International Journal of Research in Interdisciplinary Studies
Volume 3, Issue 5, May 2025
https://www.ijris.com | ISSN (Online): 2584-1017


# Developing and Deploying an LLM-Based Tool for Generating Human-Like Responses in Offline Networks

Saira Banu Atham[1], Yenduri Naveen[2*], B. V. Varun Kumar[3], H. S. Arun[4], Shaik Abdul Sameer[5]
[1,2,3,4,5]*Department of Computer Science, Presidency University, Bangalore, India*

**Abstract—The Gemini MultiPDF Chatbot represents a significant advancement in natural language processing with the fusion of Retrieval-Augmented Generation with the Gemini Large Language Model. This chatbot has the ability to handle the complexity of multiple document retrieval and response generation, providing the user with accurate and contextually pertinent information across various PDF documents. Leverage the power of RAG, the system effectively enhances its understanding and response accuracy by dynamically drawing upon external knowledge bases. The process employs PDF parsing and extraction using chunking, followed by embedding creation through vector representation using FAISS indexing to achieve efficient similarity searching and correct recovery of the data. Streamlit is the interface, with a simple, user-friendly platform for uploading documents, querying, and obtaining complete answers.**

***Index Terms*— Semantic Communication, Large Language Models (LLMs), Edge Computing.**

## 1. Introduction

The Gemini MultiPDF Chatbot is an advanced natural language processing (NLP) system that integrates Retrieval-Augmented Generation (RAG) with the Gemini Large Language Model (LLM) to enhance document search, retrieval, and response generation. This chatbot is designed to handle multiple PDF documents, providing accurate and contextually relevant responses while significantly improving information accessibility.

By leveraging RAG techniques, the chatbot retrieves and generates responses dynamically, ensuring up -to-date, accurate, and context-aware answers. Unlike traditional LLMs, which rely solely on their pre-trained knowledge, this system dynamically accesses external knowledge sources, making it a powerful tool for handling large-scale document processing, research analysis, legal documentation, and academic studies. The Gemini LLM excels in various NLP tasks, including:

- Text Summarization – Condensing long documents into concise, meaningful insights.
- Sentiment Analysis – Understanding the emotional tone of text data.
- Language Translation – Converting text between different languages.
- Question Answering – Providing fact-based responses from document sources.

By integrating these functionalities, the Gemini MultiPDF Chatbot significantly enhances document -based knowledge retrieval, reducing the time and effort required for manual information extraction.

Traditional LLMs do not have direct access to external knowledge bases. Their responses are limited to pre - trained data, which may become outdated over time. As a result, they struggle with: Providing real-time, document-specific responses. Handling domain-specific knowledge retrieval. Generating contextually accurate answers from large document repositories. How RAG Enhances Chatbot Efficiency By incorporating Retrieval- Augmented Generation (RAG), the Gemini MultiPDF Chatbot can:

Dynamically retrieve relevant information from multiple PDF sources. Provide precise, real-time responses based on current data. Enhance research efficiency by automating document searching and extraction. Support multiple applications, including academic research, customer support, and legal document analysis. This integration allows the chatbot to overcome traditional LLM constraints, making it a valuable tool for knowledge retrieval in dynamic environments.

The primary objectives of the Gemini MultiPDF Chatbot are Core Functionalities Efficient PDF Parsing and Extraction Develop a robust mechanism to process, extract, and chunk text from PDF files for optimal retrieval. Optimized Data Storage and Indexing Implement FAISS (Facebook AI Similarity Search) indexing to store document embeddings, enabling fast and accurate similarity searches. Enhanced Response Generation with RAG & Gemini LLM Improve response accuracy and relevance using real-time document retrieval in combination with Gemini LLM. User-Friendly Interface for Seamless Interaction Build an interactive a nd accessible UI using Streamlit, allowing users to upload documents, submit queries, and receive instant responses. Extended Goals Scalability and Performance Optimization Ensure the chatbot remains efficient even when handling large document repositories. Multi-Document Retrieval and Summarization Enable the chatbot to summarize and extract key insights from

*Corresponding author: naveenyanduri2439@gmail.com

multiple PDFs simultaneously. Security and Privacy Compliance Ensure that the system functions offline, safeguarding sensitive documents from externa l threats. Support for Multiple Languages and Domains Expand capabilities to support multiple languages and domain-specific terminologies.

