A Survey on Biomedical RAG: Recent Advances and New Frontiers

This repository provides a comprehensive survey on the application of Retrieval Augmented Generation (RAG) in the medical and healthcare domains. We present the basic framework of medical RAG, detailing its commonly used components, datasets, and evaluation methods. Additionally, we've compiled a collection of state-of-the-art (SOTA) approaches and highlighted some literature that explores new frontiers in this field. We are committed to regularly updating this repository and welcome any feedback or suggestions.

Introduction

With the emergence of large language models (LLMs) in recent years, numerous natural language processing (NLP) tasks have seen remarkable advancements. Their impressive capabilities in generating and understanding human-like text have resulted in outstanding performance in tasks such as summarization, question answering, information retrieval, and more. The exceptional performance of LLMs in core NLP tasks is prompting their exploration in the medical domain, ranging from aiding clinicians in making more accurate decisions to enhancing patient care quality and clinical outcomes.

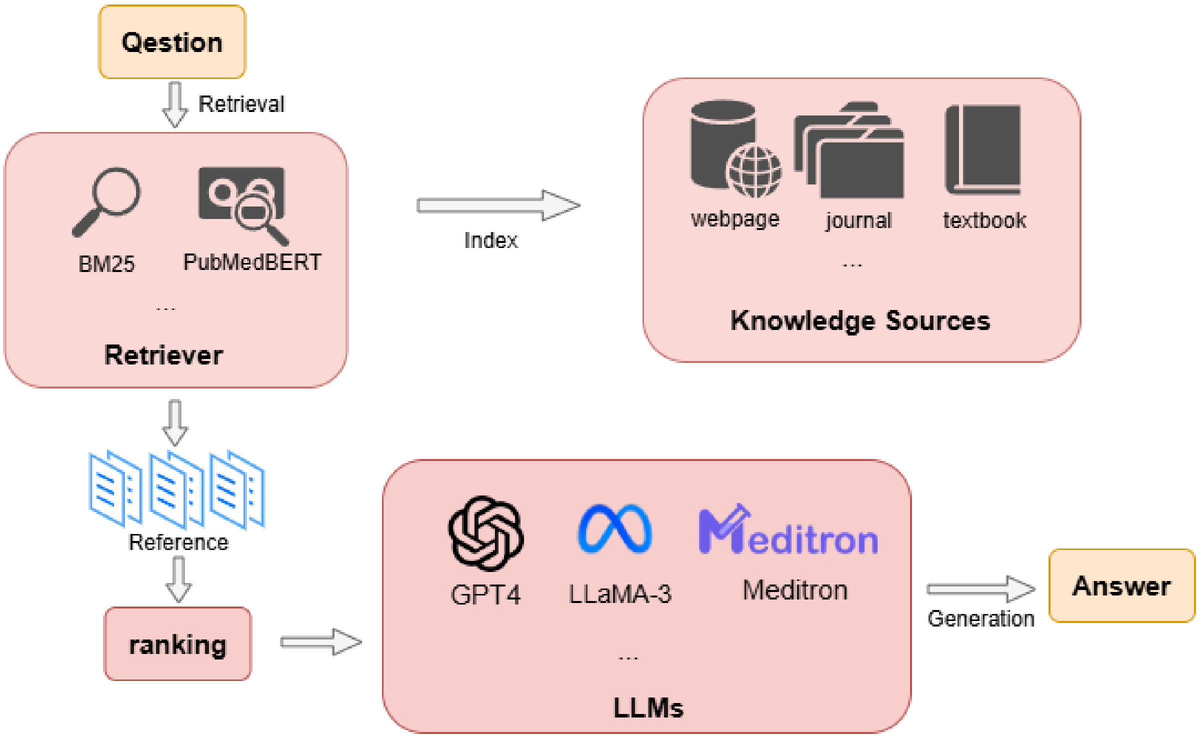
However, LLMs often generate plausible-sounding but factually incorrect responses, a phenomenon commonly known as hallucination. Additionally, once the training process is complete, the parameters of LLMs are fixed, resulting in a lack of up-to-date knowledge. Retrieval Augmented Generation (RAG) has the potential to alleviate these critical challenges because it can provide the rationale behind its generation and readily access the latest knowledge.

