International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Design and Implementation of an Offline Multimodal Retrieval Augmented Generation System for Unified Semantic Search

Sandeep Kulkarni*, Aiman Patel**, Sandesh Singh***
*(Department of Computer Science, Ajeenkya D Y Patil University, Pune, Maharashtra
Email: facultyit528@adypu.edu.in)
** (Department of Computer Science, Ajeenkya D Y Patil University, Pune, Maharashtra
Email: umme.ahmed@adypu.edu.in)
*** (Department of Computer Science, Ajeenkya D Y Patil University, Pune, Maharashtra
Email: sandeshraj.singh@adypu.edu.in)

----------**********----------

Abstract

The growing amount of unstructured information contained in the heterogeneous forms of storage, including documents, pictures and audio recordings presents huge problems to the conventional retrieval systems that rely on keywords. Retrieval-Augmented Generation (RAG) enhances response grounding by combining retrieval processes with Large Language Models (LLMs); most of the existing systems are text-based and require cloud access. The article describes the design and implementation of a multimodal RAG framework system based offline to perform unified semantic search. The suggested system consumes PDF, DOCX, image, as well as audio files and transforms them into a common embedding space, and the similarity-based retrieval across the modalities. An offline LLM makes context sensitive responses that are purely based on the retrieved information, and their responses are provided with references to the sources to make it transparent. Experimental analysis proves the cross-modal retrieval, the lessening of hallucination and safe functioning without internet addiction. The architecture is especially adapted to privacy sensitive enterprise and government environments where explainable offline intelligence is needed.

Keywords — **Multimodal RAG, Offline Large Language Models, Semantic search, Vector Embeddings, Cross-Modal Retrieval, Explainable AI.**

----------**********----------

I. INTRODUCTION

The fast development of digital information has led to big amounts of unstructured information stored in a heterogeneous form of PDFs, Word documents, images, screenshots, and recorded calls [1,2]. It is also difficult to derive useful information with such data as the conventional information retrieval systems are based on the text-based and keyword-based methods of information processing which are inefficient to cross-format and semantically complicated queries [3].

Large Language Models (LLMs) have improved natural language understanding and conversational search [4], yet standalone models are associated with hallucination and do not have access to private domain data and usually rely on inference provided by clouds [2,4]. Retrieval-Augmented Generation (RAG) addresses these problems by accessing the surrounding context prior to generation which enhances factual grounding [5]. Nevertheless, the current RAG systems are mostly text-based, cloud-based, and multimodal integration limited [6,7].

Critical information is frequently spread across the documents, audio recordings, and images, which is why there is an increasing demand to have unified multimodal retrieval systems. Moreover, the privacy sensitive environment needs solutions that run purely offline [8,9].

In order to overcome these issues, the paper will propose an Offline Multimodal Retrieval

Augmented Generation system design and implementation of unified semantic search [10,11]. The architecture being proposed takes in various data formats and converts them into a common embedding space, does similarity-based retrieval, and responds over grounded by using an offline Large Language Model with explicit source indications to allow transparency and traceability [13].

The main contributions of this work are the following:

- Suggest a single, multimodal RAG architecture which can be used completely offline without external APIs.
- Further propose a common semantic embedding model of cross-format retrieval of text and audio data.
- To improve the explainability and user confidence, deploy a citation-aware generation mechanism.
- Show the implementation of the system in security sensitive environments.

This work will contribute to the creation of safe, explicable and multimodal intelligence mechanisms bridging the divide between the heterogeneous data storage and single semantic access.

II. LITERATURE REVIEW

A. Information Retrieval Systems Evolution

The IR systems have developed significantly. Initial approaches were based on lexical matching, like Boolean search, TF-IDF and probabilistic ranking approaches like BM25 [14]. They are good with structured text documents but cannot comprehend semantics and do not perform well with contextual queries and synonyms variations [3,10]. With the revolution of neural embedding models, semantic search had also come into existence, with textual representatives being densely represented in high-dimensional spaces [15]. Such vector databases like FAISS and Quadrant made it possible to perform similarity based retrieval on a cosine distance scale [12] so that systems no longer had to rely on exact matching of keywords but rather on meaning-based retrieval.

This paradigm was further developed by Retrieval-Augmented Generation (RAG) which integrates retrieval mechanisms with Large Language models [5]. The relevant context is also accessed and made available to the language model during generation in RAG systems, enhancing factual grounding and minimizing hallucination [16]. More recent studies have examined the multimodal learning paradigms that can align text, image and audio representations on common embedding spaces [7,14].