The Gemini MultiPDF Chatbot makes several noteworthy contributions to the field of NLP, AI, and information retrieval by addressing key challenges and introducing innovative solutions: Advancing RAG-Driven Document Retrieval Demonstrates the potential of RAG techniques in improving document-based knowledge retrieval. Provides a blueprint for integrating LLMs with external knowledge sources to enhance response quality. Improved Accuracy in Multi-Document Contexts Enhances contextual relevance and response accuracy by dynamica ly retrieving and incorporating information from multiple PDF documents. Minimizes hallucinations and improves factual correctness by grounding responses in document data. Reduction of Information Overload and Manual Effort Automates document parsing, analysis, and response generation, reducing the time and effort required for manual document review.

The development of the Gemini MultiPDF Chatbot addresses critical gaps in document retrieval and NLP-driven response generation by introducing a dynamic, context-aware, and efficient system capable of managing large- scale document repositories. The study is significant in the following ways: Revolutionizing Information Retrieval Improves document accessibility by enabling quick and accurate searches across multiple PDF documents. Reduces manual workload in research, legal analysis, and academic settings by automating information extraction. Enhanced User Experience with NLP Provides human-like, coherent responses by combining context-a ware retrieval with accurate text generation. Ensures that users receive relevant and concise responses, reducing information overload. Enabling Real-Time Knowledge Discovery Empowers users with real-time insights from diverse document sources. Facilitates faster decision-making in high-stakes environments such as legal, healthcare, and corporate sectors. Bridging the Gap Between Static and Dynamic Knowledge Overcomes the limitations of traditional LLMs that are restricted to pre-trained data by dynamically incorporating real-time knowledge from uploaded documents. Enables the chatbot to be domain -specific by customizing its responses to match the content of user-provided documents.

## 2. Background

### *A. Semantic Communication Traditional Communication Systems*

Semantic Communication Traditional communication systems focus on providing high data transmission rates and minimizing symbol (bit) error rates. In such system, Shannon only focused on the technical level, which deals with encoding the source message into a bit sequence and recovering the same sequence at the receiver passing through a noisy channel without considering the underlying meaning of the message. In contrast, recently, a new paradigm has emerged in wireless communications, which shifted the f ocus of communication towards the semantic level, leading to the development of semantic communication. Semantic communication targets the exchanging of semantically equivalent information without necessarily requiring an identical match to the original transmitted information. The core concept of semantic communications is to extract the “meanings” or “features” from the transmitted information at the source and to “interpret” this semantic information at the destination. However, semantic communication requires conventional methods to encode relevant information. To extract the meanings or features, semantic communication should have a Knowledge Base (KB), which needs to be updated periodically based on the information it receives from the user’s input to retrieve and interpret new information as user inputs keep on changing. Overall, the semantic communication system is complex, where both the source and destination need to use the KB using ad vanced DNN or RNN techniques along with the functions of traditional communication. Additionally, the semantic source and destination can perceive their environment and operate autonomously.

### *B. Large Language Models (LLMs)*

LLMs are advanced AI-based systems designed to understand and generate human-like text based on vast amounts of data. The training process of these kinds of AI models requires adjusting millions or even billions of parameters through a technique called back-propagation, and finally, optimizes the models’ ability to understand and generate language. During training, LLMs are exposed to large datasets Switch %$9the lights Decoding noise Encoding Channel Source containing diverse text sources, which allows them to learn the statistical properties of language. After training, the model can generate text by taking an initial prompt as input and predicting subsequent words using diffusion techniques. This approach allows the model to iteratively refine its predictions, maintaining coherence and context throughout the generated sequence based on its learned parameters. Fine-tuning specific tasks or domains can further enhance the model’s performance, making it adept at specialized applications such as technical support, content generation, or conversational agents. The ability to handle diverse linguistic patterns and contexts makes LLMs a powerful tool for a wide range of natural language processing tasks. Recent LLM architectures like GPT-4 and Llama 3.1, can generate coherent and contextually relevant responses, making them valuable tools for applications in natural language processing and human-computer interaction. processing sensory data and receiving a response has significantly reduced compared to other distributed com putting approaches, such as cloud computing. In brief, edge computing improves response times and saves bandwidth by processing the data locally and not sending the requests to the cloud. Thus, Edge computing can significantly enhance the performance and utility of LLMs by bringing computation and data storage closer to the data sources, thereby reducing latency and bandwidth usage. By deploying LLMs on edge devices, it is possible to

