DOI: https://doi.org/10.15276/aait.07.2024.10
UDC 004.032.26:004.946

Effective documentation practices for enhancing user interaction through GPT-powered conversational interfaces

Oleksii I. Sheremet[1)]
ORCID: https://orcid.org/0000-0003-1298-3617; sheremet-oleksii@ukr.net. Scopus Author ID: 57170410800
Oleksandr V. Sadovoi[2)]
ORCID: https://orcid.org/0000-0001-9739-3661; sadovoyav@ukr.net. Scopus Author ID: 57205432765
Kateryna S. Sheremet[1)]
ORCID: https://orcid.org/0000-0003-3783-5274; artks@ukr.net. Scopus Author ID: 57207768511
Yuliia V. Sokhina[2)]
ORCID: https://orcid.org/0000-0002-4329-5182; jvsokhina@gmail.com. Scopus Author ID: 57205445522
[1)] Donbas State Engineering Academy, 39, Mashinobudivnykiv Blvd. Kramatorsk, 84313, Ukraine
[2)] Dniprovsky State Technical University, 2, Dniprobudivska Str. Kamyanske, 51918, Ukraine

ABSTRACT

The article presents a detailed overview of the integration of ChatGPT with PDF documents using the LangChain infrastructure, highlighting significant advances in natural language processing and information retrieval. This approach offers the advantage of not being limited to working exclusively with PDF documents. By leveraging the special capabilities of the LangChain infrastructure, it is possible to interact with any data files containing text information. The literature review highlights the transformative impact of OpenAI's GPT series of models on natural language processing, with advancements in GPT-4 significantly enhancing the generation of human-like text and setting new standards for interactive artificial intelligence applications. The analysis of OpenAI's application programming interface demonstrates its significant role in advancing the integration of artificial intelligence into various applications by providing accessible and robust tools that enable developers and enterprises to seamlessly incorporate sophisticated artificial intelligence functionalities. Despite their advantages, these interfaces face challenges such as latency, processing capacity limitations, and ethical considerations, which necessitate strategic implementation and continuous evaluation to fully harness their potential. The article examines the role of vector data representations, particularly vector embeddings, in enhancing the functionality of artificial intelligence and machine learning systems. These embeddings transform complex textual data into high-dimensional numerical formats, enabling artificial intelligence models to perform tasks such as language understanding, text generation, and data analysis with increased precision and depth. Vector databases play a critical role in managing and leveraging high-dimensional data, specifically vector embeddings, to enhance the operational efficiency of large language models. These specialized storage systems are optimized for handling complex data representations, enabling advanced applications such as text summarization, translation, and question-answering with high accuracy and contextual understanding. LangChain provides a versatile framework that bridges large language models and diverse data sources by utilizing vector databases. This integration enhances the AI's capabilities in data analysis and natural language processing, enabling sophisticated applications that can efficiently interpret and respond to user queries across various datasets. Developing a comprehensive application using LangChain and ChatGPT for PDF document interaction requires meticulous technical considerations. Key elements include efficient data management through LangChain's data loaders and text splitters, which transform PDFs into manageable formats and ensure coherent segmentation for accurate AI interaction. Additionally, implementing vector embeddings enhances the AI's ability to comprehend and analyze textual data, while a user-friendly interface and robust security measures ensure optimal user engagement and data protection. The practical implications of this technology are significant, with potential improvements in customer support by reducing resolution times by up to 40 %, streamlining academic literature reviews by approximately 60%, and boosting productivity in data analysis by saving an estimated 50 % of the time spent on manual data extraction.

Keywords: ChatGPT; LangChain; vector embeddings; data analysis; retrieval augmented generation

***For citation*: Sheremet O. I., Sadovoi O. V., Sheremet K. S., Sokhina Yu. V. "Effective documentation practices for enhancing user interaction through GPT-powered conversational interfaces". *Applied Aspects of Information Technology.* 2024; Vol. 7 No.2: 135–150. DOI: https://doi.org/10.15276/aait.07.2024.10**

INTRODUCTION

The integration of ChatGPT with portable document format (PDF) files using LangChain represents a significant advancement in the field of natural language processing (NLP) and information retrieval.This technology combines the power of OpenAI's GPT-4, a cutting-edge language model, with LangChain's versatile framework to create a system capable of interacting intelligently with PDF documents.

At the core of this integration is the creation of sophisticated chatbots that can comprehend, analyze, and respond to queries based on the content of PDF documents. This capability is achieved by leveraging GPT-4's advanced language understanding, which allows it to interpret and generate human-like text,

and LangChain's efficient management and processing of PDF data. By combining these technologies, developers can build applications that provide dynamic interactions with PDF content, ranging from simple data retrieval to complex analysis and conversation-based information extraction.

The practical applications of this technology are extensive and varied. In customer support, the ability to quickly access and relay information from PDFs can dramatically improve service quality and efficiency. In academic research, the technology can streamline the process of sifting through extensive databases, making the literature review process more manageable and less time-consuming. In data analysis, automating the extraction of data points from thousands of documents can significantly enhance workflow efficiency and productivity. This interactive capability allows users to engage with PDF documents as if they were conversing with a knowledgeable assistant, asking questions and receiving precise answers extracted from the PDFs. This not only accelerates the process of information retrieval but also makes it more accessible to a broader range of users, regardless of their technical expertise.

Moreover, this technology is not just limited to reading and extracting data from PDFs. It extends to the realm of web applications, where it can be used to create interactive and user-friendly interfaces. By incorporating elements like chat interfaces, text boxes for queries, and components for rendering PDF pages, users can enjoy a seamless and integrated experience while interacting with PDFs.

In essence, the integration of ChatGPT with PDF files through LangChain is a leap forward in making digital information more accessible and interactive. It stands as a testament to the ongoing evolution of artificial intelligence (AI) and its increasing role in enhancing human ability to manage and utilize the vast amounts of data in the digital world.

THE PURPOSE OF THE ARTICLE

The purpose of this article is to provide a comprehensive review and detailed analysis of the integration of ChatGPT with PDF documents using the LangChain infrastructure, aiming to highlight significant advancements in NLP and information retrieval. This integration seeks to explore practical applications and implications across various sectors.

To achieve this aim, the following tasks are formulated:

– examine the integration of ChatGPT with PDF documents, analyzing the combination of OpenAI's GPT-4 and LangChain's framework to enhance interaction with PDF files and other text-containing data files;

– assess the role of vector data representations, particularly vector embeddings, in enhancing the functionality of artificial intelligence and machine learning systems, and their applications in tasks such as text summarization, translation, and question-answering (QA);

– evaluate the capabilities of vector databases in managing and leveraging high-dimensional data, specifically vector embeddings, to enhance the operational efficiency of large language models (LLMs);

– analyze the capabilities of the LangChain framework in bridging LLMs with diverse data sources using vector databases, and its innovative features like query chains and retrieval-augmented generation (RAG) for sophisticated data analysis;

– discuss technical considerations for developing applications using LangChain and ChatGPT, including efficient data management through LangChain's data loaders and text splitters, and the importance of implementing vector embeddings, user-friendly interfaces, and robust security measures.

– explore the practical implications of this technology on customer support, academia, and data analysis, including potential improvements in service efficiency, literature review processes, and workflow productivity.

LITERATURE REVIEW

The evolution of conversational AI, particularly with OpenAI's GPT series, has fundamentally transformed capabilities in natural language understanding and generation. Brown, T. B., et al. [1] introduced GPT-3, illustrating its ability to generate coherent, contextually relevant text across a broad array of topics and formats. Other researchers [2, 3] extended these advancements with GPT-4, which improved the generation of human-like text for complex interactions [4], setting a new standard for interactive AI applications [5].

The integration of AI with document processing has moved from traditional methods to sophisticated deep learning techniques. Lample, G., et al. [6] demonstrated how deep learning is crucial for understanding unstructured data, pivotal for AI's integration with PDF documents. Techniques such as named entity recognition and optical character recognition, discussed by Ye, Q. and Doermann, D. [7] have refined AI's capability to parse and interpret complex documents, essential for conversational

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

interfaces interacting with PDFs.

