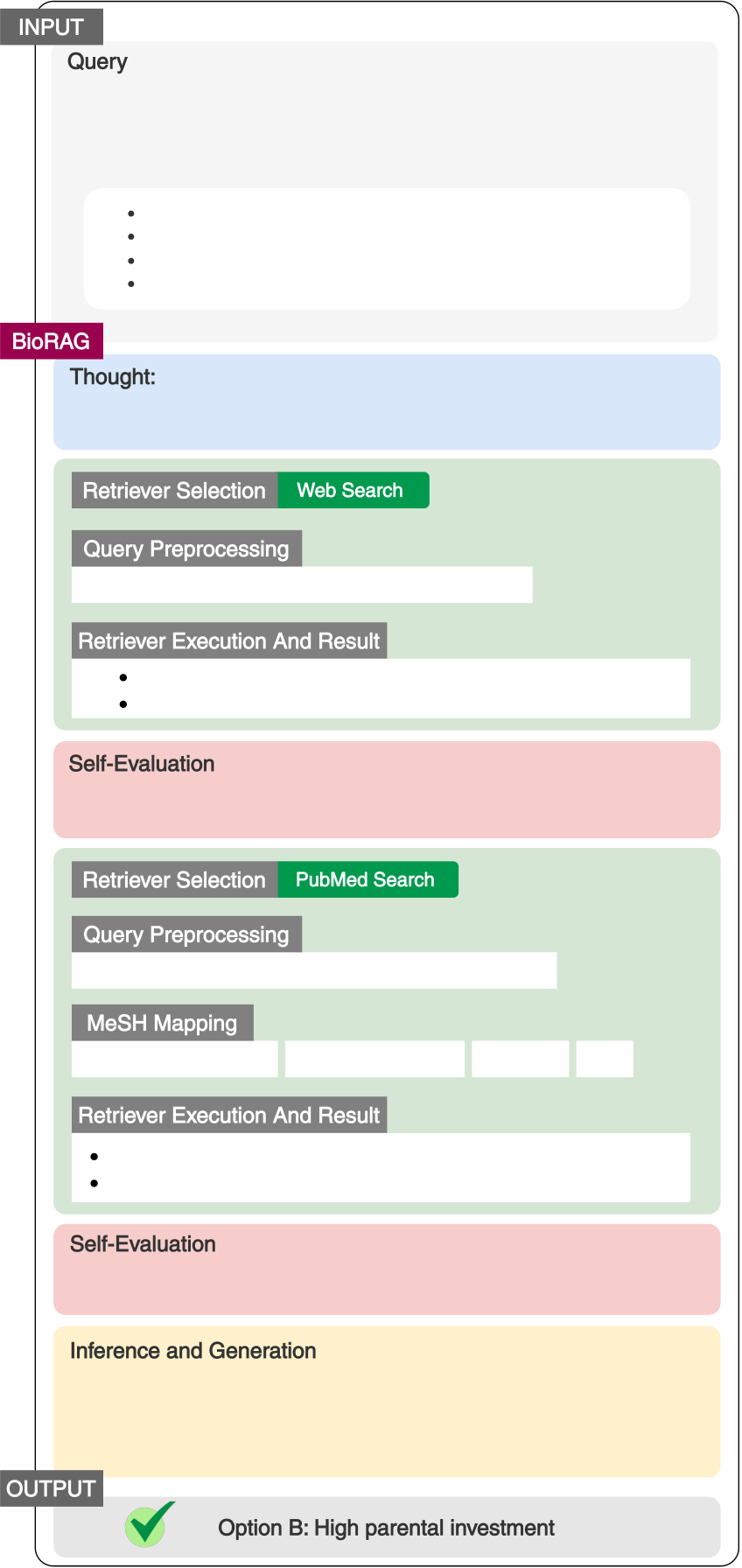
original dataset but based on *exact* individual gene matches. For the protein-coding genes task, we consider *exact* matches as correct after applying a simple vocabulary mapping that converts model-predicted "yes" / "no" to "TRUE" / "NA" and Latin species names to their informal names, respectively. The final answer of other datasets is "yes" / "no".