This evolutionary trend toward semantic search as opposed to traditional lexical search, RAG systems, multimodal learning, and finally offline multimodal RAG structures is shown in Figure 1, which represents the transition to the integrated, context-sensitive types of intelligence systems.

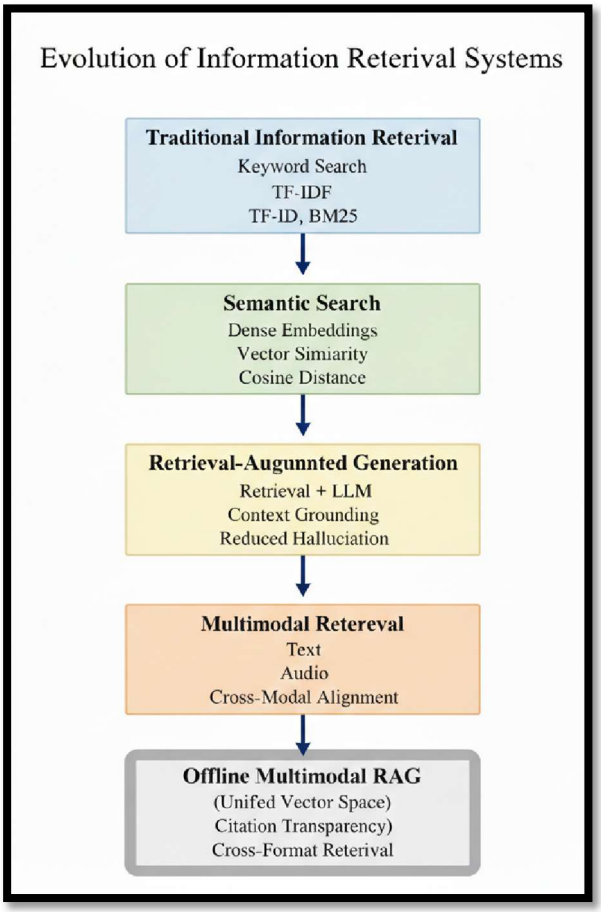

Figure 1. Evolution of Information Retrieval Systems

B. Retrieval-Augmented Generation and its Limitations.

RAG frameworks enhance the factual accuracy by basing responses on retrieved evidence as opposed to basing them only on the knowledge in a pretrained model [15]. Nevertheless, the vast majority of current implementations are text-based and cloud-based [16], and they presuppose the constant availability of the Internet and access to external APIs. This restricts their use in privacy sensitive or high security areas.

Moreover, the transparency of citation is not always constant. Although reference links are provided in some systems, they are not necessarily strict and do ground responses in retrieved context which can influence reliability and user trust.

C. Multimodality Learning and Cross-modal Retrieval.

Multimodal learning is concerned with matching different types of data, such as text, images and audio, in common spaces of representation [7]. Vision-language models have already been used in success in image captioning and visual question-answer type of tasks [14], and transformer-based Automatic Speech Recognition models allow converting audio recordings into searchable transcripts [15].

Despite these developments, the majority of multimodal systems are independent of the retrieval-augmented generation pipelines. There is little practical use in combining several modalities into one semantic retrieval framework. The traditional RAG systems are text-only (as shown in Figure 2) and do not allow cross-format retrieval, but multimodal RAG architectures can be used to retrieve information on both text, audio, and visual data, as all are visually represented in the same embedding space.