Vector embeddings have revolutionized the way machines process human language. Mikolov, T., et. al. [8, 9] introduced Word2Vec, a technique transforming words into vector spaces, thus enabling semantic understanding.

The LangChain framework [10] offers a structured approach for integrating LLMs with diverse data sources, including PDF documents, enhancing AI's interaction with and analysis of document content through modular components like data loaders and text splitters.

Recent studies have demonstrated the expediency and effectiveness of using LLMs in various practical applications. For instance, an automated evaluation method for personalized text generation (AuPEL) employs metrics such as BLEU, ROUGE, and classification accuracy to objectively measure the quality, relevance, and personalization of generated text, significantly saving time and effort compared to manual evaluations [11]. Additionally, the transformative potential of LLMs in streamlining data science processes is evidenced by their ability to reduce preprocessing time by up to 50 %, and automate tasks like data exploration and model building with substantial accuracy improvements [12]. Further, statistical methods like McNemar's test and 5×2 cross-validation are recommended for evaluating model performance, ensuring observed improvements are statistically significant [13].

In addition to centralized AI models, there is growing interest in free and open-source LLMs that can be run locally on personal or on-premise servers. These models, like EleutherAI's GPT-Neo and GPT-J [14], provide developers with the flexibility to deploy advanced AI capabilities without relying on cloud services, addressing both privacy concerns and reducing latency in applications that require immediate processing. Such models are especially valuable in environments where data security is paramount, offering the necessary tools to build personalized, secure AI solutions.

Ethical considerations, especially in processing sensitive information, continue to demand rigorous standards and careful implementation, highlighted by Bender, E. M., et al. [15]. These considerations are now as crucial as technological advancements in AI development.

As AI and document interaction technologies advance, innovations in vector embeddings and machine learning models continue to enhance the accuracy and efficiency of information extraction and interaction [16]. The development of intuitive and seamless interfaces for AI-document interactions is crucial for maximizing technology utility and user experience.

ANALYSIS OF OPENAI'S API

The landscape of artificial intelligence has undergone transformative changes with the advent of OpenAI's Application Programming Interface (API), which serve as critical enablers for developers and enterprises aiming to integrate sophisticated AI capabilities into their applications. These APIs act as crucial bridges, linking OpenAI's advanced machine learning models to user-specific software solutions, thereby democratizing the use of state-of-the-art AI technologies without necessitating the in-house development of complex models. This accessibility broadens the application of AI, enabling diverse sectors to adopt and integrate cutting-edge technology effectively.

Diverse capabilities and strengths. OpenAI offers an array of APIs, each designed to cater to different facets of artificial intelligence. Notably, the GPT-3 API is engineered for advanced language generation capabilities, facilitating a wide range of linguistic tasks such as text summarization, language translation, and QA. Additionally, the DALL-E API extends OpenAI's prowess into the realm of computer vision, enabling the generation of novel visual content from textual descriptions. Similarly, the Codex API is tailored for the automation of coding tasks, exemplifying the versatility and depth of OpenAI's API suite in tackling varied challenges across AI domains.

These APIs excel in automating traditionally labor-intensive tasks, enhancing operational efficiency and freeing up human resources for higher-level functions. The precision and speed with which these tools operate exemplify their technical sophistication and potential to revolutionize business workflows.

Flexibility in integration. A significant advantage of OpenAI's APIs lies in their robust integration capabilities. They are designed to be seamlessly integrated into diverse platforms, including web applications, mobile apps, and cloud-based infrastructures [17]. This versatility is supported by compatibility with prevalent programming languages, which amplifies their accessibility and utility across the global development community.

Challenges and limitations. Despite their advantages, OpenAI's APIs are not devoid of challenges. Technical constraints such as latency issues and limited processing capacity can adversely affect the performance of applications reliant on real-time data processing. Operational hurdles like geographic

availability, cost implications for smaller entities, and scalability concerns further complicate their adoption. Additionally, the APIs must address functional consistency to maintain reliability across varied deployments [18].

Ethical considerations. The ethical landscape in which these APIs operate is complex, highlighted by issues surrounding bias, fairness, transparency, and accountability in automated decision-making. The utilization of extensive data sets, including customer data, for model training necessitates rigorous attention to data privacy and security. These concerns demand a meticulous and informed strategy for API deployment, emphasizing the need for ethical AI practices [19].

Navigating the limitations. To navigate these challenges effectively, developers and enterprises should adopt a strategic approach that includes choosing the appropriate API for specific needs, adhering to best practices in API integration, and continuous performance evaluation. Collaborative engagements with OpenAI for support and guidance, along with investments in AI-centric development tools, can enhance the capabilities and performance of these APIs.

Thus, OpenAI's APIs represent a significant advancement in the integration of AI technology, offering substantial opportunities for innovation across various sectors. However, the realization of their full potential is contingent upon overcoming the inherent limitations and addressing the ethical considerations associated with their use.

VECTOR DATA REPRESENTATIONS

Building upon the advanced capabilities of OpenAI's APIs, which facilitate seamless integration and utilization of AI in diverse applications, it is imperative to understand the underlying technologies that enable such sophisticated interactions. One of the foundational technologies is vector data representations, particularly vector embeddings, which are central to the functionality of AI and machine learning systems (Fig. 1).

Vector data representations, especially in the form of vector embeddings, are pivotal in the realm of AI and machine learning. These embeddings are numerical representations of data that encapsulate semantic relationships and similarities, thus enabling mathematical operations and comparisons on the data. This transformation is crucial for a multitude of tasks, including text analysis, recommendation systems, and beyond, essentially converting data into a format that can be efficiently processed and understood by AI models.

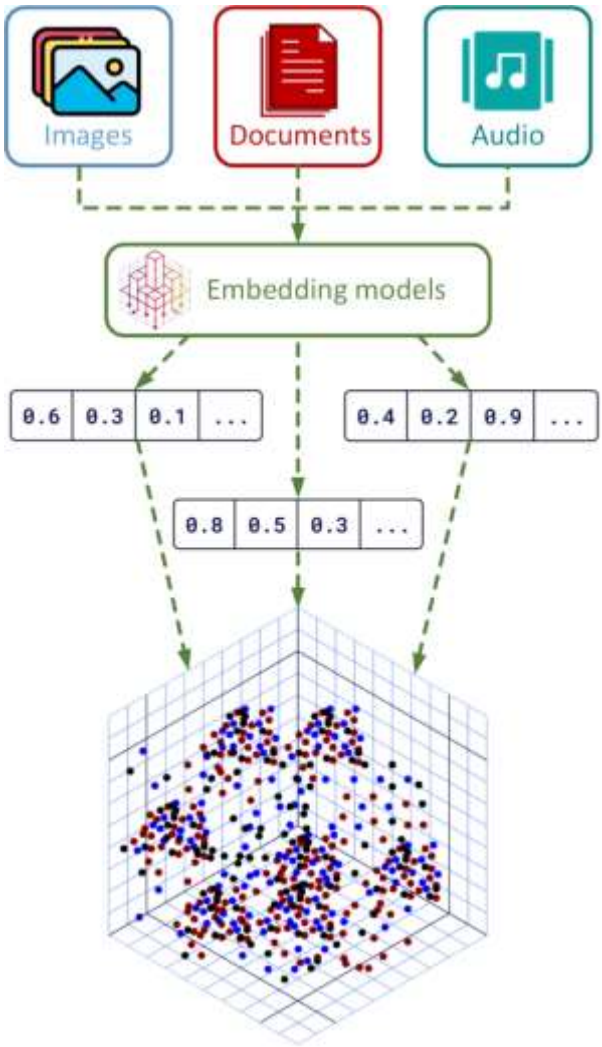

Fig. 1. **Vector embeddings**
Source: **compiled by the authors**

Vector embeddings function by assigning an n-dimensional vector to each data point. For instance, in NLP, models like Word2Vec represent words within a high-dimensional space – often extending into hundreds or thousands of dimensions [20]. This representation allows for the plotting of words in a manner that reflects semantic proximity; words with similar meanings are located closer together, illustrating the semantic relationships inherent in the data.