This survey focuses on useful techniques and the latest advances in medical and healthcare RAG. We first illustrate its basic framework and important components, and then we detail some useful improvements to these components separately. Next, we introduce datasets commonly used to evaluate medical and healthcare RAG, along with widely used knowledge sources. Finally, we present some evaluation metrics commonly used in experiments and explore new frontiers in this field.

(Please note that these new frontiers are constantly evolving. We strive to stay updated with the latest work and welcome any suggestions.)

Basic framework

Here we present a basic framework of the vanilla medical RAG. As shown in the following figure, there are four key components in medical RAG: the retriever, knowledge source, ranking method, and large language model (LLM). A question is first processed by the retriever, which indexes some relevant documents from a variety of knowledge sources composed of webpages, academic papers, textbooks and so forth. After retrieval, we obtain references, also referred to as context or background knowledge in some literature. RAG uses ranking methods to sort these references based on their relevance to the original question. Finally, the top-k relented references, along with the original question, are sent to the LLM as input to generate the final result.



Basic Framework of Medical RAG

Retriever

The retriever is a key component to decide the relevance of references to the question. A good retriever can identify the most relevant and useful documents to answer the question, while a poor one may fail to be helpful and introduce noisy information. Here we divide these retrievers into following 3 different types.

Lexical Retriever

BM25 [pdf] is a ranking function used in information retrieval to estimate the relevance of documents to a given search query. It is commonly treated as a baseline for comparison with other retrievers. However, in many tasks, experimental results demonstrate that it still offers competitive performance.

Search Engine Retriever

Using a search engine provides access to a wide range of external knowledge sources, making the search engine retriever a promising component in RAG (Retrieval-Augmented Generation). Below, we list some tools that are commonly used as retrievers in medical RAG, along with relevant literature that utilizes these tools.

NCBI Tool

> The National Center for Biotechnology Information (NCBI) provides many useful products, including PubMed, PubMed Central, PubChem, Gene, and Genome. In addition to the web interfaces to these products, NCBI also provides an API allowing programmatic access to the underlying databases and search technology.

Entrez API, also known as Entrez Programming Utilities (E-utilities), is a set of web-based tools provided by the National Center for Biotechnology Information (NCBI). These tools allow researchers and developers to access and retrieve data from NCBI's comprehensive suite of biological databases programmatically.

PubMed API provides access to the PubMed database when you specify the database as "PubMed" in your search query. Note that the PubMed API is part of the Entrez API system. You can also specify other databases, such as PubMed Central or Gene, in your search queries.

Wikipedia Tool

Wikipedia API is a set of application programming interfaces (APIs) that allows developers to access and interact with Wikipedia's vast content programmatically.

Question2Query

Sometimes, an LLM (Large Language Model) is used to transform a user's question or dialogue history into a search engine query, which is then executed in the search engine database. We refer to this method as 'Question2Query.' This approach is often used in combination with a search engine retriever.

Literature

- An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. [pdf]
  using Entrez API as retriever, Question2Query
  https://github.com/som-shahlab/Clinfo.AI/tree/main
- Tool calling: Enhancing medication consultation via retrieval-augmented large language models.[pdf]
  Distilling the key information and forming the searching query (Question2Query), using search engine as retriever