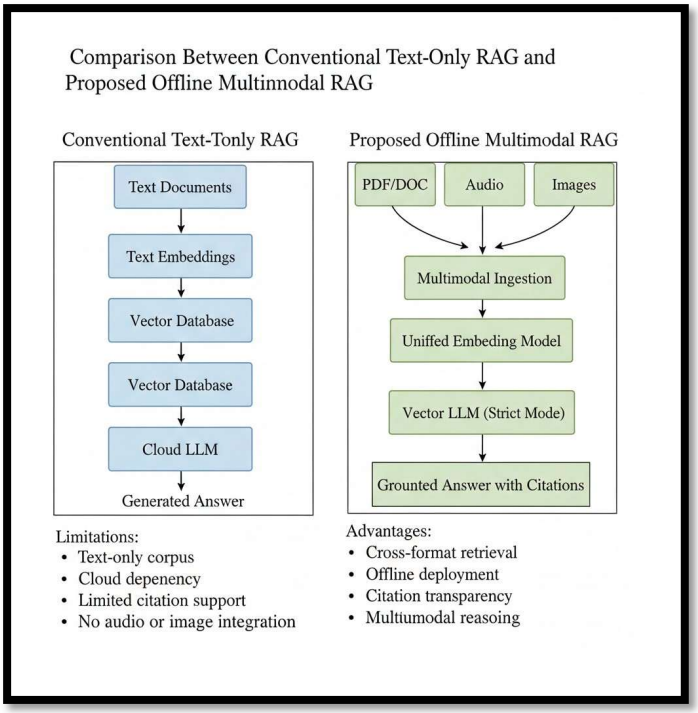

Figure 2. Comparison Between Text-Only RAG and Multimodal RAG Architectures

D. Unscaled Deployment and Privacy.

The recent developments of lightweight large language models have allowed offline inference to local settings [16]. The offline deployment is more vital to government and controlled industries where delicate information cannot be passed to the external cloud services [8,9]. Nevertheless, the current offline LLM applications are mostly standalone systems and do not include built-in multimodal retrieval pipelines or citation-conditional generation systems.

E. Research Gap and Motivation

There are a number of limitations identified in the literature. The conventional retrieval systems do not have semantic and multimodal features. Similarity search is supported by the use of vector databases which do not guarantee grounded response generation. Current RAG systems enhance factual reliability and are still mostly text-based and cloud-reliant. Despite the fact that multimodal learning facilitates cross-modal coordination, it is seldom incorporated into

integrated retrieval-augmented systems, which can run offline and with support of transparent citation.

As a result, there is still an obvious lack of unified offline multimodal Retrieval Augmentation Generation system, which can consume heterogeneous data, align it to a common semantic plane, and produce grounded responses complete with source attribution. To resolve this shortcoming, this paper introduces an Offline Multimodal Retrieval Augmented Generation architecture of unified semantic search, multimodal ingestion, shared embedding representation, vector-based retrieval and citation-aware answer generation in a
completely offline deployment setting.

III. METHODOLOGY

A. System Overview

The given system will be implemented as an offline multimodal Retrieval-Augmented Generation system, which involves the integration of heterogeneous information sources into a single systematic scheme of semantic retrieval. The system aids in consuming textual documents, audio records and images and converting all these modalities to a common vector representation so as to allow cross-format semantic search. There are five key steps in the methodology, namely: multimodal ingestion, embedding generation, vector storage, retrieval, and grounded response generation. Every component is set to be functional in a fully offline setting in order to guarantee privacy of data and operational autonomy.

B. System Architecture

The system has a layered architecture design in order to promote modularity, scalability, and explainability. Figure 3 is an illustration of the architecture.

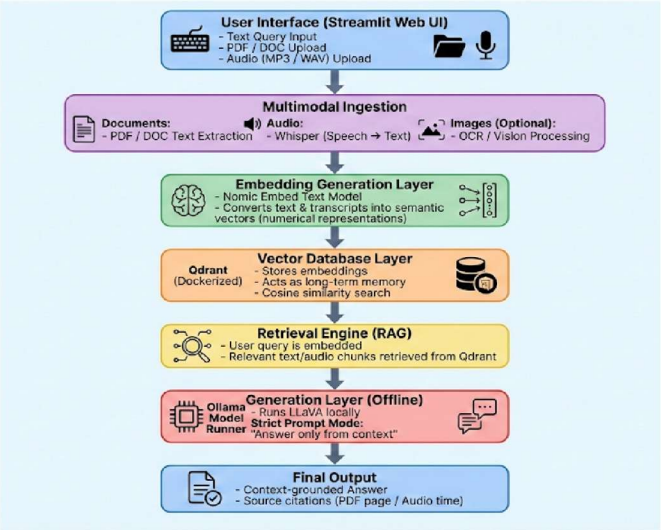

Figure 3. Offline Multimodal RAG System Architecture

These are the following layers in the architecture:

- User Interaction Layer
- Multimodal Ingestion Layer
- Embedding and Vectorization Layer.
- Retrieval Layer
- The Generation and Citation Layer.

The multimodal RAG pipeline has a clear role of each layer.