OpenAI's approach to vector embeddings exemplifies a sophisticated method in AI, translating complex textual data into interpretable, high-dimensional numerical formats. These embeddings are typically generated by advanced language models and are designed to capture the semantic nuances and contextual meanings of text. This enables AI models to perform tasks such as language understanding, text generation, and complex analytical tasks with notable precision and depth. The strategic use of vector embeddings by OpenAI is crucial in bridging the gap

between human language and machine interpretation, underpinning various AI-driven applications that demand an in-depth understanding and manipulation of data.

The applications of vector embeddings are extensive and varied. In recommendation systems, for example, items are represented as vectors within an n-dimensional space. By comparing these vectors, the system can identify and recommend items that are similar to one another, thereby enhancing user experience through personalized suggestions [21].

Search functionalities also leverage vector embeddings extensively. An example is Google's reverse image search, where images are transformed into vector representations that allow for efficient and accurate retrieval based on visual similarities [22]. This method applies not only to images but also to textual content, where search engines employ vector embeddings to improve the relevance and precision of search results.

Furthermore, vector embeddings are foundational in detecting anomalies or fraudulent activities. By analyzing the vectorized representation of transaction data, AI systems can detect patterns that deviate from the norm, thus identifying potential fraud. Similarly, in data processing and transformation, vector embeddings facilitate the mapping and interpretation of large datasets, enabling more effective data visualization and analysis [23].

Lastly, vector embeddings are integral to the operation of conversational interfaces, such as chatbots. By converting user input into vector form, these systems can better understand and respond to queries based on the semantic understanding facilitated by vector embeddings [24, 25].

VECTOR DATABASES

After exploring vector data representations, the need for effective database management systems becomes clear. Vector databases are such systems, designed specifically to handle and leverage the capabilities of vector embeddings in practical applications. These databases are essential for storing and processing the high-dimensional data typical of modern AI workflows, thereby enhancing the operational efficiency of LLMs [26, 27].

Vector databases are specialized storage systems that differ significantly from traditional relational databases (Fig. 2). They are optimized to store and manage vector embeddings, which are high-dimensional representations of data points, including text, images, audio, and video. These embeddings encapsulate the semantic relationships within the data, making vector databases particularly suited for tasks requiring deep semantic analysis.

Fig. 2. **Vector database input**
Source: **compiled by the authors**

Some of the well-known vector databases and libraries for similarity search and clustering are:

Milvus. An open-source vector database, Milvus is engineered for scalability and high-performance similarity searches. It supports multiple index types and can efficiently handle billions of vectors, making it an excellent choice for large-scale machine learning applications where rapid query response is critical [28].

Chroma. Designed to enhance LLM applications, Chroma focuses on making knowledge, facts, and skills easily accessible and integrable with LLMs. It streamlines the development of AI-driven applications by providing a robust infrastructure for embedding management and retrieval [29].

Facebook AI similarity search (FAISS). Developed by Facebook, Faiss is a library for efficient similarity search and clustering of dense vectors. It is particularly adept at managing large sets of vectors and supports several types of indices for fast retrieval, crucial for applications like image recognition and multimedia retrieval [30].

Weaviate. An open-source vector database that supports both object and vector storage, Weaviate allows for a hybrid approach combining vector search with traditional database operations like filtering. This flexibility makes it suitable for a range of applications, from AI-powered search engines to complex data analysis platforms [31].

Pinecone. A managed vector database service, Pinecone is designed to be user-friendly and highly scalable, making it ideal for businesses looking to enhance their data infrastructure with AI capabilities. Its architecture supports seamless integration with existing systems, facilitating advanced data analysis tasks [32].

Vespa. Both a search engine and vector database, Vespa is capable of performing approximate nearest neighbor searches, lexical searches, and structured data queries within the same framework. Its ability to incorporate machine-learned model inference in real time makes it uniquely positioned for

dynamic AI applications requiring immediate data processing [33].

Elasticsearch. Traditionally known for its full-text search capabilities, Elasticsearch has expanded to support vector similarity searches, integrating with third-party text embedding models. This enhancement allows it to serve a wider array of search-based applications, from e-commerce to academic research [34].

In practical scenarios, vector embeddings processed through these databases enable LLMs to perform complex tasks such as text summarization, translation, QA, and creative writing with high accuracy and efficiency. The embeddings provide a nuanced understanding of language by capturing the intricate relationships between words and phrases, thereby enabling LLMs to navigate the complexities of human language with remarkable proficiency.

Moreover, vector embeddings are instrumental in the continuous learning process of LLMs. They form the core of the model's learning architecture, allowing it to adapt and refine its responses based on new inputs and interactions. This ongoing learning process is vital for developing applications that evolve with their users, continually improving in accuracy and relevance [35, 36].

Thus, vector databases not only facilitate the efficient handling of high-dimensional data but also enhance the capabilities of LLMs, driving their ability to process and interact with human language in a contextually aware and meaningful manner. This integration of vector databases with AI applications underscores the transformative impact of advanced data management technologies in the realm of artificial intelligence.

LANGCHAIN: BRIDGING AI AND DATA

LangChain is a framework that uses vector databases and serves as a crucial bridge between LLMs and diverse data sources [37, 38] (Fig. 3).

LangChain stands as a comprehensive and versatile framework designed specifically to enhance the integration of LLMs like GPT-4 with various data formats, including PDFs. This framework distinguishes itself by facilitating not just the connection between AI models and diverse datasets but also by significantly augmenting the AI's capabilities in data analysis.

In the realm of NLP, LangChain offers an advanced and efficient mechanism for processing and managing language data. It particularly excels in integrating AI models with data sources like PDFs, employing sophisticated NLP techniques to interpret and respond to user queries. This enables the development of complex language-based applications that are robust and responsive [40].

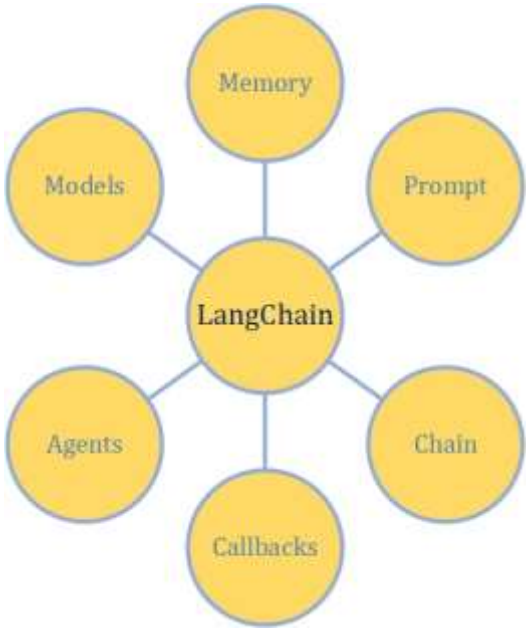

Fig. 3. **Modules in LangChain**
Source: **compiled by the [39]**

LangChain incorporates specialized document loaders, such as the PyPDF loader, which is optimized for the efficient parsing of PDF documents. This loader not only processes documents rapidly but also preserves essential metadata about the PDFs, enhancing the data's usability in AI applications.

The framework employs TextSplitters to segment long texts into smaller, semantically coherent chunks. This segmentation is vital for maintaining the context and coherence of the information being processed, ensuring that even lengthy texts are handled without losing meaning or continuity. By finely tuning parameters such as chunk size and overlap, LangChain can optimize the processing of complex documents, maintaining a seamless narrative flow.

A crucial component of LangChain is its use of embeddings and vector databases, such as Chroma, to transform text into vector representations within high-dimensional spaces. This transformation is key to performing semantic analysis and enabling functionalities like semantic search. The integration with vector databases enhances the framework's capability for efficient storage and rapid retrieval of these representations, thus bolstering analytical processes.

LangChain is also equipped with innovative features like query chains and QA chains. Query chains are designed as sequences of queries that methodically extract specific information from documents, whereas QA chains are structured to facilitate direct QA from texts. These features are particularly advantageous for applications in customer support, academic research, and data analysis, where precise and contextually relevant information retrieval is critical.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

The incorporation of RAG within LangChain marks a significant evolution in AI-driven data processing. RAG combines the strengths of dense vector retrieval with advanced generative models, enhancing the framework's ability to conduct sophisticated data querying and analysis. By utilizing external data sources, including comprehensive document corpora, RAG allows LangChain to augment the responses and insights generated by LLMs, delivering outputs that are not only context-rich but also highly accurate [41].

