STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 12, Issue 3, 2025
September–December
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

Leveraging Large Language Models for Personalized Document Summarization and Question Answering: An Architecture for Stoner-Friendly Chatbots

Niket Singh[1,*], Raj Singh[2]

Abstract

This study presents a detailed framework for developing personalized chatbots that utilize large language models (LLMs) to process and extract information from extensive documents while effectively responding to user inquiries. The proposed system is designed to mitigate information overload by employing advanced natural language processing techniques, leveraging technologies such as OpenAI, LangChain, and Streamlit. By integrating these tools, the framework enhances knowledge retrieval, simplifies document comprehension, and improves overall productivity. The study delves into the architecture, implementation, and practical applications of the framework, demonstrating its ability to streamline access to critical information. Furthermore, it explores how developers and researchers can leverage this system to create end-to-end solutions for document summarization and automated question-answering. Through a structured and step-by-step approach, this research provides valuable insights into constructing intelligent chatbot systems capable of efficiently managing vast amounts of textual data. Ultimately, the proposed framework contributes to the advancement of AI-driven knowledge management and human-computer interaction.

Keywords: Large language models, langchain, chatbots, streamlit, open AI

INTRODUCTION

The capacity to divert substantial attention away from the massive volume of textbooks has peaked in the period marked by the explosive growth of digital data. Chatbots, which are created using artificial intelligence (AI) and natural language processing (NLP) technology, have been as an adaptable remedy for this issue. From customer service to instructional support, chatbots offer vibrant avenues for automating tasks like document recapitulation and question response. Productivity and knowledge reclamation rise as a result [1]. Large language models (LLMs) are a crucial part of current developments in natural language processing (NLP) because they change how robots understand and produce textbooks that appear human. Because these models have been trained on big datasets and shown exceptional abilities in comprehending and generating natural language, they are perfect for enabling technical chatbots for tasks like document summarization and question answering [2]. It is impossible to overestimate the significance of LLMs in NLP activities since they enable chatbots to examine and synthesize complicated textual information with unmatched delicacy and efficacy [3].

***Author for Correspondence**
Niket Singh
E-mail: niketsingh199@gmail.com

[1]Research Scholar, MCA, Department of Computer Science, Thakur Institute of Management Studies, Career Development & Research TIMSCDR), Mumbai, Maharashtra, India
[2]Research Scholar, MCA, Department of Computer Science, Thakur Institute of Management Studies, Career Development & Research TIMSCDR), Mumbai, Maharashtra, India

Received Date: March 06, 2025
Accepted Date: April 06, 2025
Published Date: July 08, 2025

Citation: Niket Singh, Raj Singh. Leveraging Large Language Models for Personalized Document Summarization and Question Answering: An Architecture for Stoner-Friendly Chatbots. Journal of Artificial Intelligence Research & Advances. 2025; 12(3): 88–93p.

© STM Journals 2025. All Rights Reserved

General Study

This review looks at how NLP technology is developing, how important LLMs are in helping to reevaluate chatbot capabilities, and how important it is to have validated chatbots that can answer queries and summarize documents in order to help with the problems brought on by an excessive amount of data. In order to solve the information overload in the scientific literature, the composition used extractive summarizers to improve the key components of exploratory reports [1]. Through experimentation, Balage *et al.* established improvements taking into account the textbook's intricate organization and superior summaries, particularly for languages lacking advanced NLP tools. This methodology addressed the difficulties of creating relevant receptivity from large volumes of scientific textbooks, highlighting the necessity of a thorough comprehension of texts for superior summaries [2]. Bang *et al.* investigated chatbots and conversational interfaces in the context of AI ethics. This study investigated the disparities in recommendation generation between script grounded and large language model (LLM)-grounded chatbots, as well as the ethical defenses of LLM-grounded suggestions. This study emphasized that enterprises have comparable translucency, justice, sequestration, and accountability by comparing their traits and limits.

This study emphasized the need to take ethics into account when creating and evaluating conversational AI systems, which calls for more research and support when putting ethical principles into practice. In an effort to overcome the difficulties with automatic textbook summarizing, Prasad *et al*. used machine literacy to create a dynamic connectionist textbook summarizer [3]. The authors demonstrated that adaptive structures in connectionist infrastructures improve the efficacy of summarization by acknowledging the dynamic character of linguistic qualities. The limitations of this technique highlight the need for language-independent solutions for reliable textbook summarization in modeling dynamic systems. In response to the exponential expansion of online material, rapid advancements in automatic textbook summary technologies have been explored by Brilliant *et al*. [4].