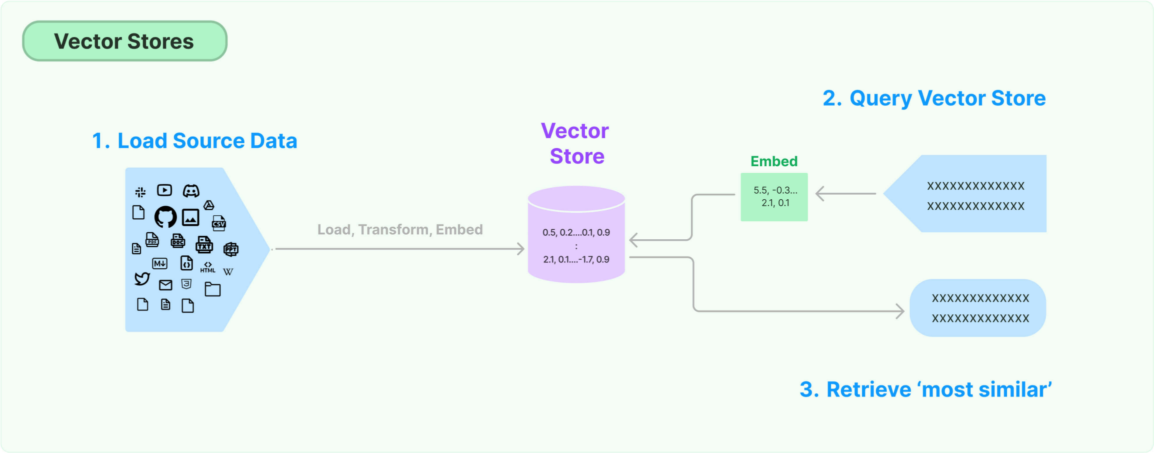
Semantic Retriever

Due to recent advancements in deep learning, semantic retrievers, also known as dense retrievers, have achieved impressive performance and are widely used in Biomedical RAG. These retrievers encode and match queries and documents as dense vectors (document embeddings). This approach often utilizes Pre-trained Language Models (PLMs) to encode documents, treating the nearest documents in vector space as the most relevant at a semantic level.

Vector Stores

> One of the most common ways to store and search over unstructured data is to embed it and store the resulting embedding vectors, and then at query time to embed the unstructured query and retrieve the embedding vectors that are 'most similar' to the embedded query. A vector store takes care of storing embedded data and performing vector search for you.

Vector stores are an important component in semantic retrievers, offering efficient search methods like K-nearest neighbors (KNN) for RAG developers. They enable rapid retrieval of semantically similar documents, enhancing the performance of Biomedical RAG systems. The process of embedding, storing, and searching documents is illustrated in the following picture. Here, we list two commonly used vector stores in Biomedical RAG.



> **Chroma** is an AI-native open-source vector database. It comes with everything you need to get started built in, and runs on your machine.

> **Faiss** is a library for efficient similarity search and clustering of dense vectors. It contains algorithms that search in sets of vectors of any size, up to ones that possibly do not fit in RAM. It also contains supporting code for evaluation and parameter tuning.

Embedding-based Retriever

Also called dense retriever. We can construct a semantic retriever by combining an embedding method with a vector store. However, it is crucial to select an appropriate embedding model, as the training corpora of these models vary, leading to differing abilities to encode various types of documents.

General Embedding Models

General embedding models are trained on general corpora and are widely used in various information retrieval systems. There are a substantial number of open-source general embedding models, and they are often treated as baselines in Biomedical RAG experiments. The following table shows some representative models.

Model	Feature	Data	Link
LDA	Machine Learning based	2003	pdf; Github
Doc2Vec	Deep Learning based	2014	pdf; Github
FastText	Deep Learning based	2017	pdf; Github
Sent2Vec	Deep Learning based	2018	pdf; Github
RoBERTa	BERT based	2019	pdf; Hugging Face
ColBERT	BERT based	2020	pdf; Github; Hugging Face
SimCSE	Contrastive Learning	2021	pdf; Github

Commercial Embedding Models

Thanks to recent advances in Large Language Models, many AI companies now provide commercial embedding APIs, which are popular among biomedical researchers and developers. Although these services may be costly, especially with large datasets, their attractive performance and convenience (call API only, not need for train) has led to widespread use. The following table lists some popular commercial embedding models. Note that each company offers models of various sizes, so the maximum input and embedding dimensions may vary.

Model	Max Input Token	Dimension	Company	Link
text-embedding	8191	1536-3072	OPEN AI	document
voyage-2	4000~1600	1024-1536	Anthropic	document
Vertex AI	3072	768	Google	document
bge-large	512	1024	Baidu	document
tao-8k	8192	1024	Baidu	document

Embedding with open source LLMs

Due to their larger number of parameters, large language models have a superior ability to understand text. Some researchers choose to use the embedding layers of open-source LLMs for document embedding. There are many LLMs available for embedding purposes. More details about LLMs in RAG can be found in Generation Model section. Here, we provide an example that uses MedLLaMA as an embedding model.

Biomedical Embedding Models

In many biomedical NLP tasks, language models trained on biomedical-related corpora outperform general-domain language models due to their superior ability to understand biomedical language. Therefore, using a biomedical language model as an embedding model is a common approach to building a semantic retriever. Below, we list some popular biomedical embedding models.