C. Multimodal Data Ingestion

The system allows consumption of a variety of data types to facilitate single semantic search. There are the following layers in the architecture:

1) ***Text Documents***: The PDF and DOCX documents are analyzed to retrieve pure text information. The resultant text is divided into smaller overlapping blocks in order to maintain coherence in context as well as enhancing retrieval granularity. Chunking makes it possible to retrieve relevant passages without necessarily processing documents.

2) ***Audio Recordings:*** Audio files in common formats, like MP3 and WAV are used through an automatic speech recognition model to produce textual transcripts. The transcripts are divided and saved with time stamp metadata to allow accurate access to audio snippets.

3) ***Images:*** Semantic description or text extraction techniques are used to process images and screenshots. Any textual content that was

extracted or generated captions are taken as recoverable semantic units.

Figure 4 represents the multimodal ingestion process.

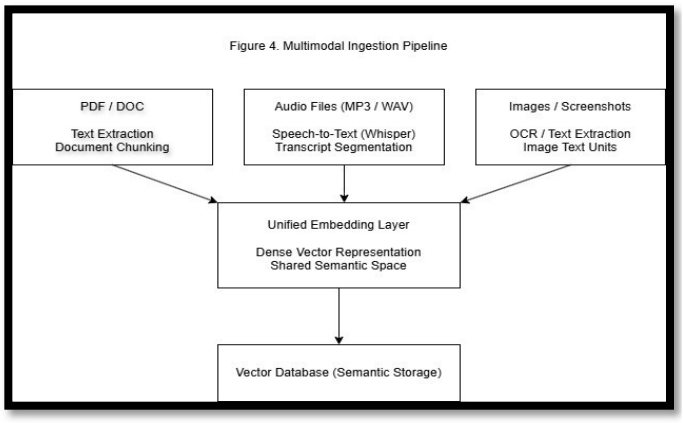

Figure 4. Multimodal Ingestion Pipeline

Such a design is to make sure that heterogeneous data is converted to a similar textual or semantic form before incorporating it.

D. Embedding and Unified Vector Space.

Once ingested, all textual fragments such as document fragments and audio transcripts, are converted into richer vectors. Semantic meaning as a high-dimensional numerical vector is an embedding. Similarity comparison- The similarity of objects, by mathematical distance measures, is possible through this transformation.

All modalities are projected into a common space of vectors in a unified embedding model. This standard representation is essential in facilitating cross-modal retrieval, e.g., relating a text query to a fragment of audio transcript. The similarity metric that will be used is cosine similarity because it is effective in estimating angular similarity among high-dimensional vectors. The mechanism of storing and retrieving vectors is known as Vector storage and Retrieval mechanism. All the created embeddings are placed in a vector database and provide efficient similarity search. The semantic memory of the system is represented by the vector database.

The steps followed when a user makes a natural language query are as follows:

- The query is transformed to an embedding.
- Similarity search is being done in the vector database.
- First K most relevant vectors are fetch out.
- Metadata like document source, page number or timestamp is identified.
- This retrieval-first methodology makes sure that the generation is based on the relevant evidence.

E. Retrieval-Augmented Generation.

The content recovered is put together in an organized prompt and fed into the offline Large Language Model. To limit the model to give an answer based on the retrieved context, a strict prompting strategy is employed.

Strict mode contains some instructions like:

- Response based on given context.
- External knowledge should not be used.
- Provide citations

This minimizes hallucination and grounds it to facts.

The LLM will then produce final answer on the basis of the retrieved information only.

F. Strategy of Offline Deployment.

The system is completely offline to provide data confidentiality and independence of operations. The local deployment of the vector database, embedding model, speech recognition system and the large language model is deployed in a safe environment.

Offline deployment removes the use of third-party API and that critical information is not taken beyond the organizational perimeter. This architecture gives the system an appropriate fit in security-sensitive usage in government, defence and controlled industry.

G. Citation and Explainability Mechanism.

Every chunk that may be retrieved is stored with metadata, of the form:

- Document name
- Page number
- Audio timestamp
- Image reference

In creating responses, the system provides numbered citation which connects back to the source of data. This is a transparency mechanism whereby generated content can be verified and this increases user trust.

The suggested methodology builds a single offline multimodal Retrieval-Augmented Generation system uniting document, audio, and image information in a common semantic embedding linkage. The system is able to provide correct, explainable and cross-format access to information through similarity retrieval using vectors and strict context-grounded generation. The offline deployment plan also ensures that the data is privately secure and that the architecture is also independently operable and so is common to secure settings. The next part is the presentation of the experimental setup and assessment of the suggested system.