Liu *et al.* created GPT-3, an API for joyful generation that exercises the OpenAI language model, to let people and corporations create content more quickly [5]. Equipped with sophisticated machine-learning techniques, such as a recurrent neural network (RNN) architecture, the instrument endeavors to effectively generate superior content for various vibrant platforms. It featured a range of options akin to Facebook ads, LinkedIn postings, Amazon product descriptions, and blogs, all with a dashboard that was pleasant to stoners. Technology highlighted its effectiveness in streamlining content development across several platforms, underscoring its usefulness in supporting pharmaceuticals with varying content generating requirements. It also addressed the issues of limited writing skills and time restrictions. LangChain, a query system that exercises LLMs for efficient information reclamation from PDF documents, was introduced by Gaur and Saunshi [6]. The authors demonstrated that LangChain improved information reclamation and expedited the querying process by utilizing StreamLit and natural language processing methods. This method solved the problems of extracting relevant data from PDFs and provided a useful tool for efficient data access. Mansurova *et al*. investigated a novel method for textbook summarization models by using large language models (LLMs) as gold-standard oracles [7], like GPT-3.5 [8–10].

Investigated the rebuttals to the claims that model training and assessment procedures should be based on LLMs. The investigation examined videlicet GPT Score and GPT Rank, two LLM grounded approaches for evaluating summary quality, in conjunction with contrastive literacy training techniques that make use of LLM-guided signals. Experiments on the CNN/Daily Mail and Sum datasets showed that, when estimated with LLM-grounded criteria, lesser summarization models might achieve performance comparable to LLMs. However, moral evaluation showed a discrepancy, indicating that lower models have not yet attained the performance position of LLMs despite improvements in the suggested training techniques. The investigation emphasized the drawbacks and advantages of LLM as a reference environment, highlighting the need for more research and improvement. It made a contribution by highlighting the limits of LLM-grounded training and assessment methods and by showcasing empirical improvements in lower models taught with LLM references and contrastive

© STM Journals 2025. All Rights Reserved

literacy. The experimenters investigated the use of LLMs for rapid-fire operation development using LangChain, an open-source software library, as the focal point of their investigation, which is described by Pokhrel and Banjade [11]. Emphasizing LLMs, like OpenAI's ChatGPT, which is well-known for doing tasks like writing essays and creating laws, the investigation highlighted the modular design of LangChain.

Through real-world examples spanning chatbots, autonomous agents, and document-based question answering, the research demonstrated LangChain's ability to accelerate operational development.

It highlighted how LLMs have revolutionized AI geography and established LangChain as an essential tool for expediting the development process, encouraging continued research and innovation in the field. The investigation of Shibi *et al*. focused on the ability of LLMs, including LaMDA, OPT, and GPT-3, to solve calculating word problems [12]. The "da Vinci-002" model from GPT-3 performed well on both symbolic and numerical tasks using the STAMP dataset. In the numerical test set, a two-step method was used to perfect delicacy. Certain egging techniques improve the model's capacity to clarify the research process and resolve challenging issues. Large LLMs have the potential to shatter symbolic computation issues, according to the study, however there is still space for improvement. To improve Large Language Models (LLMs) like ChatGPT, Nalini *et al.* included an external knowledge operation module that allowed users to access data from vector databases and the Internet [13]. Using the double diamond design process, a chatbot prototype was created with frontend and backend layers, taking into account elements such as prompt templates, chains, memory, models, and agents in order to increase awareness of blockchain technology in Kazakhstan. The GPT-3.5 from OpenAI and Python were selected due to their adaptability in natural language processing. The chatbot's capacity to deliver precise information in real time was enhanced by the integration of web hunt and semantic technologies.