Model	Base	Date	Link
BioBERT	BERT	2019	pdf; Github
PubMedBERT	BERT	2021	pdf; Hugging Face
UmlsBERT	BERT	2021	pdf; Github
BioBART	BART	2022	pdf; Github

More information about biomedical embedding models can be found in Table I of this survey.

Recent Advanced Retriever

Although the aforementioned off-the-shelf retrievers are readily available, the task of searching for relevant and accurate documents in the medical domain remains challenging. Consequently, customized retrievers tailored specifically to these tasks have been developed. These advanced retrievers also achieve comparable performance. Here, we introduce some of them.

- **Contriever**: An unsupervised general dense retriever based on contrastive learning. Experiments show that its zero-shot capability (trained on general corpora and applied to new domains) is promising. It is often used as a strong baseline in biomedical RAG.
  [pdf]; [Github]
- **SPECTER**: A model designed to generate document-level embeddings of scientific documents. It utilizes the citation graph as a training signal to capture inter-document relatedness, showing an advantage in generating document embeddings, particularly for scientific papers. In the biomedical domain, scientific papers are a vital part of external knowledge sources, so using SPECTER as a retriever in biomedical RAG demonstrates comparable performance.
  [pdf]; [Github]
- **MedCPT**: A specialized retriever trained on user click logs from PubMed. It includes a query encoder (QEnc), a document encoder (DEnc), and a ranking model (CrossEnc). The initial embedding model used in MedCPT is PubMedBERT. Contrastive loss is employed to train both the MedCPT retriever (QEnc, DEnc) and the MedCPT ret-ranker (CrossEnc).
  [pdf]; [Github]
- **MEDRAG toolkit**: A systematic implementation of RAG for medical QA. It provides a convenient way to search multi-source medical documents. Its retriever consists of four different components, including BM25, Contriever, SPECTER, and MedCPT.
  [pdf]; [Github]
- **Llama2Vec**: An LLM-based embedding method that fine-tunes Llama on novel unsupervised adaptation tasks, enabling it to serve as an effective backbone encoder for dense retrieval. Although originally developed for the general domain, it can be applied to or further enhanced for biomedical RAG.
  [pdf];

Literature

- BiomedRAG: A Retrieval augmented Large Language Model for Biomedicine [pdf]
  Embedding with open source LLMs, using MedLLaMA
- Zero-shot ecg diagnosis with large language models and retrieval-augmented generation [pdf]
  Commercial Embedding Models, using OPEN AI API
- Chatent: Augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery [pdf]
  Commercial Embedding Models, using OPEN AI API
- Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented [pdf]
  Recent Advanced Retriever, using MedCPT
- ReMAG-KR: Retrieval and Medically Assisted Generation with KnowledgeReduction for Medical Question Answering [pdf]
  Recent Advanced Retriever, using MedCPT

Ranking Method

Actually, most of the aforementioned retrievers , e.g., BM25, can provide a score for each retrieved document, allowing us to rank these documents based on their retrieval scores. However, for tasks in the biomedical domain, ranking may involve additional considerations. For instance, in the 2021 TREC Healthcare Misinformation Track, a retrieved document is assessed based on three criteria: *Usefulness*, *Supportiveness*, and *Credibility*.

> *Usefulness*: The extent to which the document contains information that a search user would find useful in answering the topic's question.
> *Supportiveness*: Does the document support or dissuade the use of the treatment in the topic's question
> *Credibility*: whether the document is considered credible by the assessor.

A good ranking method can accurately identify the rationale behind a user's query while filtering out noise and toxic information. Here, we introduce some practices.

Reciprocal Rank Fusion

Reciprocal Rank Fusion (RRF) [pdf] is an algorithm that evaluates search scores from multiple, previously ranked result sets to produce a unified result set. RRF can be utilized when using multiple retrievers, and you want to combine their results into a single ranking. This approach gives higher weight to documents that are ranked highly in at least one list, even if they aren't ranked as highly in others.

Trained Scorer & Distilled Scorer from LLM

Some choose to train a neural-network-based scorer model to rank retrieval results according to specific criteria. This approach requires an additional dataset to train a model that can evaluate documents based on different criteria. We introduce some datasets used to train a ranking scorer in the Dataset section.