Retrieval augmented generation models (Fig. 4) represent a significant evolution in NLP by integrating a transformer-based generative component with a neural retrieval system. This hybrid approach enhances the capabilities of AI systems to deliver more precise and contextually appropriate responses [42].

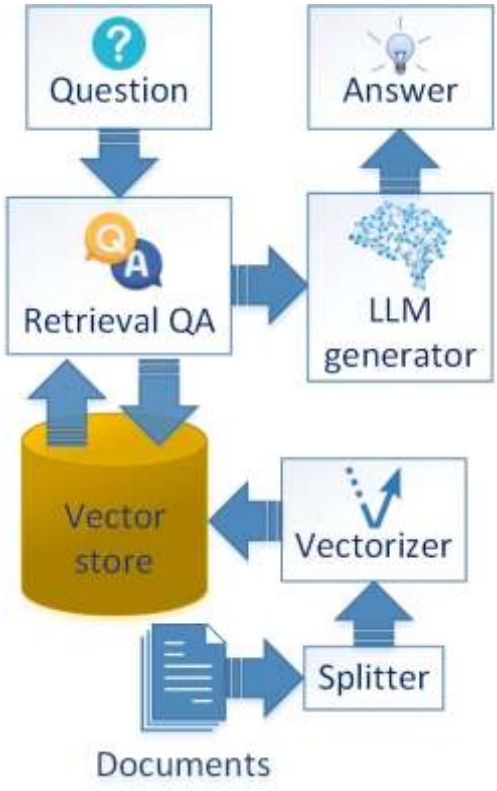

Fig. 4. **RAG workflow**
Source: **compiled by the authors**

TOKEN LIMITATIONS IN LARGE LANGUAGE MODELS

In summary, LangChain exemplifies a robust and dynamic framework that substantially enhances the utility of LLMs in analyzing and processing data from diverse sources. Its ability to integrate seamlessly with external databases and computational resources combined with its use of vectorized representations for precision and advanced components like models, prompts, chains, and agents, positions it as a formidable tool in the AI landscape. The potential applications of LangChain are vast and varied, encompassing everything from AI-driven email assistants to sophisticated data analysis tools and customer service chatbots.

Building on the discussion of LangChain's capabilities in bridging AI and diverse data sources, it is important to consider the inherent limitations of the underlying AI technologies, particularly those related to token constraints in models like OpenAI's GPT-3.5 and GPT-4. These limitations present significant challenges, especially when handling extensive datasets or long conversation histories, which are critical for applications requiring detailed documentation and user interaction.

OpenAI's GPT-3.5 and GPT-4 models have predefined token limits that pose challenges for maintaining context and coherence in extended interactions. GPT-3.5 is constrained to 4,000 tokens, while GPT-4 allows for 8,192 tokens, with a variant extending up to 32,768 tokens – approximately equivalent to 50 pages of text. Managing these token limits is crucial for ensuring smooth operation in applications like chatbots, customer service interfaces, or any system reliant on generating or processing extensive text [12].

To mitigate issues related to token constraints and enhance the continuity and context in interactions, several strategies can be implemented:

Limiting response length. One basic approach is to set a cap on the maximum length of responses by specifying the max_tokens parameter in the API call. This method ensures that the responses do not exceed the designated token count but does not address the accumulation of tokens due to extensive input history [43].

History truncation. Efficiently managing conversation history involves truncating older segments as newer interactions add to the token count. By calculating the expected token count for upcoming responses, the input history can be adjusted dynamically to ensure it remains within the model's processing capabilities without sacrificing necessary context [44].

Using summarization. Employing summarization techniques to condense the conversation history is an effective way to reduce token usage while retaining essential information. This summarized context can then be used as the basis for subsequent interactions, minimizing token expenditure and preserving the continuity of the dialogue [45].

Automating token management. Implementing automated systems to track and manage token usage can significantly streamline operations. For

instance, a script could monitor token counts for each interaction, storing this data for optimal management of conversation histories. This proactive approach helps maintain a balance between detailed interactions and token limitations [46].

From a scientific perspective, these strategies highlight the balance between computational efficiency and the quality of user interactions. Research into more advanced token management techniques, such as adaptive token allocation based on the complexity of the query or the criticality of the information requested, could further refine these interactions. Additionally, ongoing advancements in model efficiency and token economy are likely to alleviate some of these constraints, enabling more robust and seamless interactions even within token-limited environments.

Thus, while LangChain and similar frameworks provide powerful tools for integrating AI into diverse applications, the limitations of underlying technologies like token constraints in LLMs necessitate thoughtful management strategies to fully leverage these advanced capabilities in enhancing documentation practices and user interaction.

ADVANCED INTEGRATION TECHNIQUES FOR CONVERSATIONAL AGENTS

In the realm of artificial intelligence, agents refer to systems or software that act autonomously to perform tasks or achieve goals based on the environment they are interacting with [47]. Conversational agents, often implemented as chatbots or virtual assistants, are designed to simulate human-like interactions and provide users with information or assistance through text or voice communication. These agents are central components of platforms like LangChain.

The fundamental concept of agents involves utilizing a LLM to select a sequence of actions [48]. In chains, this sequence is predetermined and explicitly defined within the code. However, in agents, the sequence is dynamically determined by using the language model as a reasoning mechanism to decide the appropriate actions and their order (Fig. 5).

Conversational agents are crucial for enhancing user experience by providing dynamic and context-aware interactions. They interpret user queries and commands, generate relevant responses, and maintain an ongoing dialogue that mimics human conversation. The sophistication of these agents is not merely in understanding and responding to direct inputs but also in their ability to adapt their responses based on the context, user preferences, and interaction history [49].

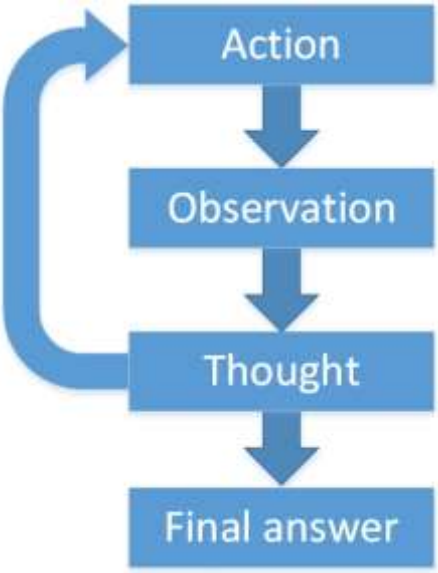

Fig. 5. **Agent scheme**
Source: **compiled by the authors**

Customization of conversational agents. Customization is a key strategy in the deployment of conversational agents, particularly on platforms like LangChain. This process involves tailoring the responses of language models based on specific user feedback, regional nuances, and industry-specific jargon. Through advanced machine learning techniques, these agents dynamically adjust their outputs to better align with the context or the specific requirements of the documentation process [50].

Dynamic response adjustment. Machine learning algorithms enable conversational agents to evolve their response mechanisms over time. This is achieved by continuously training the models on varied datasets that capture a wide range of user interactions. Techniques such as transfer learning and continuous adaptation [51] are employed to ensure that the models not only respond accurately but also reflect the evolving usage patterns and language specific to different user groups.

Contextual relevance. Ensuring that conversational agents maintain context and coherence throughout interactions is fundamental for user satisfaction. Advanced NLP techniques, such as context tracking and discourse analysis, are utilized to allow these agents to keep track of the conversation history. This capability ensures that responses are not only relevant but also contextually connected to previous interactions, thereby enhancing the natural flow of dialogue [52, 53].