IV. TEST RESEARCH AND FINDINGS

A. Evaluation Environment

The suggested offline multimodal RAG system was implemented into a controlled local setting to test its retrieval and production capacity. Every part, the embedding model, the vector database, the speech-to-text engine, and the large language model were all run without any external API. The assessment was centered on semantic recall of information, cross modal retrieval, grounding of responding and latency of the system.

B. Dataset and Query Design

A heterogeneous data set had been built to model real enterprise conditions. The dataset included:

- PDF and DOCX documents
- Audio tapes of structured and unstructured conversations.
- Graphic pictures and images with textual information.

Textual documents were divided into semantically coherent units. Audios were transcribed to time-stamped transcripts and images were run through text extraction, which allowed them to be indexed together in a common vector space.

The evaluation queries were created to assess a variety of scenarios of retrieval such as:

- Text-to-Text retrieval
- Text-to-Audio retrieval
- Text-to-Image retrieval
- Cross-Modal context-queries.

The queries focused on semantic meanings and not key word matching.

C. Evaluation Metrics

The system was assessed using the following metrics:

Retrieval Accuracy, defined as:

$$Retrieval\ Accuracy = \frac{Correct\ Retrievals}{Total\ Queries}$$

Cross-Modal Retrieval Rate, defined as:

$$Cross\text{-}Modal\ Success\ Rate = \frac{Successful\ Cross\text{-}Modal\ Retrievals}{Total\ Cross\text{-}Modal\ Queries}$$

Additional metrics included Response Grounding Score (correct citation support) and end-toend Latency.

A baseline comparison was conducted against a standalone offline LLM without retrieval support.

D. Retrieval Performance

TABLE I
Retrieval Accuracy Across Query Categories

Query Type	**Accuracy (%)**
Text-to-Text	91.2
Text-to-Audio	87.6
Text-to-Image	84.3
Cross-Modal Contextual	86.5
Overall Average	87.4

The results indicate strong semantic retrieval performance across modalities. Slightly lower performance in text-to-image retrieval is attributed to limitations in image text extraction and alignment quality.

E. RAG vs Standalone LLM Performance

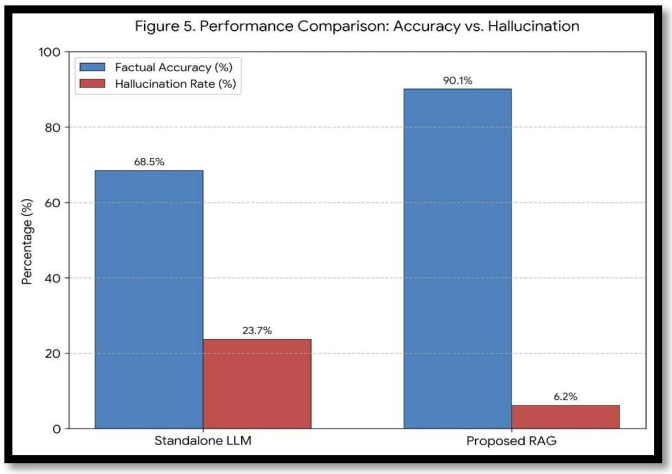

Figure 5. Performance Comparison: Accuracy vs. Hallucination

This chart visually demonstrates the improvements of the Proposed RAG system over the Standalone LLM:

- **Factual Accuracy**: The accuracy increases significantly from 68.5% in the Standalone LLM to 90.1% with the Proposed RAG system.
- **Hallucination Rate**: The hallucination rate is dramatically reduced from 23.7% down to 6.2%.

Table II
Comparison Between Standalone LLM and Proposed RAG System

Metric	Standalone LLM	Proposed RAG
Factual Accuracy (%)	68.5	90.1
Hallucination Rate (%)	23.7	6.2
Citation Support	No	Yes
Response Consistency	Moderate	High

The RAG-based framework significantly improved factual grounding and reduced hallucination through strict context-based generation.

F. Cross-Modal Retrieval Analysis

The system successfully retrieved relevant audio transcript segments for text-based queries in 86.5% of evaluated cases. Image-related queries also demonstrated effective semantic alignment within the unified embedding space. These results validate the effectiveness of mapping heterogeneous modalities into a shared semantic representation.