Additionally, since large models like ChatGPT have strong evaluation capabilities, some developers opt to use large language models (LLMs) as scorers to filter relevant documents. However, because LLMs can be costly to run, distilling a model from an LLM can also be a good option.

Literature

- Benchmarking retrieval augmented generation for medicine [pdf]
  the aforementioned MEDRAG toolkit, using RRF to incorporate its four retrievers
- Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation [pdf]
  trained scorer, utilizing BERT as the backbone model, using three additional datasets to evaluate documents based on *Usefulness*, *Supportiveness*, and *Credibility*, RRF is then used to obtain the final score.
- BiomedRAG：A retrieval augmented large language model for biomedicine [pdf]
  distilled scorer from LLM, its Tailored Chunk Scorer is trained to align with the LLM.

Generation Model

Generation, based on the references found by the retriever and the prompt, is a key step in providing the final response to the user. Large language models (LLMs) are often used as the generation component. There are a substantial number of LLMs available for biomedical RAG. We categorize these LLMs into three different types and list some widely used LLMs for each.

General domain open source LLMs

Model	Param Size	Link
T5	0.06-11B	[Github]; [Checkpoints]
ChatGLM3	6B	[Github]; [Checkpoints]
OpenLLaMA	3,7,13B	[Github]; [Checkpoints]
LLaMA 2	7-70B	[Download Link]
LLaMA 3	8-70B	[Introduction]; [Checkpoints]
MPT-7B	7B	[Introduction]; [Checkpoints]
Phi-3 Mini	3.8B	[pdf]; [Checkpoints]
Mistral 7B	7B	[Introduction]; [Checkpoints]

General domain commercial LLMs

Model	Company	Link
ChatGPT-3.5/4	Open AI	document
Claude-3.5	Anthropic	document
Gemini-2.0	Google	document
ERNIE-4	Baidu	document

Biomedical domain open source LLMs

Model	Base	Param Size	Link
ChatDoctor	LLaMA	7B	[Github]
MedAlpaca	LLaMA	7, 13B	[Github]
PMC-LLaMA	LLaMA	7B	[Github]
GatorTronGPT	GPT-3	5, 20B	[Github]
BioMistral	Mistral	7B	[Hugging Face]
MEDITRON	LLaMA	7,70B	[Webpage]; [Hugging Face]

As answering user questions primarily relies on the generative capabilities of LLMs, we only list LLMs that are good at generation and can serve as backbone models for biomedical RAG.

Knowledge Source

Data quality of knowledge sources is crucial for large language models (LLMs) to effectively answer medical questions. Here, we list some valuable knowledge sources, including medical research articles, clinical guidelines, textbooks, drug databases, and knowledge graphs. Practice shows that increasing the number of knowledge sources does not always lead to improvement. Therefore, we should select appropriate knowledge sources based on the specific problem at hand.

- **PubMed** is the most widely used literature resource, containing over 36 million biomedical research articles. Many relevant studies solely use PubMed as the retrieval corpus.
  link: https://pubmed.ncbi.nlm.nih.gov/
- **StatPearls** is a point-of-care clinical decision support tool. There are 9,330 publicly available StatPearls articles accessible through NCBI Bookshelf.
  link: https://www.ncbi.nlm.nih.gov/books/NBK430685/.
- **Textbooks** is a collection of 18 widely used medical text books, which are important references for students taking the United States Medical Licensing Examination (USLME).
  link: https://github.com/jind11/MedQA?tab=readme-ov-file
- **DrugBank** is a comprehensive and freely accessible online database containing information on drugs and drug targets. It integrates detailed drug data (chemical, pharmacological, and pharmaceutical) with comprehensive information on drug targets (sequence, structure, and pathway).
  link: https://go.drugbank.com/releases/latest
- **Medical Subject Headings (MeSH)** is a comprehensive controlled collection for indexing journal articles and books in the life sciences. It organizes information on biomedical and health-related topics into a hierarchical structure.
  link: https://www.nlm.nih.gov/databases/download/mesh.html
- **PrimeKG** offers a comprehensive overview of diseases, medications, side effects, and proteins by merging 20 biomedical sources to detail 17,080 diseases across ten biological levels.
  link: https://github.com/mims-harvard/PrimeKG
  Xu et al. provide a template for converting a knowledge graph (KG) into sentences as follows:

  ```
  candidate_relation = ["disease_phenotype_positive", "disease_protein",
  "disease_disease", "drug_effect", "drug_protein"]

  relations = {
  "phenotype present": "[ent1] has the phenotype [ent2]",
  "carrier": "[ent1] interacts with the carrier [ent2]",
  "enzyme": "[ent1] interacts with the enzyme [ent2]",
  "target": "The target of [ent1] is [ent2]",
  "transporter": "[ent2] transports [ent1]",
  "associated with": "[ent2] is associated with [ent1]",
  "parent-child": "[ent2] is a subclass of [ent1]",
  "side effect": "[ent1] has the side effect of [ent2]"
  }
  ```