Sentiment analysis. Integration with sentiment analysis tools allows agents to detect and adapt to the emotional tone of user inputs. This feature is particularly beneficial in customer service settings, where recognizing and responding to customer emotions

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

can significantly affect the outcome of interactions [54]. LLMs have demonstrated remarkable performance across a wide range of NLP tasks, including sentiment analysis. Leveraging LLMs, such as ChatGPT, for sentiment analysis allows for more accurate classification of emotions. For example, using ChatGPT to analyze student feedback about their teachers has shown immense promise. In a study, ChatGPT achieved an impressive overall F1-score of 88 % in classifying student feedback into positive, negative, or neutral sentiments, outperforming state-of-the-art deep learning and transformer-based models. These findings highlight the potential of LLMs to enhance sentiment analysis in various contexts, providing valuable insights for decision-making and improving user interactions [55].

Automated summarization. Integrating automated summarization capabilities allows conversational agents to quickly present dense information in a digestible format. This is especially useful in scenarios where users seek quick insights from detailed documents or lengthy discussions [56].

These integrations provide a more nuanced understanding of user queries and allow for personalized, context-aware responses. Sentiment analysis helps adjust communication tone, facial recognition enables personalized interactions based on user recognition, and automated summarization assists in condensing complex information into digestible summaries.

TECHNICAL CONSIDERATIONS FOR APPLICATION DEVELOPMENT

To develop a comprehensive application that leverages the capabilities of LangChain and ChatGPT for interacting with PDF documents, several technical considerations must be meticulously addressed [57].

Utilizing LangChain's data loaders and text splitters. The foundation of a successful application involves the efficient management and processing of PDF files. Developers should utilize LangChain's data loaders, which are specifically engineered to import PDF documents and transform them into formats amenable to processing by conversational AI models like ChatGPT. Furthermore, to handle extensive documents effectively, text splitters play a crucial role. These tools segment lengthy PDF content into smaller, coherent chunks, preserving the necessary context for ChatGPT to perform accurate and meaningful interactions. This segmentation ensures that the AI can manage and respond to the document's contents systematically, enhancing the user experience by maintaining continuity and relevance in the interactions [58].

Incorporating vector embeddings. The implementation of vector embeddings is vital for enriching the AI's comprehension of textual data within PDFs. By transforming text into high-dimensional vector spaces, these embeddings allow ChatGPT to discern and analyze the underlying semantic meaning of the text [59]. This capability enables the AI to respond to user queries with higher precision, tailoring its responses to reflect the content and nuances captured in the document's vector representations. This step not only improves the accuracy of the AI's outputs but also its ability to engage in more complex, content-specific discussions based on the user's queries.

Building a user-friendly interface. For optimal user engagement, the application must be designed with a clear, intuitive user interface that simplifies the process of uploading and interacting with PDF documents. Key features might include a drag-and-drop facility for easy file uploads a responsive chat interface for real-time interaction with ChatGPT, and dynamic display areas that visually highlight sections of the PDF relevant to the user's inquiries. These interface components should be designed with the end-user in mind, ensuring ease of use and enhancing the overall interactivity of the application [60].

Ensuring scalability and security. To accommodate a broad range of users and document types, scalability is essential. The application should be engineered to efficiently handle varying file sizes and different types of PDF documents, ensuring consistent performance across all user interactions. Security is equally critical, especially given the potential sensitivity of the data contained in user-uploaded PDFs. Robust security measures, including data encryption and secure data storage practices, must be implemented to protect against unauthorized access and ensure user data privacy [61].

CONCLUSIONS

The integration of ChatGPT with PDF documents using the LangChain infrastructure represents a significant advancement in NLP and information retrieval. This integration not only enhances interaction with PDF files but also extends to other text-containing data files, leveraging the combined strengths of OpenAI's GPT-4 and LangChain's framework. This technology enables sophisticated applications that facilitate dynamic and interactive engagements with PDF content, ranging from simple data retrieval to complex analysis and conversation-based information extraction.

Vector data representations, particularly vector

embeddings, play a crucial role in enhancing the functionality of artificial intelligence and machine learning systems. These embeddings transform complex textual data into high-dimensional numerical formats, enabling AI models to perform tasks such as text summarization, translation, and QA with increased precision and depth. Vector databases, optimized for handling these complex data representations, are critical in managing and leveraging high-dimensional data, boosting the operational efficiency of LLMs. This capability allows for more accurate and contextually relevant responses in various AI-driven applications.

LangChain excels in bridging LLMs with diverse data sources using vector databases. It incorporates innovative features like query chains and retrieval-augmented generation for sophisticated data analysis. The framework's use of specialized document loaders and text splitters ensures efficient data management and coherent segmentation of PDFs, facilitating accurate AI interaction. These elements underscore LangChain's versatility and effectiveness in integrating AI models with varied datasets, enabling the development of complex language-based applications that are robust and responsive.

Developing applications using LangChain and ChatGPT requires meticulous technical considerations. Efficient data management through LangChain’s data loaders and text splitters is essential. Implementing vector embeddings enriches the AI’s comprehension of textual data within PDFs, while user-friendly interfaces and robust security measures ensure optimal user engagement and data protection. These technical aspects are critical for building applications that perform well, provide a seamless user experience, and safeguard sensitive data.

LangChain also introduces agents, which add an additional layer of functionality and flexibility to AI applications. These agents dynamically determine the sequence of actions to take based on the input they receive, enhancing the AI's ability to interact with complex workflows and data sources. This dynamic action selection enables the development of more adaptive and intelligent systems, further expanding the practical applications of this technology.

The practical implications of this technology are profound, impacting various sectors including customer support, academia, and data analysis. In customer support, the ability to quickly access and relay information from PDFs can dramatically improve service quality and efficiency, potentially reducing resolution times by up to 40%. In academic settings, researchers can sift through extensive databases quickly, making the literature review process more manageable and reducing the time needed for comprehensive reviews by approximately 60%. For data analysis, automating data extraction from thousands of documents can significantly streamline workflows, boosting productivity and saving an estimated 50% of the time spent on manual data extraction [11, 12].

The integration of ChatGPT with PDF files through LangChain exemplifies the ongoing evolution of artificial intelligence. This technology sets a new standard for interactive AI applications, showcasing its potential to transform interactions with digital information. As AI continues to advance, the methodologies and technologies discussed here will pave the way for even more innovative and impactful applications, further enhancing efficiency, productivity, and user engagement across various sectors.

In conclusion, the detailed review and comprehensive analysis highlight the transformative impact of integrating ChatGPT with PDF documents using LangChain, demonstrating the potential to revolutionize interactions with digital information and setting the stage for future advancements in artificial intelligence.

REFERENCES

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., et al. “Language models are few-shot learners”. *arXiv preprint* arXiv:2005.14165. 2020. DOI: https://doi.org/10.48550/arXiv.2005.14165.

2. Open AI. “GPT-4 technical report”. *arXiv preprint* arXiv:2303.08774. 2023. DOI: https://doi.org/10.48550/arXiv.2303.08774.

Al-Janabi, O. M., Alyasiri, O. M. & Jebur, E. A. “GPT-4 versus Bard and Bing: LLMs for fake image detection”. *3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA).* Denpasar, Bali: Indonesia. 2023. p. 249–254, https://www.scopus.com/authid/detail.uri?authorId=57321859600. DOI: https://doi.org/10.1109/ICICyTA60173.2023.10429022.

3. Bitri, R. & Ali, M. “A comparative review of GPT-4’s applications in medicine and high decision making”. *International Conference on Computing, Networking, Telecommunications & Engineering Sciences Ap-*

plications (CoNTESA). Zagreb: Croatia. 2023. p. 61–68, https://www.scopus.com/authid/detail.uri?authorId=58627058100. DOI: https://doi.org/10.1109/CoNTESA61248.2023.10384948.

4. Chen, B. et al. "On the use of GPT-4 for creating goal models: an exploratory study". *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. Hannover: Germany. 2023. p. 262–271, https://www.scopus.com/authid/detail.uri?authorId=57220072911.
DOI: https://doi.org/10.1109/REW57809.2023.00052.

5. Lample, G. & Conneau, A. "Cross-lingual language model pretraining". *arXiv preprint* arXiv:1901.07291. 2019, https://www.scopus.com/authid/detail.uri?authorId=57156540200.
DOI: https://doi.org/10.48550/arXiv.1901.07291.

6. Ye, Q. & Doermann, D. "Text detection and recognition in imagery: a survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015; 37 (7): 1480–1500, https://www.scopus.com/authid/detail.uri?authorId=7003658285. DOI: https://doi.org/10.1109/TPAMI.2014.2366765.