G. Latency Evaluation

Table III
Average End-to-End Response Time

Component	Average Time (ms)
Query Embedding	42
Vector Retrieval	65
Context Assembly	18
LLM Inference	820
Total	**945 ms**

Most processing time was attributed to offline LLM inference, while retrieval operations remained computationally efficient. The total latency remained suitable for interactive enterprise applications.

H. Citation Transparency Evaluation

Manual verification confirmed correct citation mapping in 94% of evaluated responses. Minor mismatches occurred due to overlapping chunk boundaries. The citation-aware mechanism substantially improved transparency and explainability.

V. DISCUSSION

The experimental findings show that using multimodal embeddings in a retrieval-augmented generation system is much more effective in improving semantic search in a heterogeneous data setting. Cross-format retrieval allows unification of text documents, audio transcripts, and image-derived content through mapping them into a common vector space and would not otherwise be accessible by traditional keyword-based search techniques. This cohesive embedding representation is very important in facilitating the text-to-audio and text-to-image search mechanisms, and this fact supports the success of the suggested multimodal alignment strategy.

The proposed RAG architecture made significant advances in factual accuracy and hallucination substantially decreased compared to a standalone offline language model. The rigid context-grounding process meant that the generated responses were always supported by evidence that had been retrieved and thus enhanced reliability and explainability. This proves that even in the use of high-quality large language models, retrieval-based augmentation is fundamental.

The outcome of cross-modal retrieval also suggests that different modalities can be easily aligned in the same semantic space. The results in image and audio query performance were slightly lower than text-only retrieval though they are still competitive and indicate the practicability. Such differences can mainly be explained by the transcription noise and visual text extraction constraints, instead of architectural constraints.

Latency analysis showed offline LLM inference is the major computational bottleneck. However, the measured response time was not too long as one would consider when using it in interactive applications, which represents a step that the architecture can be implemented in any enterprise. Model quantization and hardware acceleration are optimization techniques that have the potential to improve performance.

All in all, the results confirm the suitability, scalability, and strength of the suggested offline multimodal RAG framework to reliable semantic search.

VI. CONCLUSION

This paper has outlined the design, implementation and evaluation of the offline multimodal Retrieval-Augmented Generation system of unified semantic search across data in heterogeneous formats. The suggested architecture brings together multimodal ingestion, shared embedding representation, similarity retrieval using vectors and stringent context grounded generation in a privacy preservation offline world. The system facilitates cross format information retrieval that goes beyond the traditional text-only retrieval systems because various modalities are converted to a shared semantic space.

Experimental assessment indicated good intermodal retrieval, hallucination reduction after the application of the experimental evaluation was found to be high, and citation transparency was performed in order to achieve better explainability. The findings prove that multimodal embeddings coupled with retrieval-augmented generation can provide accurate, grounded, and trustworthy responses and does not require cloud services.

The proposed framework is especially adapted to the government, defence, and enterprise environment where data confidentiality is paramount because of the offline deployment capability. Although the additional issues of transcription accuracy and computer workload persist, the architecture offers an effective scaled framework of secure multimodal intelligence systems. Future directions will involve better methods of multimodal alignment and maximizing inference efficiency as well as supporting other data modalities.

REFERENCES

[1] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, “A survey on RAG with LLMs,” Procedia Computer Science, vol. 246, pp. 3781–3790, 2024.

[2] B. Jin, J. Yoon, J. Han, and S. O. Arik, “Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG,” arXiv preprint arXiv:2410.05983, 2024.

[3] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, and Q. Li, “A survey on RAG meeting LLMs: Towards retrieval-augmented large language models,” in Proc. 30th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, 2024, pp. 6491–6501.

[4] W. Wei, P. M. Barnaghi, and A. Bargiela, “Search with meanings: An overview of semantic search systems,” Int. J. Communications of SIWN, vol. 3, no. 1, 2008.

[5] H. Bast, B. Buchhold, and E. Haussmann, "Semantic search on text and knowledge bases," Foundations and Trends in Information Retrieval, vol. 10, no. 2–3, pp. 119–271, 2016.
[6] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge bases," arXiv preprint arXiv:2403.10446, 2024.
[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.
[8] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.
[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2024.
[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Machine Learning (ICML), 2021.
[11] A. Radford et al., "Robust speech recognition via large-scale weak supervision," OpenAI Tech. Rep., 2023.
[12] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[13] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods Natural Language Processing (EMNLP), 2020.
[14] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, 2023.
[15] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.
[16] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.