Cosine similarity was used to assess the verdicts, and a summary was created from the top five. Elmo embedding makes it easier to provide correct summaries. Perfecting summary for various papers and increasing delicacy and speed are among the unborn work's objectives. The proposal proposes to enhance communication and decision making capabilities of multiagent systems (mass) by using LLMs, such as GPT-grounded technologies. Based on the MADE-K model, a new agent armature has been created to improve reasoning, decision-making, and conversational characteristics. Integration demonstrated the implicit reworking of agent relations and Mass's capacity for problem-solving through a business script. Obstacles akin to computational output and opinion interpretability were acknowledged for future development. A web operation with the goal of effectively overcoming the difficulties in sifting through videotape material by generating brief summaries of YouTube videotape reiterations utilizing NLP techniques and the Flask frame. By providing capabilities similar to restatement and textbook-to-speech, the writers showed that the summarizer boosted stoner access to videotape material, prostrating issues in comprehending long, complex content and enhancing overall understanding. The architecture of the model is shown is Figure 1.

To overcome the shortcomings of the available technologies for setting up large-scale model operations, a novel approach was presented By integrating the LangChain frame, this technology makes it possible for AI models such as Chat GLM-6B to seamlessly communicate with the original data sources (Figure 2). Furthermore, the system included the NASA frame, which improved its capacity to carry out tasks like reality identification and intent brackets. Using these fabrics, the system increased the efficiency of textbook creation and allowed for more efficient operations in vibrant environments, akin to modern network scheduling and monitoring systems. The literature study makes it clear that combining AI and NLP technologies especially large language models (LLMs) offers interesting solutions for dealing with information overload.

These technologies enable chatbots to perform very well in activities like question answering and document summarizing, showcasing their efficacy in handling enormous amounts of digital data from a variety of fields.

© STM Journals 2025. All Rights Reserved

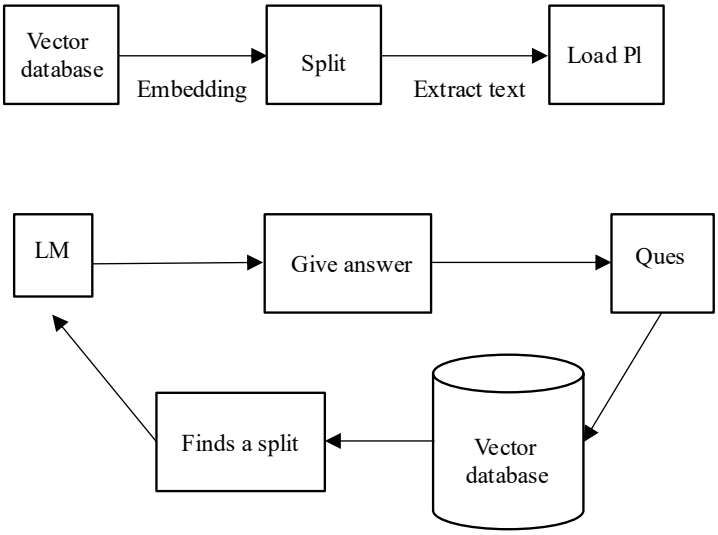

Figure 1. Architecture of the model.

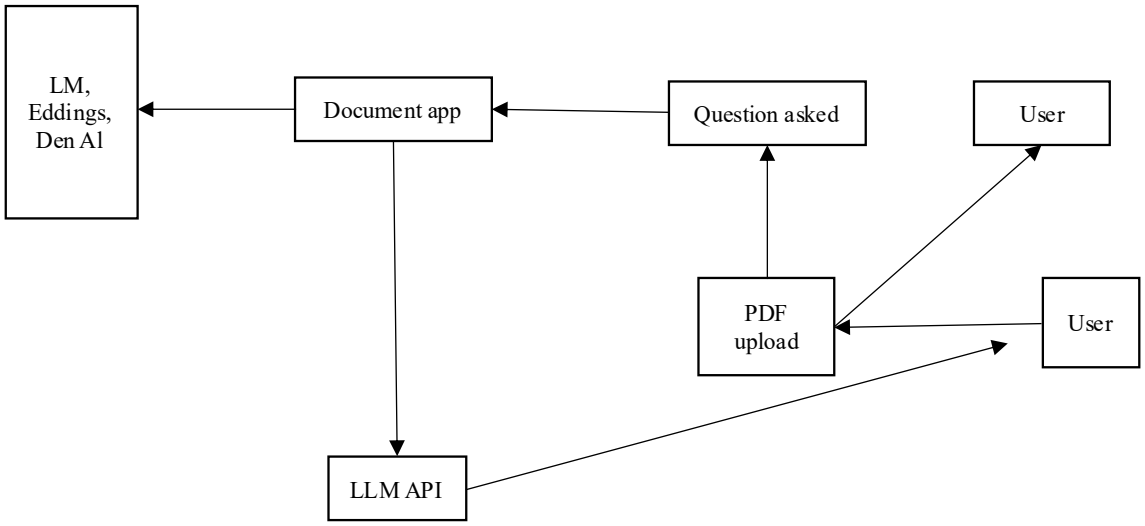

Figure 2. Block diagram.