**3.4 Results on Biological-related Tasks**

To verify the effectiveness of the proposed model, we first conduct biological QA tasks. Results are shown in Table 2. We conclude with the follow-ing findings: (1) Based on the results of BioLLMs and GPT-3.5, we conclude that fine-tuning domain-specific data is helpful for domain-specific tasks. As the size of BioLLMs is much smaller than GPT-3.5, their performance is on par with GPT-3.5. (2) BIORAG performs better than BioLLMs and GPT-3.5, it indicates the effectiveness of local and exter-nal data sources. (3) Though the size of BIORAG is much smaller than SciRAG (NewBing), it has better performance. The gain comes from two as-pects. The first one is our customized prompts. The second aspect lies in the local and external information sources. NewBing has no access to specialized databases and lacks technical biologi-cal descriptions for reasoning. (4) GeneGPT scores 0% accuracy in this task, because it is a customized model for the GeneTuring dataset, resulting in poor generalization capabilities.

**3.5 Specialized Biological Reasoning Results**

The GeneTuring dataset contains more specialized biological questions, and the corresponding reason-ing process highly relies on technical biological corpus and descriptions. Results are shown in Ta-



|  |  |  |
| --- | --- | --- |
| INPUT  Query **：**A species of small rodent eats seeds from only one species of pine. In normal years, a pair of these rodents will have a litter of two or three. It is unusual for small rodents to have such small litter sizes. The rodents are most likely to exhibit which other characteristic? | INPUT | |
| BioRAG | Query**：**What is the official gene symbol of SGEF? |
| Thought:  I can use the Gene tool to find the official gene symbol of SGEF. | |

*Option A: Moderate sexual size dimorphism*

|  |  |  |  |
| --- | --- | --- | --- |
| *Option B: High parental investment*  *Option C: Precocial young*  *Option D: Frequent extrapair matings*  BioRAG  Thought:  This question is about rodent behavior in biology and ecology. I'll start with a web search for general information. | | Retriever Selection | Gene |
| Query Preprocessing  SGEF  Retriever Execution And Result  Official Symbol: ARHGEF26  Name: Rho guanine nucleotide exchange factor 26 Other Aliases: CSGEF , HMFN1864, SGEF | |
| Retriever Selection | Web Search |

Query Preprocessing   
Small rodent species with small litter sizes

Retriever Execution And Result   
Most rodents are small animals with robust bodies...

The largest litter I have heard of was 32 babies...

Self-Evaluation**：**  
The necessary information has been obtained to form an answer.

Inference and Generation**：**  
The official gene symbol for SGEF is ARHGEF26.

OUTPUT

ARHGEF26

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Self-Evaluation**：** Small litter sizes in rodents may adapt to their specialized diet,  needing more parental investment. Exploring further. | | | | NewBing  Inference and Generation**：** The official gene symbol for SGEF is SGEF (Src homology 3 domain-containing Guanine nucleotide Exchange Factor).  Search from: https://en.wikipedia.org/wiki/SGEF OUTPUT  SGEF  GeneGPT |
| Retriever Selection | PubMed Search | | |
| Query Preprocessing  Parental investment and litter size in rodents  MeSH Mapping | | | |
| Parental Investment | | Parental Behavior Rodentia | ... |

Retriever Execution And Result

The average litter sizes and reproductive performance... Previous analysis of the rules regarding how much more...

Self-Evaluation**：**  
Small litter sizes in rodents may be an adaptation to their specialized diet, requiring more parental investment.

Inference and Generation**：**  
Since the rodent species has a small litter size, it's possible that they invest more in each offspring. This could lead to a higher parental investment, which might be related to option B: High parental investment.

Few-shot**：**  
Hello. Your task is to use NCBI Web APIs to answer genomic

questions. Here are some examples (Entrez API) ...