7. Mikolov, T., Chen, K., Corrado, G. & Dean, J. "Efficient estimation of word representations in vector space". *arXiv preprint* arXiv: 1301.3781. 2013. DOI: https://doi.org/10.48550/arXiv.1301.3781.

8. Pita, M. & Pappa, G. L. "Strategies for short text representation in the word vector space". *7th Brazilian Conference on Intelligent Systems (BRACIS).* Sao Paulo: Brazil. 2018. p. 266–271, https://www.scopus.com/authid/detail.uri?authorId=55664324600.
DOI: https://doi.org/10.1109/BRACIS.2018.00053.

9. Asyrofi, R., Dewi, M. R., Lutfhi, M. I. & Wibowo, P. "Systematic literature review langchain proposed". *International Electronics Symposium (IES)*, Denpasar: Indonesia. 2023. p. 533–537, https://www.scopus.com/authid/detail.uri?authorId=57220087356.
DOI: https://doi.org/10.1109/IES59143.2023.10242497.

10. Wang, Y., Jiang, J., Zhang, M., Li, C., Liang, Y., Mei, Q. & Bendersky, M. "Automated evaluation of personalized text generation using large language models". *arXiv preprint* arXiv:2310.11593. 2023. DOI: https://doi.org/10.48550/arXiv.2310.11593.

11. Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B. & Abiodun, O. I. "A Comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity". *Information.* 2023; 14 (8): 462, https://www.scopus.com/authid/detail.uri?authorId=57190409845.
DOI: https://doi.org/10.3390/info14080462.

12. Dietterich, T. G. "Approximate statistical tests for comparing supervised classification learning algorithms". *Neural Computation.* 1998; 10 (7): 1895–1923. DOI: https://doi.org/10.1162/089976698300017197.

13. Hirway, C., Fallon, E., Connolly, P., Flanagan, K. & Yadav, D. "A comparative study of intent classification performance in truncated consumer communication using GPT-Neo and GPT-2". *International Conference on Emerging Techniques in Computational Intelligence (ICETCI).* Hyderabad: India. 2023. p. 97–104, https://www.scopus.com/authid/detail.uri?authorId=57193133745.
DOI: https://doi.org/10.1109/ICETCI58599.2023.10331337.
Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. "On the dangers of stochastic parrots: can language models be too big?" *Proceedings of FAccT.* 2021, https://www.scopus.com/authid/detail.uri?authorId=57222183688. DOI: https://doi.org/10.1145/3442188.3445922.

14. Jacobs, C. & Kamper, H. "Leveraging multilingual transfer for unsupervised semantic acoustic word embeddings". *IEEE Signal Processing Letters*, 2024; Vol. 31: 311–315, https://www.scopus.com/authid/detail.uri?authorId=57222742499. DOI: https://doi.org/10.1109/LSP.2023.3347154.

15. Wen, X. & Zheng, Y. "The application of artificial intelligence technology in cloud computing environment resources". *Journal of Web Engineering.* September 2021; 20 (6): 1853–1866, https://www.scopus.com/authid/detail.uri?authorId=57457805700. DOI: https://doi.org/10.13052/jwe1540-9589.2067.

16. Baziyad, M., Kamel, I. & Rabie, T. "On the linguistic limitations of ChatGPT: an experimental case study". *International Symposium on Networks. Computers and Communications (ISNCC).* Doha: Qatar. 2023. p. 1–6, https://www.scopus.com/authid/detail.uri?authorId=6602232897.
DOI: https://doi.org/10.1109/ISNCC58260.2023.10323661.

17. Ghandour, A., Woodford, B. J. & Abusaimeh, H. "Ethical considerations in the use of ChatGPT: an exploration through the lens of five moral dimensions". *IEEE Access.* 2024; 12: 60682–60693, https://www.scopus.com/authid/detail.uri?authorId=25521012800.

DOI: https://doi.org/10.1109/ACCESS.2024.3394243.

18. Hendrawan, I. R., Utami, E. & Hartanto, A. D. "Comparison of Word2vec and Doc2vec methods for text classification of product reviews". *6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE).* Yogyakarta: Indonesia. 2022. p. 530–534, https://www.scopus.com/authid/detail.uri?authorId=35796613800.
DOI: https://doi.org/10.1109/ICITISEE57756.2022.10057702.

19. Heimerl, F. & Gleicher, M. "Interactive analysis of word vector embeddings". *Computer Graphics Forum.* 2018; 37 (3): 253–265, https://www.scopus.com/authid/detail.uri?authorId=12781821100.
DOI: https://doi.org/10.1111/cgf.13417.

20. Bitirim, Y., Bitirim, S., Celik Ertugrul, D. & Toygar, O. "An evaluation of reverse image search performance of Google". *IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC).* Madrid: Spain. 2020. p. 1368–1372, https://www.scopus.com/authid/detail.uri?authorId=21742198800.
DOI: https://doi.org/10.1109/COMPSAC48688.2020.00-65.

21. Pande, A. & Ahuja, V. "WEAC: Word embeddings for anomaly classification from event logs". *IEEE International Conference on Big Data (Big Data).* Boston: MA, USA. 2017. p. 1095–1100, https://www.scopus.com/authid/detail.uri?authorId=43461160700.
DOI: https://doi.org/10.1109/BigData.2017.8258034.

22. Yager, K. G. "Domain-specific chatbots for science using embeddings". *Digital Discovery.* 2023; 2: 1850–1861, https://www.scopus.com/authid/detail.uri?authorId=8047068200.
DOI: https://doi.org/10.1039/D3DD00112A.

23. Hassani, H. & Silva, E. S. "The role of ChatGPT in data science: how AI-assisted conversational interfaces are revolutionizing the field". *Big Data and Cognitive Computing.* 2023; 7 (2): 62, https://www.scopus.com/authid/detail.uri?authorId=25521675700. DOI: https://doi.org/10.3390/bdcc7020062.

24. Singh, P. N., Talasila, S. & Banakar, S. V. "Analyzing embedding models for embedding vectors in vector databases". *IEEE International Conference on ICT in Business Industry & Government (ICTBIG).* Indore: India. 2023. p. 1–7, https://www.scopus.com/authid/detail.uri?authorId=57223019615.
DOI: https://doi.org/10.1109/ICTBIG59752.2023.10455990.
Taipalus, T. "Vector database management systems: fundamental concepts, use-cases, and current challenges". *Cognitive Systems Research*, 2024; 85, https://www.scopus.com/authid/detail.uri?authorId=57202507438. DOI: https://doi.org/10.1016/j.cogsys.2024.101216.

25. Wang, J. et al. "Milvus: a purpose-built vector data management system". *SIGMOD '21: Proceedings of the 2021 International Conference on Management of Data.* 2021. p. 2614–2627. DOI: https://doi.org/10.1145/3448016.3457550.

26. Janakiram, M. S. V. "Exploring Chroma: the open source vector database for LLMs". *The New Stack.* 2023. – Available from: https://thenewstack.io/exploring-chroma-the-open-source-vector-database-for-llms. – [Accessed: May, 2023].

27. George, G. and Rajan, R. "A FAISS-based search for story generation". *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India, 2022, p. 1-6, https://www.scopus.com/authid/detail.uri?authorId=57197984309. DOI: https://doi.org/10.1109/INDICON56171.2022.10039758.

28. "Welcome to Weaviate docs". – Available from: https://weaviate.io/developers/weaviate. – [Accessed: May, 2023].

29. "Pinecone docs". – Available from: https://docs.pinecone.io/guides/get-started/quickstart. – [Accessed: May, 2023].

30. "Vespa documentation". – Available from: https://docs.vespa.ai/. – [Accessed: May, 2023].

31. Guo, Q., Hu, J. & Liang, Z. "A scalable target indexing and retrieval system for massive video data processing based on Elasticsearch and Hadoop". *IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC).* Chongqing: China. 2024. p. 1805–1809, https://www.scopus.com/authid/detail.uri?authorId=57200090896.
DOI: https://doi.org/ 10.1109/IAEAC59436.2024.10503896.