process data locally rather than relying on centralized cloud servers. Additionally, edge computing improves privacy and security by keeping sensitive data on local devices rather than transmitting it over the Internet. Moreover, it enables better scalability, as processing loads are distributed across multiple edge nodes rather than concentrated in a central cloud, which can prevent bottlenecks and reduce the risk of downtime of the central cloud server refers to a network of interconnected physical devices that communicate and exchange data with each other and distributed computing systems such as Edge and Cloud over the Internet. Apart from having sensors (to sense the environment) and actuators (to act on the environment) devices must have a communication facility for connectivity. Thus, semantic communication, LLMs, edge computing, and networking can work together to create intelligent, user-friendly, efficient, and quick responsive systems. In such a system devices collect data from the environment and transmit it to edge computing nodes for immediate processing. After that, offline LLM in the edge device can analyze and interpret the received data as a part of semantic communication to understand the context and intent of the transmitted data. Hence, such an architecture can enable real-time, context aware responses and decisions, a user-friendly and efficient network eco-system to improv crossover.

## 3. Proposed Methodology

### *A. System Model*

The proposed methodology integrates Large Language Models (LLMs), semantic communication, and edge computing to enhance networks' efficiency, responsiveness, and intelligence. Unlike traditional communication systems that rely on raw bit transmission, our approach focuses on meaning-based communication while leveraging LLMs at the network edge to improve data processing, reduce bandwidth usage, and ensure context- aware decision-making. The framework consists of the following core components:

Semantic Communication Module – Extracts meaningful information instead of transmitting raw data. LLM Processing Unit – Uses LLMs for feature extraction, intent recognition, and response generation. Edge Computing Infrastructure – Processes data near devices to reduce latency and improve efficiency. Device Layer – Sensors, actuators, and smart devices generate real-time data for semantic communication. Security & Privacy Layer – Ensures data integrity, encryption, and decentralized processing to mitigate risks. This methodology enables intelligent systems capable of real-time adaptation, efficient resource utilization, and privacy-preserving communication.

### *B. Advantages of using LLM for Semantic Communication*

*Networks Efficient Data Transmission*: Semantic communication for cusses on transmitting the meaning or the semantics of the data rather than the raw data itself. This can significantly reduce the amount of data that needs to be transmitted, leading to more efficient use of bandwidth and lower energy consumption of the senders, which is crucial for many devices that operate on limited power sources. Enhanced Decision

*Making*: Semantic communication can improve the quality of decision-making processes in various applications, such as smart cities, industrial automation, and healthcare by focusing on the meaning of the data, which can lead to more relevant and actionable information. Additionally, semantic communication enables systems to understand the context in which data is generated and used. This context -awareness can lead to the development of more intelligent and adaptive services that can respond dynamically to changing conditions. Quicker responsiveness: As LLM in semantic communication can extract information in depth from minimal input size, therefore, it helps in quicker decision-making as transmission timing, processing timing, and queuing timing get reduced with the size of the data that has to be transmitted. Enhanced User Experience Users can issue simple, natural language commands, which the system interprets and executes accurately, thereby minimizing the likelihood of errors. By considering historical data, the system can further consider user preferences and adjust settings, accordingly, leading to a more personalized and intuitive experience. Thus, natural language interactions powered by LLMs make it easier for users to communicate with devices, enhancing the overall user experience.

## 4. System Design and Implementation

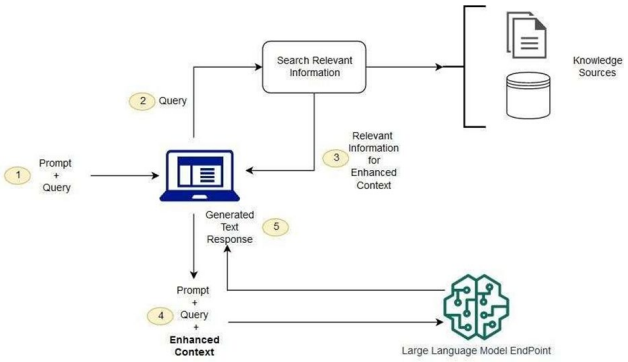


Fig. 1. System architecture

This section discusses some of the examples of how LLMs can be integrated into Edge-based systems as a part of semantic communication: Smart Home Assistants: LLMs can make smart home assistants much more powerful by enabling them to understand and respond to more natural and context -aware interactions. For instance, with the help of an LLM, a smart assistant can interpret and execute complex voice commands. This means users can speak more naturally and still have their commands accurately followed, making the interaction with smart home devices smoother and more intuitive. LLMs can also manage multiple devices, like lights, thermostats, and security systems, seamlessly integrating their functions based on user preferences and context. Furthermore, using historical data enables the system to learn user preferences and habits over time. For example, it can adjust temperature settings based on past behavior or suggest frequently used commands. Additionally, historical data can be used to understand the