Request :

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?

db=gene&term=SGEF

Request :

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?

db=gene&id=1956

Inference and Generation**：**  
The official gene symbol for SGEF is SGEF

|  |  |  |  |
| --- | --- | --- | --- |
| OUTPUT | Option B: High parental investment | OUTPUT | SGEF |

Figure 5: A case study selected from the College Biol-ogy dataset.

ble 1. As this dataset does not contain the train data, BioLLMs are performed directly without fine-tuning. Their bad results indicate their poor gener-alization. In this dataset, we focus on the analyses of GeneGPT, NewBing, and BIORAG (1) For the **nomenclature tasks**, the performance of BIORAG and GeneGPT rank first and second respectively, as both of them have access to the Gene database. BIORAG integrates the results of search engines while GeneGPT does not, and this brings the gap. (2) The reasoning behind **genomic location tasks** relies on the highly specialized Gene and dbSNP database. BIORAG and GeneGPT achieve 100% accuracy in the gene SNP association sub-task, as

Figure 6: A case study conducted on the gene alias task in the GeneTuring dataset.

both of them have access to the dbSNP database. However, NewBing has no access to the dbSNP database, thus it gets 0% accuracy in this task. For the gene location subtask, the challenge is the vari-ants of gene names. The interface of GeneGPT does not support advanced search, thus the re-trieved names are not comprehensive. In contrast, general search engines, such as NewBing, have better retrieved results when the query entity has variants or ambiguities. Thus NewBing has a better performance in this task than GeneGPT. BIORAG supports the above two kinds of interfaces, and achieves the best results in this task. (3) **Functional analysis tasks** rely on both the Gene database and



|  |  |  |
| --- | --- | --- |
| BioRAG | INPUT | Query**：**What are genes related to B-cell immunodeficiency? |
| NewBing |

Thought:

|  |  |  |
| --- | --- | --- |
| Use Gene tool to find genes related to B-cell immunodeficiency. | | Inference and Generation**：** Here are some notable B-cell immunodeficiency syndromes and  associated genes: |
| Retriever Selection | Gene |

Query Preprocessing   
B-cell immunodeficiency.

Retriever Execution And Result   
Official Symbol: TOP2B   
Name: DNA topoisomerase II beta Other Aliases: BILU, TOPIIB, top2beta

Self-Evaluation**：**  
Use the PubMed tool to conduct further searches on genes and

X-linked agammaglobulinemia (XLA):   
Gene: Bruton tyrosine kinase (BTK)   
Function: Crucial for B cell development; mutations lead to a lack of mature B cells and antibody production1.

Common variable immunodeficiency (CVID):   
Genes: Multiple genes (e.g., CTLA4, IKZF1, PIK3CD, STAT3)   
Features: Preserved B cell numbers but impaired antibody production and low immunoglobulin levels2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| diseases. | | | ... | Other B-cell disorders: |
| Retriever Selection | PubMed Search | |
| Gene panel: ADA, ADA2, AICDA, BLNK, CD19, CD27, |
| CD40, CD40LG, and more |
| Query Preprocessing | | | Purpose: Detect single nucleotide variants and copy |
| number changes associated with B-cell disorders3. |
| TOP2B B-cell immunodeficiency | | |
| MeSH Mapping | | | Remember that B cells play a crucial role in immune responses, |
| Immunoglobulin G1 Fragments | | Immunologic Factors | and defects in these genes can lead to recurrent infections and |
| Retriever Execution And Result | | | other complications14. If you have specific symptoms or |
| concerns, consider consulting a healthcare professional for |
| B cell development is a highly ... | | | personalized advice. 😊 |
| Topoisomerase 2β (TOP2B) introduces ... | | |
| Self-Evaluation**：** Use the web search tool to find general information. | | | Search from: https://geekymedics.com/b-cell-and- |
| immunoglobulin-immunodeficiencies |
| Retriever Selection | Web Search | | OUTPUT |
| Query Preprocessing | | | BTK |

TOP2B B-cell immunodeficiency

Retriever Execution And Result GeneGPT

Sequencing of patients with B cell-restricted...