32. Tissera, M. N. S., Asanka, P. P. G. D. & Rajapakse, R. A. C. P. "Enhancing customer segmentation using large language models (LLMs) and deterministic, independent-of-corpus embeddings (DICE)". *4th International Conference on Advanced Research in Computing (ICARC).* Belihuloya: Sri Lanka. 2024. p. 73–78, https://www.scopus.com/authid/detail.uri?authorId=59031553900.

DOI: https://doi.org/10.1109/ICARC61713.2024.10499784.

33. Bacon, G. & Menon, V. “Use of large language model embeddings to predict research topic suitability based on organizational capabilities”. *SoutheastCon*, Atlanta: GA, USA. 2024. p. 1376–1381, https://www.scopus.com/authid/detail.uri?authorId=59007900600.
DOI: https://doi.org/10.1109/SoutheastCon52093.2024.10500152.

34. Madhav, D., Nijai, S., Patel, U. & Champanerkar, K. “Question generation from PDF using LangChain”. *11th International Conference on Computing for Sustainable Global Development (INDIACom).* New Delhi: India. 2024. p. 218–222, https://www.scopus.com/authid/detail.uri?authorId=57735529200.
DOI: https://doi.org/10.23919/INDIACom61295.2024.10499105.

35. Duan, Z. “Application development exploration and practice based on LangChain+ChatGLM+Rasa”. *2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE).* Chengdu: China. 2023. p. 282–285, https://www.scopus.com/authid/detail.uri?authorId=58921225000.
DOI: https://doi.org/10.1109/CBASE60015.2023.10439133.

36. “Build chatbot webapp with LangChain”. – Available from: https://www.geeksforgeeks.org/build-chatbot-webapp-with-langchain. – [Accessed: May, 2023].

37. Pillai, M. & Thakur, P. “Developing a website to analyze and validate projects using LangChain and Streamlit". *2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT).* Bengaluru: India. 2024. p. 1493–1501, https://www.scopus.com/authid/detail.uri?authorId=57773449600.
DOI: https://doi.org/10.1109/IDCIoT59759.2024.10467765.

38. Topsakal, O. & Akinci, T. C. “Creating large language model applications utilizing LangChain: a primer on developing LLM apps fast”. *International Conference on Applied Engineering and Natural Sciences.* 2023; 1: 1050–1056, https://www.scopus.com/authid/detail.uri?authorId=9738690400.
DOI: https://doi.org/10.59287/icaens.1127.

39. Silva, L. & Barbosa., L. “Improving dense retrieval models with LLM augmented data for dataset search”. *Knowledge-Based Systems.* 2024; 294, https://www.scopus.com/authid/detail.uri?authorId=18036441700. DOI: https://doi.org/10.1016/j.knosys.2024.111740.

40. Mao, A. “Large language model settings: temperature, top p and max tokens”. – Available from: https://vectorshift.ai/blog/large-language-model-settings-temperature-top-p-and-max-tokens. – [Accessed: May, 2023].

41. “How to truncate chat history to a fixed token count in an LCEL+RunnableWithMessageHistory RAG chain”. – Available from: https://github.com/langchain-ai/langchain/discussions/21041. – [Accessed: May, 2023].

42. Dunn, C. “Infinite chat with history summarization”. – Available from: https://devblogs.microsoft.com/surface-duo/android-openai-chatgpt-18. – [Accessed: May, 2023].

43. Liu, X., Guo, C., Yao, B. & Sarikaya, R. “A self-learning framework for large-scale conversational AI systems". *IEEE Computational Intelligence Magazine*, May 2024; 19 (2): 34–48, https://www.scopus.com/authid/detail.uri?authorId=57219624307. DOI: https://doi.org/10.1109/MCI.2024.3363971.

44. Kowsher, Md., Panditi, R., Prottasha, N. J., Bhat, P., Bairagi, A. K., et al. “Token trails: navigating contextual depths in conversational AI with ChatLLM”. *arXiv preprint* arXiv:2404.02402. 2024. DOI: https://doi.org/10.48550/arXiv.2404.02402.

45. Wason, R., Arora, P., Arora, D., Kaur, J., Singh, S. P. & Hoda, M. N. “Appraising success of LLM-based dialogue agents”. *11th International Conference on Computing for Sustainable Global Development (INDIACom).* New Delhi: India. 2024. p. 1570–1573, https://www.scopus.com/authid/detail.uri?authorId=55965162900. DOI: https://doi.org/10.23919/INDIACom61295.2024.10498880.

46. De Vito, G., Lambiase, S., Palomba, F. & Ferrucci, F. “Meet C4SE: your new collaborator for software engineering tasks”. *49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).* Durres: Albania. 2023. p. 235–238, https://www.scopus.com/authid/detail.uri?authorId=57219252610. DOI: https://doi.org/10.1109/SEAA60479.2023.00044.

47. “Conversational agents get personal. How to take advantage of them”. *Artificial intelligence*. – Available from: https://www.tomorrow.bio/post/conversational-agents-get-personal-how-to-take-advantage-of-them-2023-09-5135067488-ai. – [Accessed: May, 2023].

48. Yao, S., Zhao, Y., Deng, Q., Ma, J. & Kang, Q. "Multi-objective neural architecture adaptation in transfer learning". *IEEE International Conference on Networking, Sensing and Control (ICNSC).* Shanghai: China. 2022. p.1–5, https://www.scopus.com/authid/detail.uri?authorId=57205426603.
DOI: https://doi.org/10.1109/ICNSC55942.2022.10004146.

49. Gao, J., Chen, J., Zhang, S., He, X. & Lin, S. "Recognizing biomedical named entities by integrating domain contextual relevance measurement and active learning". *IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).* Chengdu: China. 2019. p. 1495–1499, https://www.scopus.com/authid/detail.uri?authorId=57209472530.
DOI: https://doi.org/10.1109/ITNEC.2019.8728991.

50. Saikia, K. P., Mukherjee, D., Mahapatra, S., Nandy, P. & Das, R. "Unveiling deeper petrochemical insights: navigating contextual question answering with the power of semantic search and LLM fine-tuning". *International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).* Greater Noida: India. 2023. p. 881–886, https://www.scopus.com/authid/detail.uri?authorId=58916129000.
DOI: https://doi.org/10.1109/ICCCIS60361.2023.10425564.

51. Wan, B., Wu, P., Yeo, C. K. & Li, G. "Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies". *Information Processing & Management.* 2024; 61 (2), https://www.scopus.com/authid/detail.uri?authorId=58775685500.
DOI: https://doi.org/10.1016/j.ipm.2023.103609.

52. Shaikh, S., Daudpota, S. M., Yayilgan, S. Y. & Sindhu, S. "Exploring the potential of large-language models (LLMs) for student feedback sentiment analysis". *International Conference on Frontiers of Information Technology (FIT).* Islamabad: Pakistan. 2023. p. 214–219, https://www.scopus.com/authid/detail.uri?authorId=57210376659. DOI: https://doi.org/10.1109/FIT60620.2023.00047.

53. Nguyen, H., Chen, H., Maganti, R., Hossain, K. T. & Ding, J. "Measurement and identification of informative reviews for automated summarization". *IEEE International Conference on Artificial Intelligence Testing (AITest).* Athens: Greece. 2023. p. 146–151, https://www.scopus.com/authid/detail.uri?authorId=58109081300. DOI: https://doi.org/10.1109/AITest58265.2023.00031.

54. Elmeghni, Z. "GPT-4 and LangChain: building Python chatbot with PDF integration". – Available from: https://blog.nextideatech.com/chat-with-documents-using-langchain-gpt-4-python. – [Accessed: May, 2023].

55. Saeed, A., Dhanda, N., Rao, A. S. & Verma, R. "AI-enabled semantic web". *2nd International Conference on Disruptive Technologies (ICDT)* Greater Noida: India. 2024. p. 1136–1141, https://www.scopus.com/authid/detail.uri?authorId=57225817328.
DOI: https://doi.org/10.1109/ICDT61202.2024.10489028.