Architecture

OpenAI is an exploring association that performs research across several AI areas, including NLP, robotics, underpinning literacy, and beyond, with the goal of enhancing artificial intelligence technology for the benefit of society. The creation of AI systems capable of performing a new set of activities with intellect comparable to that of a mortal is the main goal. With the use of large-scale language models like the GPT series, which may cause textbooks to suggest new language based on input data, they have managed to create artificial intelligence systems that are capable of considerable capabilities. These models function in terms of restatement, textbook creation, and natural language appreciation. The system's framework is to create a trap operation that uses the stream lighted, LangChain, and OpenAI APIs to summarize the PDF; Figure 1 illustrates this process and shows the model's armature.

The reader class was imported from the PyPDF2 module once the PDF file was initially submitted to the system. Working with PDF lines is made easier with this Python archive, which enables activities like interpretation, manipulation, and textbook creation from PDF documents. Every knob in the page is resolved into a gobbet, which is then saved in the vector database. This process creates an embedding from every gobbet. A member of the data that has been isolated from a larger dataset is referred to as a

© STM Journals 2025. All Rights Reserved

knob in the context of data processing. It is often anticipated that it will be divided into more manageable sections for storage, processing, or dissection. Quantitative data representations that synchronize their semantic meanings are called embeddings Word embeddings are widely assumed in natural language processing (NLP) to represent words as vectors in a high-dimensional space.

Words with comparable meanings are arranged closer together in the vector room according to this representation. Several techniques, such as Word2Vec, Glove, or deep literacy models like mills, can be used to create these embeddings. We used the Open Aim embedding class from LangChain embeddings in this investigation It is a part of the LangChain package and offers embeddings for textbook data that are based on OpenAI's language models. Vectors that define data are saved in a repository called a vector store, sometimes referred to as a knowledge base. The vector store serves as a knowledge warehouse where the system may access the embeddings generated from the indexed data while performing searches This vector storage enables semantic searching and the retrieval of relevant data based on stoner queries. In response to a query prompt, the stoner searches the vector store semantically and queries the gobbets to find ranking effects.

The OpenAI API was integrated at the backend and provides the user with replies, all based on the extensive language model. But they have the option to exercise it. If not, the stoner will question with a more thorough personalized prompt and remain for the particular response, provided the stoner finds satisfaction in the rejoinder. Figure 2 shows the system block that is suggested.

By providing access to slice-bite language processing capabilities, the OpenAI API facilitates the development of enhanced LLMs that resemble GPT-3, GPT-3.5, and GPT-4. These models do exceptionally well at producing textbooks in a variety of vibrant styles and tones, which makes activities like content coinage, summary, and handwriting production easier. In addition to the OpenAI API, LangChain is an open-source archive that works well with other natural language processing (NLP) tools. This makes it easier to create reliable channels for activities like data drawing and summarization. By streamlining the development of interactive trap operations, Streamlite improves this ecosystem by enabling developers to showcase work produced by OpenAI and LangChain with the least amount of software required. When combined, these technologies provide enhanced language processing performance, optimize processes, and generate operations that seem professional.

Outcome

A kind of technology that is included into multiagent systems (MASs) to improve rigidity and communication is a tone-adaptive large language model (LLM). It makes use of slice-edge large language models, like to GPT-4, to help agents adapt to challenging jobs and respond wisely to evolving circumstances. During our trial assessment, we used a variety of papers from a wide range of diverse fields. On PDF summaries, this approach showed notable efficacy. This technique outperformed conventional styles in extracting the important information from papers in brief summaries. With a great deal of stoner pleasure, the question-answering chatbot provided quick solutions to each query in a matter of seconds. It retrieves documents from a retriever first, then uses a question-answering chain to answer inquiries based on the documents that were recovered. In particular, this system used the 'stuff' document chain type, in which a prompt is created from a list of documents and then reused by a Large Language Model (LLM) [14]. By comprehending the intent and context of hunt queries, semantic hunt improves the capabilities of hunt machines and produces more relevant and accurate hunt outcomes.