The role of TOP2B in the immune system was revealed ...

Self-Evaluation**：**  
Use the PubMed tool to conduct further searches on genes and diseases, specifically focusing on TOP2B B-cell   
immunodeficiency syndrome.

Few-shot**：**  
Hello. Your task is to use NCBI Web APIs to answer genomic

questions. Here are some examples (Entrez API) ...

1. Gene **：**  
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?

db=gene&retmax=5&retmode=json&sort=relevance&term=LMP10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Retriever Selection | PubMed Search | | ... | 2. dbSNP**：** https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi? |
| Query Preprocessing | | | db=snp&retmax=10&retmode=json&id=1217074595 |
| TOP2B B-cell immunodeficiency syndrome | | | 3. BLAST: |
| MeSH Mapping | | | https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi? |
| CMD=Put&PROGRAM=blastn&.........ATCGGATGACAGAG........ |
| Immunologic Deficiency Syndromes | | B-Lymphocytes | Request : https://www.ncbi.nlm.nih.gov/gene/?term=B- |
| Retriever Execution And Result | | |
| DNA Topoisomerase IIβ (TOP2B) ...  Background: BCL11B encodes B cells ... | | |
| cell+immunodeficiency |

Result: HTML + CSS + JavaScript Code

Self-Evaluation**：**  
TOP2B is associated with B-cell immunodeficiency syndrome

Inference and Generation**：**  
TOP2B is associated with B-cell immunodeficiency syndrome

Inference and Generation**：**  
B-cell immunodeficiency is the gene associated with B-cell

immunodeficiency

|  |  |  |  |
| --- | --- | --- | --- |
| OUTPUT | TOP2B | OUTPUT | B-cell immunodeficiency |

Figure 7: A case study conducted on the gene disease association task in GeneTuring dataset.

relative PubMed papers. The PubMed corpus pro-vides detailed gene-disease relationships. Although NewBing retrieves the metadata, BIORAG com-bines the local PubMed database with other spe-cialized databases to achieve the best results.

nents within the BIORAG framework to achieve optimal performance in biological question reason-ing tasks. By understanding the contribution of each component, we can better optimize BIORAG for different tasks and datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **3.6** | **Ablation Study** | **3.7** | **Case Study** |

To evaluate the contribution of each component of BIORAG, we performed an extensive ablation study using the GeneTuring dataset, systematically removing individual components to assess their im-pact on performance across various tasks. This study was designed to isolate the effects of dif-ferent databases, components, and base models, with the experiments categorized as follows: (1) **Databases**: We consider three variations to eval-uate the effectiveness of each data sources of our database: **D1**: BIORAGwithout the Gene database; **D2**: BIORAGwithout general search engines. **D3**: BIORAGwithout the local PubMed database. (2) **Model Components**: We investigate the impact of specific components of our proposed frame-work: **C1**: BIORAGwithout the MeSH Filter; **C2**: BIORAGwithout the Query Rewrite component; **C3**: BIORAGwithout the Self-Evaluation mecha-nism. (3) **Base Models**: We compare the perfor-mance when using two different base LLM models: **M1**: take Llama-3-8B as the basic LLM, and **M2**: take Llama-3-70B as the basic LLM of BioRAG. Based on the results of ablation study, we high-lights the following key findings: (1) **Impact of Databases:** The results indicate that the Gene database (D1) plays a crucial role in performance. For instance, the accuracy significantly drops in tasks such as Gene\_location when this compo-nent is removed. The general search engines (D2) and local PubMed database (D3) also contribute positively, but their impact is less pronounced compared to the Gene database. (2) **Component Contributions:** Among the components, the Self-Evaluation mechanism (C3) is vital for maintaining high accuracy across most tasks. The MeSH Filter (C1) and Query Rewrite (C2) also enhance perfor-mance, but their absence does not degrade the re-sults as severely as the removal of Self-Evaluation. (3) **Effects of Basic Language Models:** Compar-ing the two base models, Llama-3-70B (M2) gener-ally outperforms Llama-3-8B (M1) across all tasks, indicating that the larger model size contributes to better handling of complex biological queries. These findings underscore the importance of inte-grating diverse data sources and advanced compo-