context of the command in a better way. For instance, if it knows that a user typically dims the lights and plays soft music in the evening, it can proactively make these adjustments without explicit commands. Overall, the entire system can create a more personalized and intuitive user experience. Industrial In industrial devices are deployed to continuously monitor machinery and equipment. Sensor attached devices collect vast amounts of data, including temperature readings, vibration levels, and operational cycles, etc. By understanding the intricate patterns and contexts within the data, LLMs can identify possible issues that may not be determined by the existing system and accordingly, LLMs can suggest detailed maintenance schedules and recommendations. Preemptive replacement recommendations before the actual failure and real-time alerts can significantly improve the efficiency of the Industrial ecosystems by preventing unexpected downtime. The efficiency of industries can be further improved by allowing LLMs to optimize the overall maintenance strategy by learning from historical data. They can identify trends such as recurring failures or seasonal variations in equipment performance, allowing for more strategic planning. Furthermore, LLM generated responses can simplify complex information, making it easy for non-technical individuals to understand and act on maintenance needs. Ensures timely medical intervention, potentially preventing more serious health complications. Moreover, LLM-based semantic communication can significantly improve coordination and efficiency in medical environments such as surgical operations. For example, in a scenario where a doctor performs a complex surgery and uses voice commands to interact with various medical devices and communicate with the surgical team. For example, the doctor might say, “Increase the oxygen level by 10%” and the LLM would interpret this command, adjust the settings on the relevant device, and confirm the change. This allows the doctor to focus on the surgery without manually adjusting equipment. These examples illustrate how LLM-based semantic communication can add significant value to systems by providing advanced data analysis, context-aware interactions, and predictive capabilities.

## 5. Conclusion

In conclusion, the Gemini MultiPDF Chatbot sets a new benchmark in AI -powered document retrieval and response generation by blending cutting-edge NLP, RAG, and LLM technologies. Its ability to handle large-scale document repositories with speed, accuracy, and contextual awareness makes it an invaluable tool across various domains. By automating document processing tasks, improving knowledge retrieval, and ensuring data privacy, the chatbot revolutionizes the way information is accessed and utilized. As the system continues to evolve with future enhancements such as multi-language support, OCR integration, and hybrid retrieval methods, it is poised to become an indispensable resource for professionals seeking efficient, secure, and intelligent document management solutions.

## References

[1] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. University of Illinois Press, 1949.

[2] X. Luo, H. Chen, and Q. Guo, “Semantic communications: Overview, open issues, and future research directions,” IEEE Wireless Communications, vol. 29, no. 1, pp. 210–219, 2022.

[3] E. Grassucci, J. Park, S. Barbarossa, S.-L. Kim, J. Choi, and D. Comminiello, “Generative AI meets semantic communication: Evolution and revolution of communication tasks,” 2024.

[4] J. Park, S. Samarakoon, A. Elgabli, J. Kim, M. Bennis, S.-L. Kim, and M. Debbah, “Communication- efficient and distribute learning over wireless networks: Principles and applications,” Proceedings of the IEEE, vol. 109, no. 5, pp. 796– 819, 2021.

[5] A. Hazra, A. Kalita, and M. Gurusamy, “Meeting the requirements of internet of things: The promise of edge computing,” IEEE Internet of Things Journal, vol. 11, no. 5, pp. 7474–7498, 2024.

[6] A. Kalita and M. Khatua, “6TiSCH – IPv6 enabled open stack IoT network formation: A review,” ACM Trans. Internet Things, vol. 3, no. 3, pp. 24:1–24:36, Jul. 2022.

[7] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.

[8] J. Dai, P. Zhang, K. Niu, S. Wang, Z. Si, and X. Qin, “Communication beyond transmitting bits: Semantics- guided source and channel coding,” IEEE Wireless Communications, vol. 30, no. 4, pp. 170–177, 2023.

[9] D. Gu¨ndu¨z, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, “Beyond transmitting bits: Context, semantics, and task-oriented communications,” IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 5–41, 2023.

[10] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, “Pushing large language models to the 6G edge: Vision, challenges, and opportunities,” 2024.

[11] S. Barbarossa, D. Comminiello, E. Grassucci, F. Pezone, S. Sardel- litti, and P. Di Lorenzo, “Semantic communications based on adaptive generative models and information bottleneck,” IEEE Communications Magazine, vol. 61, no. 11, pp. 36–41, 2023.

[12] H. Nam, J. Park, J. Choi, M. Bennis, and S.-L. Kim, “Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation,” 2023.