CONCLUSION

This study culminates in a comprehensive method for creating validated chatbots derived from large language models (LLMs), with a focus on question answering and document summarization through the integration of technologies like StreamLit, LangChain, and OpenAI. This effectively tackles the problem of information overload by facilitating the emergence of receptivity from documents. This investigation provides a step-by-step explanation on how innovators may utilize the frame to do end-

© STM Journals 2025. All Rights Reserved

to-end operations for document summarizing and question answering. The combination of StreamLit's user-friendly interface design, LangChain's efficient natural language processing, and OpenAI's sophisticated language models provides a versatile outcome for researchers and developers looking to use LLMs for jobs that are more founded in textbook knowledge. Unborn suggestions include optimizing the model, including generative AI models that are adaptable, and extending the functionalities of personalized chatbots to encompass more features. This framework might alter how medications interact with and determine receptivity from textual input, increase productivity, and facilitate the retrieval of information across a variety of diverse fields.

REFERENCES

1. Balage Filho PP, Pardo TS, Nunes MD. Summarizing scientific texts: Experiments with extractive summarizers. In IEEE Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007). 2007 Oct 20; 520–524.
2. Bang J, Lee BT, Park P. Examination of ethical principles for LLM-based recommendations in conversational AI. In 2023 IEEE International Conference on Platform Technology and Service (PlatCon). 2023 Aug 16; 109–113.
3. Prasad RS, Kulkarni UV, Prasad JR. Machine learning in evolving connectionist text summarizer. In 2009 IEEE 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication. 2009 Aug 20; 539–543.
4. Brilliant M, Nurhasanah IA, Oktaria H, Handoko D. Digital heritage portal based on progressive web app: Efforts for the development of cultural heritage and tourism in Lampung. TEKNOKOM. 2024; 7(1): 165–71.
5. 8Liu Y, Shi K, He KS, Ye L, Fabbri AR, Liu P, Radev D, Cohan A. On learning to summarize with large language models as references. arXiv preprint arXiv:2305.14239. 2023 May 23.
6. 9Gaur V, Saunshi N. Symbolic math reasoning with language models. In 2022 IEEE MIT Undergraduate Research Technology Conference (URTC). 2022 Sep 30; 1–5.
7. 10Mansurova A, Nugumanova A, Makhambetova Z. Development of a question answering chatbot for blockchain domain. Sci J Astana IT Univ. 2023 Sep 30; 15: 27–40.
8. Monks T, Harper A. Improving the usability of open health service delivery simulation models using Python and web apps. NIHR Open Res. 2023 Dec 15; 3: 48.
9. 7Sai PJ. An effective query system using LLMs and LangChain. Int J Eng Res Technol. 2023 Jun; 12(06): 367–369.
10. 13Patil DD, Dhotre DR, Gawande GS, Mate DS, Shelke MV, Bhoye TS. Transformative trends in generative ai: Harnessing large language models for natural language understanding and generation. Int J Intell Syst Appl Eng. 2024; 12(4s): 309–19.
11. Pokhrel S, Banjade SR. AI content generation technology based on open AI language model. Journal of Artificial Intelligence and Capsule Networks. 2023 Dec 18; 5(4): 534–48.
12. Shibi K, Grace RK, Geetha MS. Abstractive Summarizer using Bi-LSTM. In 2022 IEEE International Conference on Edge Computing and Applications (ICECAA). 2022 Oct 13; 1605–1609.
13. Nalini N, Narayan A, Sridharan AM, Pradhan A. Automated text summarizer using google pegasus. In 2023 IEEE International Conference on Smart Systems for applications in Electrical Sciences (ICSSES). 2023 Jul 7; 1–4.
14. Topsakal O, Akinci TC. Creating large language model applications utilizing LangChain: A primer on developing LLM apps fast. In International conference on applied engineering and natural sciences. 2023 Jul 10; 1(1): 1050–1056.

© STM Journals 2025. All Rights Reserved