To compare reasoning differences among BIORAG and the baselines in a more intuitive manner, we select three typical case studies in this section. We first provide a case study to show the work-flow of BIORAG (Figure 5). It is selected from the College Biology dataset. BIORAG performs self-evaluation twice: the first time it starts with a web search for general information, but the results are insufficient to support answering the question. Thus BIORAG conducts the second self-evaluation and calls for the more specialized PubMed database. The results this time are accu-rate and sufficient to support answering the ques-tion, thus BIORAG gives the final answer based on the results.

The second case study is conducted on the gene alias task in the GeneTuring dataset (Figure 6). The challenge of this task is the variants of gene names. NewBing gets the response from the Wikimedia. However, Wikimedia is not specialized enough to provide the alias for the input gene, which leads to the wrong answer. The prompts of GeneGPT are too complicated, none of the prompts is relevant to this task. In addition, its NCBI API returns the gene IDs, instead of the gene names. The LLM is un-able to understand these IDs, and finally arrives at a wrong answer. BIORAG employs fuzzy queries, yielding a larger number of related responses with a higher error tolerance. Furthermore, each result contains detailed gene-related information and de-scriptions, such as the aliases. Thus BIORAG gets the correct answer.

The third case study is conducted on the gene-disease association task in the GeneTuring dataset, shown in Figure 7. Reasoning behind this task re-lies on both the Gene database and relative PubMed papers. The PubMed abstracts provide detailed gene-disease relationships. NewBing gets the re-sponse from the Geekymedics website. Although the Geekymedics website provides general medical information, it does not offer the correct or spe-cific details required for gene-disease associations. Consequently, NewBing’s response is inaccurate due to the reliance on a non-specialized source. GeneGPT chose the wrong NCBI API. The API’s

feedback is a complicated and interminable HTML page, with massive irrelevant information or de-scriptions. Based on the ambiguous backgrounds, GeneGPT outputs the wrong answer. In the reason-ing process of BIORAG, BioRAG uses multiple tools, i.e., Gene database, local PubMed database, and Web search, to gather and conduct mutual con-firmation on the information of genes associated with B-cell immunodeficiency. The process in-volves preprocessing queries, executing searches, and conducting self-evaluations at each step to en-sure comprehensive and accurate results. The rea-soning process is thorough, incorporating various data sources to confirm the association of specific genes with B-cell immunodeficiency.

**4 Conclusion**

This paper introduces BIORAG, an innovative framework that integrates Retrieval-Augmented Generation with Large Language Models to en-hance biological question-reasoning. The frame-work’s ability to obtain relevant and current in-formation from a blend of traditional databases, toolkits, and modern search engines ensures the accuracy of the generated answers. Through ex-tensive validation, including rigorous testing on widely recognized biology QA datasets and exten-sive case studies, BIORAG has demonstrated its su-perior ability to handle complex biological queries. These results underscore the framework’s potential as a valuable tool for the scientific community, fa-cilitating more accurate and efficient information processing.

**References**

Peter William Atkins, George Ratcliffe, Julio de Paula, and Mark Wormald. 2023. *Physical chemistry for the life sciences*. Oxford University Press.

Sören Auer, Dante AC Barone, Cassiano Bartz, Ed-uardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Letícia MF Bertoline, Angélica N Lima, Jose E Krieger, and Samantha K Teixeira. 2023. Before and after alphafold2: An overview of protein structure predic-tion. *Frontiers in Bioinformatics*, 3:1120370.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models un- derstand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Benjamin Cole, Dominique Bergmann, Crysten E Blaby-Haas, Ian K Blaby, Kristofer E Bouchard, Siobhan M Brady, Doina Ciobanu, Devin Coleman-Derr, Samuel Leiboff, Jenny C Mortimer, et al. 2021. Plant single-cell solutions for energy and the environ-ment. *Communications biology*, 4(1):962.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific lan-guage model pretraining for biomedical natural lan-guage processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. 2023. Prompt-guided retrieval augmen-tation for non-knowledge-intensive tasks. In *Find-ings of the Association for Computational Linguistics: ACL 2023*, pages 10896–10912.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language under-standing. *arXiv preprint arXiv:2009.03300*.