56. Maryamah, M., Irfani, M. M., Tri Raharjo, E. B., Rahmi, N. A., Ghani, M. & Raharjana, I. K. "Chatbots in academia: a retrieval-augmented generation approach for improved efficient information access". *16th International Conference on Knowledge and Smart Technology (KST).* Krabi: Thailand. 2024. p. 259–264, https://www.scopus.com/authid/detail.uri?authorId=59007149600.
DOI: https://doi.org/10.1109/KST61284.2024.10499652.

57. Dash, S. K. "Build a ChatGPT for PDFs with LangChain". – Available from: https://www.analyticsvidhya.com/blog/2023/05/build-a-chatgpt-for-pdfs-with-langchain. – [Accessed: May, 2023].

58. Wu, X., Duan, R. & Ni, J. "Unveiling security, privacy, and ethical concerns of ChatGPT". *Journal of Information and Intelligence.* 2024; Vol. 2, Issue 2: 102–115. DOI: https://doi.org/10.1016/j.jiixd.2023.10.007.

Conflicts of Interest: the authors declare no conflict of interest

Received 14.02.2024
Received after revision 16.04.2024
Accepted 13.05.2024

DOI: https://doi.org/10.15276/aait.07.2024.10
УДК 004.032.26:004.946

Ефективні практики документування для покращення взаємодії з користувачем за допомогою розмовних інтерфейсів на базі GPT

Шеремет Олексій Іванович[1)]
ORCID: https://orcid.org/0000-0003-1298-3617; sheremet-oleksii@ukr.net. Scopus Author ID: 57170410800
Садовой Олександр Валентинович[2)]
ORCID: https://orcid.org/0000-0001-9739-3661; sadovoyav@ukr.net. Scopus Author ID: 57205432765
Шеремет Катерина Сергіївна[1)]
ORCID: https://orcid.org/0000-0003-3783-5274; artks@ukr.net. Scopus Author ID: 57207768511
Сохіна Юлія Віталіївна[2)]
ORCID: https://orcid.org/0000-0002-4329-5182; jvsokhina@gmail.com. Scopus Author ID: 57205445522
[1)] Донбаська державна машинобудівна академія, бул. Машинобудівників, 39. Краматорськ, 84313, Україна
[2)] Дніпровський державний технічний університет, вул. Дніпробудівська, 2. Кам'янське, 51918, Україна

АНОТАЦІЯ

У статті представлено детальний огляд інтеграції ChatGPT з PDF-документами за допомогою інфраструктури LangChain, що підкреслює значні досягнення в обробці природної мови та пошуку інформації. Перевага цього підходу полягає в тому, що він не обмежується виключно роботою з PDF-документами. Використовуючи спеціальні можливості інфраструктури LangChain, можна взаємодіяти з будь-якими файлами даних, що містять текстову інформацію. Огляд літератури підкреслює трансформаційний вплив моделей серії GPT від OpenAI на обробку природної мови, при цьому прогрес, досягнутий у GPT-4 значно покращує генерацію тексту, схожого на написаний людиною, і встановлює нові стандарти для інтерактивних застосувань штучного інтелекту. Аналіз інтерфейсу прикладного програмування OpenAI демонструє його важливу роль у просуванні інтеграції штучного інтелекту в різні застосунки, надаючи доступні та надійні інструменти, які дозволяють розробникам і підприємствам легко інтегрувати складні функції штучного інтелекту. Незважаючи на свої переваги, ці інтерфейси стикаються з такими проблемами, як затримка, обмеження потужності обробки та етичні міркування, які вимагають стратегічного впровадження та постійної оцінки, щоб повністю використовувати їхній потенціал. У статті досліджується роль векторних представлень даних, зокрема векторних вбудовувань, у покращенні функціональності систем штучного інтелекту та машинного навчання. Ці вбудовування перетворюють складні текстові дані у багатовимірні числові формати, дозволяючи моделям штучного інтелекту виконувати такі завдання, як розуміння мови, генерація тексту та аналіз даних із високою точністю та глибиною. Векторні бази даних відіграють важливу роль в управлінні та використанні високовимірних даних, зокрема векторних вбудовувань, для підвищення операційної ефективності великих мовних моделей. Ці спеціалізовані системи зберігання оптимізовані для роботи зі складними представленнями даних, уможливлюючи розширені застосування, такі як узагальнення тексту, переклад і відповіді на запитання з високою точністю та розумінням контексту. LangChain надає універсальну структуру, яка поєднує великі мовні моделі та різноманітні джерела даних за допомогою векторних баз даних. Ця інтеграція розширює можливості штучного інтелекту в аналізі даних і обробці природної мови, створюючи складні застосування, які можуть ефективно інтерпретувати та відповідати на запити користувачів на різних наборах даних. Розробка комплексного застосунку з використанням LangChain і ChatGPT для взаємодії з документами PDF вимагає ретельного технічного розгляду. Ключові елементи включають ефективне керування даними за допомогою завантажувачів даних LangChain і текстових розділювачів, які перетворюють PDF-файли в керовані формати та забезпечують узгоджену сегментацію для точної взаємодії штучного інтелекту. Крім того, впровадження векторних вбудовувань покращує здатність штучного інтелекту сприймати й аналізувати текстові дані, а зручний інтерфейс і надійні заходи безпеки забезпечують оптимальне залучення користувачів і захист даних. Практичні наслідки цієї технології значні: потенційні покращення в підтримці клієнтів шляхом скорочення часу вирішення проблеми до 40 %, оптимізації оглядів академічної літератури приблизно на 60 % і підвищення продуктивності аналізу даних завдяки економії приблизно 50 % часу, витраченого на ручне вилучення даних.

Ключові слова: ChatGPT; LangChain; векторні вбудовування; аналіз даних; генерація з доповненою вибіркою

ABOUT THE AUTHORS

Oleksii I. Sheremet - Doctor of Engineering Sciences, Professor, Head of the Department of Electromechanical Systems of Automation and Electric Drive. Donbas State Engineering Academy, 39, Mashinobudivnykiv Blvd. Kramatorsk, Ukraine
ORCID: https://orcid.org/0000-0003-1298-3617; sheremet-oleksii@ukr.net. Scopus Author ID: 57170410800
***Research field*:** Machine learning and artificial intelligence in general technical problems and electromechanics; predicative analytics based on artificial intelligence technology

Шеремет Олексій Іванович - доктор технічних наук, професор, завідувач кафедри Електромеханічних систем автоматизації Донбаської державної машинобудівної академії, бул. Машинобудівників, 39. Краматорськ, Україна

Oleksandr V. Sadovoi - Doctor of Engineering Sciences, Professor, Department of Electrical Engineering and Electromechanics. Dniprovsky State Technical University, 2, Dniprobudivska Str. Kamyanske, Ukraine
ORCID: https://orcid.org/0000-0001-9739-3661; sadovoyav@ukr.net. Scopus Author ID: 57205432765
***Research field*:** Optimal control of electromechanical systems

Садовой Олександр Валентинович - доктор технічних наук, професор кафедри Електротехніки та електромеханіки Дніпровського державного технічного університету, вул. Дніпробудівська, 2. Кам'янське, Україна

Kateryna S. Sheremet - Laboratory Assistant, Department of Intelligent Decision Support Systems. Donbas State Engineering Academy, 39, Mashinobudivnykiv Blvd. Kramatorsk, Ukraine.
ORCID: https://orcid.org/0000-0003-3783-5274; artks@ukr.net. Scopus Author ID: 57207768511
***Research field*:** Machine learning; decision support systems

Шеремет Катерина Сергіївна - лаборант кафедри Інтелектуальних систем прийняття рішень Донбаської державної машинобудівної академії, бул. Машинобудівників, 39. Краматорськ, Україна

Yuliia V. Sokhina - PhD in Engineering Sciences, Associate Professor, Department of Electrical Engineering and Electromechanics. Dniprovsky State Technical University, 2, Dniprobudivska, Str. Kamyanske, Ukraine
ORCID: https://orcid.org/0000-0002-4329-5182; jvsokhina@gmail.com. Scopus Author ID: 57205445522
***Research field*:** Optimal control of electromechanical systems

Сохіна Юлія Віталіївна - кандидат технічних наук, доцент кафедри Електротехніки та електромеханіки Дніпровського державного технічного університету, вул. Дніпробудівська, 2. Кам'янське, Україна