- **Wikipedia** is a large-scale open-source encyclopedia, which frequently used as a general domain corpus in information retrieval tasks.
  link: https://huggingface.co/datasets/legacy-datasets/wikipedia

Among these knowledge sources, Xiong et al. provide a biomedical knowledge corpus called **MedCorp**, which combines documents from PubMed, StatPearls, textbooks, and Wikipedia. You can find more information about this corpus at the following link: https://huggingface.co/MedRAG.

Dataset

In this section, we introduce some datasets that can be used to evaluate the performance of biomedical retrieval-augmented generation (RAG) systems. As mentioned before, RAG consists of a retriever, a ranking method, and a generation model. For comprehensive evaluation, we have collected two types of datasets: those for biomedical information retrieval (IR) tasks and those for biomedical downstream tasks. The former focuses on evaluating retrievers and ranking methods, while the latter, which includes tasks such as information extraction, question answering (QA), multiple-choice evaluation, dialogue, and text summarization, assesses the performance of the entire biomedical RAG system on biomedical tasks.

Because our primary focus is on evaluating biomedical RAG systems, we have not included datasets for training or fine-tuning large language models (LLMs) here.

Biomedical Information Retrieval (IR) Tasks

Dataset	Year	Link
TREC-COVID	2021	[pdf]; [Wegpage]
BioASQ	2015	[pdf]; [Wegpage]
SciFact	2020	[pdf]; [Hugging Face]
NFCorpus	2016	[pdf]; [Wegpage]
SciDocs	2020	[pdf]; [Github]

Additionally, some researchers invite professional physicians to evaluate the relevance between retrieval results and the questions, although this approach can be somewhat costly. An alternative method is to use powerful commercial LLMs like ChatGPT or open-source medical LLMs like MEDITRON to assess this relevance. If using this approach, it is important to design an effective prompt.

Biomedical Downstream Tasks

Medical Information Extraction

Dataset	Year	Task	Link
GENIA	2003	Entity Recognition	[pdf]; [Hugging Face]
GENIA11	2011	Event Extraction	[pdf]; [Webpage]
ADE	2012	Relationship Extraction	[pdf]; [WebPage]

Dataset	Year	Task	Link
ShARe13	2013	Entity Recognition	[pdf]
GENIA13	2013	Event Extraction	[pdf];[Webpage]
NCBI	2014	Entity Recognition	[pdf]; [Wegpage]
ShARe14	2014	Entity Recognition	[pdf]; [Webpage]
CADEC	2015	Entity Recognition	[pdf]; [Webpage]
BC5CDR	2016	Entity Recognition	[pdf]; [Hugging Face]
PHEE	2022	Event Extraction	[pdf]; [Github]

Medical Question Answering

Dataset	Year	Task	Link
MedQA	2021	Multiple-choice	[pdf]; [Github]
MedMCQA	2022	Multiple-choice	[pdf]; [Github]
PubMedQA	2019	Multiple-choice	[pdf]; [Github]
MMLU	2021	Multiple-choice	[pdf]; [Github]
MedicationQA	2019	QA	[pdf]; [Github]
MedQuAD	2019	QA	[pdf]; [Github]
HealthSearchQA	2022	QA	[pdf]
emrQA	2018	QA	[pdf]; [Github]
MEDIQA	2020	Dialogue	[pdf]; [Hugging Face]
CORD-19	2020	Dialogue	[pdf]; [Hugging Face]
ChatDoctor	2023	Dialogue	[pdf]; [Github]
Wikidoc Patient Information	2023	Dialogue	[Hugging Face]
Medical Flashcards	2023	Dialogue	[Github]
Wikidoc	2023	Dialogue	[Hugging Face]
MIRAGE	2024	Multiple-choice & QA	[pdf]; [Github]