Andreas Holzinger, Katharina Keiblinger, Petr Holub, Kurt Zatloukal, and Heimo Müller. 2023. Ai for life: Trends in artificial intelligence for biotechnol-ogy. *New Biotechnology*, 74:16–24.

Wenpin Hou and Zhicheng Ji. 2023. Geneturing tests gpt models in genomics. *BioRxiv*.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open- source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Augustin Lecler, Loïc Duron, and Philippe Soyer. 2023. Revolutionizing radiology with gpt-based models: current applications, future possibilities and limita-tions of chatgpt. *Diagnostic and Interventional Imag-ing*, 104(6):269–274.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Dominique Lepore, Koustabh Dolui, Oleksandr Tomashchuk, Heereen Shim, Chetanya Puri, Yuan Li, Nuoya Chen, and Francesca Spigarelli. 2023. Inter- disciplinary research unlocking innovative solutions in healthcare. *Technovation*, 120:102511.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Nau-mann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural In-formation Processing Systems*, 36.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Jun-jie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback. *arXiv preprint arXiv:2403.06840*.

Qingqing Long, Yilun Jin, Yi Wu, and Guojie Song. 2021a. Theoretically improving graph neural net-works via anonymous walk graph kernels. In *Pro-ceedings of the Web Conference 2021*, pages 1204–1214.

Qingqing Long, Lingjun Xu, Zheng Fang, and Guojie Song. 2021b. Hgk-gnn: Heterogeneous graph ker-nel based graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1129–1138.

Zooey Nguyen, Anthony Annunziata, Vinh Luong, Sang Dinh, Quynh Le, Anh Hai Ha, Chanh Le, Hong An Phan, Shruti Raghavan, and Christopher Nguyen. 2024. Enhancing q&a with domain-specific fine-tuning and iterative reasoning: A comparative study. *arXiv preprint arXiv:2404.11792*.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Bob O’Donnell. 2023. New bing brings ai to search engine. *USA Today*, pages 01B–01B.

Ankit Pal, Logesh Kumar Umapathi, and Malaikan-nan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical do-main question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Kho-vanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O’Neill, Barbara Robbertse, et al. 2020. Ncbi taxon-omy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062.

Guojie Song, Qingqing Long, Yi Luo, Yiming Wang, and Yilun Jin. 2020. Deep convolutional neural net-work based medical concept normalization. *IEEE Transactions on Big Data*, 8(5):1195–1208.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmen- tation for large language models. *arXiv preprint*  *arXiv:2310.05002*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Infor-matics Association*, page ocae045.

Jasper Xian, Tommaso Teofili, Ronak Pradeep, and Jimmy Lin. 2024. Vector search with openai em-beddings: Lucene is all you need. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1090–1093.

Meng Xiao, Ziyue Qiao, Yanjie Fu, Hao Dong, Yi Du, Pengyang Wang, Hui Xiong, and Yuanchun Zhou. 2023. Hierarchical interdisciplinary topic detection model for research proposal classification. *IEEE Transactions on Knowledge and Data Engineering*.

Meng Xiao, Min Wu, Ziyue Qiao, Yanjie Fu, Zhiyuan Ning, Yi Du, and Yuanchun Zhou. Interdisciplinary fairness in imbalanced research proposal topic in-ference: A hierarchical transformer-based method with selective interpolation. *ACM Transactions on Knowledge Discovery from Data*.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Ra- jmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. *arXiv preprint*  *arXiv:2305.11541*.