Medical Generation

Dataset	Year	Task	Link
MIMIC-III	2016	Text Summarization	[pdf]; [Webpage]
MIMIC-CXR	2019	Text Summarization	[pdf]; [Webpage]

Dataset	Year	Task	Link
MeQSum	2019	Text Summarization	[pdf]; [Github]
CORD-19	2020	Text Summarization	[pdf]; [Github]
MentSum	2022	Text Summarization	[pdf]; [Webpage]
MultiCochrane	2023	Text Summarization	[pdf]; [Github]
PMC	Update	Text Summarization	[Webpage]

Evaluation Method

Here, we list some evaluation metrics used to assess RAG in different tasks. According to the task, we divide these metrics into supervised metrics and unsupervised metrics.

Supervised Metrics

Task	Metric
Text Classification	Accuracy, Recall, Precision, F1 Score, ROC Curve, AUC Value
Named Entity Recognition	Precision, Recall, F1 Score, Span-Level Metrics, Entity Type Level Metrics
Question Answering Systems	EM (Exact Match), F1 Score, Mean Reciprocal Rank (MRR), Hits@k
Information Retrieval	NDCG@k, MAP@k

Unsupervised Metrics

For unsupervised tasks like text generation and text summarization, traditional methods often use lexical-level metrics such as **BLEU**, **ROUGE**, **METEOR**, **GoogleBLEU**, and **chrF**. These metrics primarily assess the quality of results by checking vocabulary matches between candidate and reference texts. However, **BERTScore** has become increasingly popular because it evaluates the semantic relevance between the candidate document and the reference from a semantic perspective. This means BERTScore leverages the semantic representation capabilities of deep learning models like BERT to provide a deeper understanding of the semantic similarity in the texts, thus offering more comprehensive evaluation results.

Source-Augmented Metrics

Source-Augmented (SA) metrics additionally consider relevant context to evaluate the quality of model-generated outputs. These metrics go beyond comparing the generated text to reference outputs by considering how well the generated content aligns with the source content in terms of meaning, context, and information. Here we list some SA metrics.

- **UniEval (T5-large)**: UniEval uses the T5-large model, a powerful text-to-text transformer, to assess various dimensions of generated text, including fluency, coherence, and informativeness. Its capacity to evaluate multiple facets makes it effective for comprehensive analyses of text outputs.

- **COMET (XLM-RoBERTa)**: COMET leverages the XLM-RoBERTa model, which is a multilingual pretrained transformer, to improve translation task evaluations. By capturing semantic nuances across various languages, COMET offers more reliable assessments of cross-lingual semantic similarity and translation quality.
- **CTC Summary Consistency (BERT)**: This metric uses BERT to evaluate the consistency of summaries, ensuring they maintain semantic alignment and coherence with the source text. It provides a measure of semantic fidelity in text summarization, enhancing the evaluation beyond basic lexical consistency.

Frontiers

In this field, several promising methods have recently emerged, aiming to explore effective patterns for constructing a more advanced biomedical RAG. We present several noteworthy studies in this section and will continue to update it with the latest advancements in the field.

Jeong et al. [pdf] introduced a self-adaptive RAG framework tailored for the biomedical domain, termed Self-BioRAG, which empowers the LLM to autonomously determine whether a given question necessitates retrieval.If a question does not require retrieval, the backbone LLM will address it solely through its generative capabilities. Wang et al. [pdf] introduced JMLR, a synchronized training framework designed to jointly optimize the fine-tuning of both the LLM and the information retrieval component. It offers a novel approach to align the retriever and the LLM in Biomedical RAG. Frisoni et al. [pdf] present MEDGENIE, the first generate-then-read framework, which demonstrates superior performance compared to traditional retrieval-then-read approaches. Their work highlights that suboptimal retrieval can introduce noise and undermine the generative capabilities of medical LLMs.

Here, we present some of the latest works in biomedical RAG research and will continuously update this list.

- **To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering**
- **RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records(short)**
- **Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability**
- **Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented**
- **MedCoT: Medical Chain of Thought via Hierarchical Expert**
- **BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers**
- **SeRTS: Self-Rewarding Tree Search for Biomedical Retrieval-Augmented Generation**
- **BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine**