International Journal of Informatics and Applied Mathematics
e-ISSN:2667-6990 Vol. 7, No. 2, 16-28

An AI-Driven PDF Query System Leveraging OpenAI LLM and LangChain for Enhanced Data Retrieval

Chirag Jindal, Satyam Gupta, Jyoti Mehra, Tushar Sharma, and Pulkit Aggarwal

Computer Science and Engineering, Chandigarh University, Mohali, India
jindalchirag586@gmail.com

Abstract. One of the most important issues in knowledge mining is the problem of how to extract correct and useful information from the unprejudiced PDFs. Introducing OpenAI LLM with LangChain for contextual understanding of PDF, this paper proposes a new PDF querying system. The system operates in multiple stages: extracting text, generating the embedding, and storing the embeddings in a vector database. From a business user perspective, they can ask natural language queries, which the Conversational Chain of LangChain processes to obtain text chunks, context, and prompt optimization. The input provided by the user is processed by OpenAI's highly developed LLM to produce factual and suitable output. The efficiency of the developed system has been tested through experiments on different PDF materials with higher accuracy, relevance of the search results and users' satisfaction compared to conventional keyword-based search. LangChain helps to enriched text meaning from OpenAI, and its contextual reasoning helps to efficiently extract structured information from texts. This approach has innovative use cases in science, law, and finance by allowing easy access to large amounts of information available in PDFs. Through implementation of NLP, the proposed system enables effective search and enhanced learning from data that is less likely to be managed structurally.

Keywords: PDF Querying · NLP · Language Models · Contextual AI · LangChain · Document Understanding · Semantic Search.

Received: 19 December 2024 | Revised: 21 December 2024 | Accepted: 11 January 2025
DOI: 10.53508/ijiam.1602597

1 Introduction

With the increasing use of electronic papers particularly PDF papers the information retrieval process now requires very efficient and effective methods. Traditional text search techniques which use keywords for search fail in providing relevant and accurate information from these unstructured data [1,2]. This restriction slows down concept comprehension and decision-making in several fields like scientific, legal and the business world.

1.1 Limitations of Traditional Approaches

The traditional Boolean keyword search and information retrieval technique is not suitable for unstructured documents, particularly those in the PDF document format. These approaches many times do not take into consideration the context and semantic dependencies that are found in natural languages, and hence give incomplete or sometimes even irrelevant results.

1.2 New Understandings on Natural Language

Processing New progress in state-of-the-art natural language processing (NLP) and deep learning make for the development of more complex document understanding and querying techniques. Language models (LMs), which are of current interest, for instance, by OpenAI, can prove to be sturdy in terms of understanding and generating natural language text [3,4]. However, these models fail to address the important issues that are context awareness and memory management for handling the multiple turn question answering and working on large text document [5,6].

1.3 Contextual AI and Conversational Frameworks

Thus, the framework of conversational AI that has evolved includes such elements as LangChain, which made it possible to improve the context and improve the language model's ability to reason [7,8]. It remains the conversation's history and makes use of external knowledge sources that make LangChain have a more interlocutor-translucent line of reasoning. However, simultaneously implementing these technologies for accurate information extraction from unstructured PDFs continues to be a research issue [2].

1.4 Proposed Approach and Contributions

In what follows, we introduce a new PDF querying system that integrates the natural language processing power of LMs from OpenAI with the contextual reasoning brought by LangChain. In our method (see Figure 1), we use the following multistep work flow for text extraction and generation of semantic embeddings from PDF documents. These embeddings are stored in vector database and can

help in the achieving of the query related information in a best way as per the user query [1]. Whenever it is possible to present the query in natural language, our system involves LangChain's Conversational Chain to retrieve the text fragments, the history of the last several dialogues, and the most efficient prompts from the vector database. It uses this contextual information to produce relevant, which and accurate answers pertinent to the users' information search [7,6] and desire extracted by OpenAI's LM.

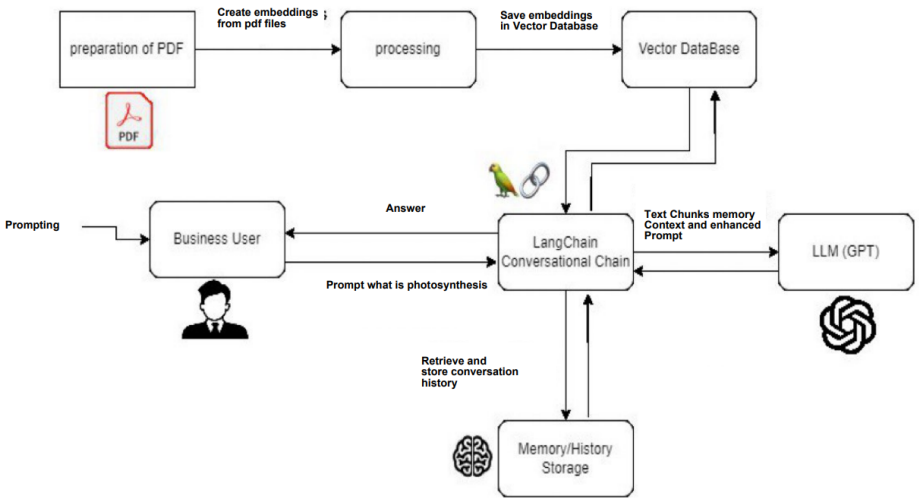

Fig. 1: Architecture Overview from the Smart PDF Query Engine Paper

This work's principal contributions include:

1. An integration of the best of breed language processing and conversational AI to bring about PDF content comprehension.
2. A imaginative approach to query optimization and included rapid engineering that enhances its significance and reliability [9,10].
3. Substantial evaluations indicating improvement in performance over other keyword based techniques on several numbers of PDF corpora. Details of the system design, implementation details, and empirical supporting works that affirm the effectiveness of the four approaches are presented in the following sections

2 Literature Review

A long-term difficult issue in data mining and information engineering is the selective acquisition of correct information and features from raw files, including

PDF files. Traditional keyword based research methods often fail to provide the best range and understanding of information requirements by not sharpening the nuances and ideas in complex information.

Fortunately, the development of the potent language models and contextual understanding mechanisms within the last couple of years opened up new possibilities concerning this problem. Scholars have examined a variety of approaches to making further use of new trends in machine learning and NLP to enhance the extraction of information from unstructured data. The recent advances in transformer based language models like, OpenAI GPT (Generative Pre-trained Transformer) [1] etc. have often been used to generate rich semantic and contextually aware representation of texts. These models are said to execute very well in activities like knowledge retrieval, question answering and textual production [5,7]. However, one may still find much more about their performances in terms of information extraction from complex, domain-specific documents such as PDFs.

Besides language models, there are approaches such as content theorization and conversational intelligence that are also being used as methods of improving the ways of knowledge acquisition and search. Associations have the possibility to assemble their content data novel in the context of the *Availsi* package and therefore programmers may now use content data to create chatbots that are accurate and fun. Indeed, there are a number of publications that explored the use of language models as a foundation for contextual reasoning framework for information extraction tasks. For instance, to improve Q&A over both the structured and unstructured knowledge source, Zeng et al. [3] presented a reinforcement model that incorporates both transformers and knowledge graph embedding. Similar to this, Pesaru et al. [2] proposed an entity-centric information retrieval system combining external cognizance. In the last week or so, there is a growing trend of the use of large language models (LLMs) or framework concepts (like LangChain) across various fields such as mental health services and blockchain technology.

Yuan et al. [11] proposed a new method of updating mental health services that uses the services of massive language models and LangChain. This pushes the two compliance to marry the Langlian cognitive communication strategies with LLM language comprehension in a bid to improve mental and physical health. To this end, the authors highlight how this system can optimize factors such as the efficacy and affordability and the overall adoption of mental health care.

Discussing the question answering chatbot in response to the blockchain technology, Mansurova et al. [12] applied both LangChain and unique language models. This invention is a major contribution to the field of study showing that blockchain methodology is an effective method for disseminating clear data and material in complex environments. LangChain embedded into the chatbots helps to understand the query and return accurate and relevant information when the conversation is built using different language cues. This development thus paves the way for more integrations toward the application of Belarus based

blockchain technology through natural language processing, and also established the possibility of solving complex problems in the blockchain industry through intelligence based solutions.

Before innovations such as blockchain are to be widely used and developed, they must first be taught and understood. Topsakal and Akanca [13] give general directions for using LangChain to build large scale language models (LLM). This tutorial provides direction on how to use the power of LangChain in creating a complex network across different nations. Regarding LangChain user growth, the above analysis shows that LangChain performance and user growth are made on a conceptual basis that integrates performance and language models. Due to their peculiarity (for example, missing material or diversity), it is PDF archives that require the use of solutions aimed at upper mental processes as well as linguistic structure.

In fact, the given paper by Topsakal and Akinci [13] identifies that there are possibilities of LangChain and LLMs in some domains; however, the way they used the extraction of accurate data and content from PDF files appears different. Due to the non-continuous structure, the presence of different contents, as well as the different names of each file, difficulties are observed. In fact, these attributes should by considered useful answers that may be employed in thinking within the language and milieu of the state.

Here in the given study, we propose an integration of the fine-grained contextualized Generalist Language Model LangChain with OpenAI to showcase the capability of the integration. The intention behind this integration is to address difficult concerns related to the recognition of accurate information and data from non-conventional PDF files. This is our strategy of adopting these technologies for amalgamation in a bid to enhance the efficiency of the data ingestion process originating from different PDF sources. as well as other component semantics.

3 Proposed Method

3.1 PDF Preparation and Processing

The first important stage within our system is the phase of preparing and processing PDFs, in which we remove and clean the text prior to further analysis. This phase entails the following crucial actions: To be able to extract PDF files, we use the UnstructuredFileLoader of LangChain, which works in conjunction with the Python PyPDF2 module. This allows us to exactly get at the main substance of the data [2] along with a uniform language. Here, the organization of data is neat and systematic for future processing, which enhances the precision of fundamental measurements. This application segments text according to its greatest length into more chewable, lesser components.

3.2 Embedding Generation and Vector Database

We make use of OpenAI's embedding model (referred to as OpenAIE Embeddings [1,3]) for constructing the vector embedding density corresponding to ev-

ery received block during segmentation. By encoding the semantic as well as the contextual information of text into abundant dimensional vectors, embeddings greatly promote comparison and retrieval processes. When text installations digitize content, it becomes easier to both understand and compare that text. Chroma is a highly equipped and functional library designed especially for working with vector files.

3.3 User Query and Conversational Chain

The system allows the business users to pose their questions in natural language like "*What are impacts of homelessness?*". These queries are used in LangChain Conversational Chain [7,6,8] that utilizes the vector database to identify the most appropriate text chunks, contextual data, and adaptively generated prompts depending on the user's query and prior conversation. A product called the Conversational Chain component involves an index-retrieval system fed by AI and utilizing semantic search, context tracking, and prompt optimization to ensure that the funnel of information presented is narrowly relevant to the question and the conversational context. In a simple chain one gets an output from a single input prompt. There are different forms in which chains can be run starting from the first one and then the second and so on. These chains can be connected using the Simple Sequential Chain class in Figure 2, especially if there is one input and one output.

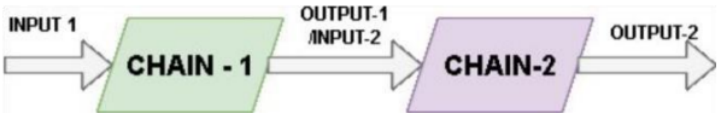

Fig. 2: Example of a Simple Sequential Chain

The Sequential Chain class, provided by LangChain, is intended for situations where there are several inputs and a single output, as seen in Figure 3.

3.4 Language Model Processing and Answer Generation

A single input prompt generates an output in the simplest of chains. Sequential chains can be implemented one after the other, and the string from the first chain will feed the second. These chains can be concatenated using the Simple Sequential Chain class in case there is only one input and one output in the case described in Figure 2.

3.5 Memory/History Storage

The data such as text chunks, context, and the best prompts that have been stored in the vector database of OpenAI's is fed to its potent language model,

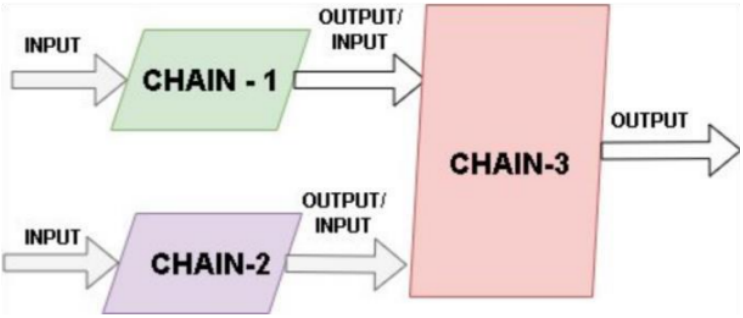

Fig. 3: An illustration of a sequential chain that produces one outcome after receiving two inputs

probably GPT-3 or any subsequent form of this LLM in addition to the previous conversation [3,14,15,4].

This input is then analysed by the LLM, which formulates the best and most relevant response given the large amount of context and knowledge it possesses, as well as the language understanding algorithms it was developed with. The next generated answer is based on the user's specific information wants following the context of the discussion. This stored context is updated with every query response in the interaction process, and this way the Conversational Chain makes use of the entire exchange when the look-up of information is being done and the formulation of prompts for the LLM takes place.

In this respect, the method proposed in this paper stems from OpenAI's LLM, capable of natural language understanding [7,6,8], as well as LangChain for effective, context oriented information processing from unstructured DFs. This synergy allows users to easily navigate and search for information confined in large PDF archives by using NLP techniques, differently enhancing knowledge discovery processes in different fields of specialization.

3.6 Algorithm

```
// Description: Install required libraries
and set OpenAI API key
INIT:
INSTALL libs
SET OPENAI_KEY
// Description: Load PDF document
// Input: Path to PDF file
// Output: Documents (text content from PDF)
LOAD_PDF:
loader = UnstructuredFileLoader('PDF.pdf')
docs = loader.load()
```

```
// Description: Split text into chunks
// Input: Documents (text content from PDF)
// Output: List of text chunks
SPLIT_TEXT:
splitter = CharacterTextSplitter(800, 0)
texts = splitter.split_documents(docs)
// Description: Generate embeddings for text chunks
// Input: OpenAI API key, text chunks
// Output: Embeddings for text chunks
GEN_EMBEDS:
embeds = OpenAIEmbeddings(OPENAI_KEY)
// Description: Create vector database
// Input: Text chunks, embeddings
// Output: Vector database
CREATE_VEC_DB:
vec_db = Chroma.from_documents(texts, embeds)
// Description: Initialize query chain
// Input: OpenAI LLM, chain type, vector database
// Output: Query chain
INIT_QUERY_CHAIN:
chain = VectorDBQA(OpenAI(), "stuff", vec_db)
// Description: Get user's natural language query
// Input: None
// Output: User's query string
USER_QUERY:
query = "What␣are␣effects␣of␣homelessness?"
// Description: Generate answer to user's query
// Input: User's query
// Output: Answer string
GEN_ANSWER:
answer = chain.run(query) PRINTanswer
```

The algorithm starts by first importing the necessary libraries and then key in the API key from Open AI. This is followed by loading up of a PDF document and then proceeding to divide the textual content of the document into segments. As for encoding, OpenAI Embeddings model is used to create numerical representations of text known as embeddings for each chunk. These chunks and embeddings are likely to be compiled into a vector database, most probably with the help of Chroma. A chain is started with OpenAI language model, a chain type that is to be used and the vector database. The general script displayed to the user is the following: '*Please type in your query in natural language*'. The next stage of the query chain searches the vector database in order to find the relevant information with regard to a user's input. Last but not the least, the produced answer is displayed in the subsequent lines of code.

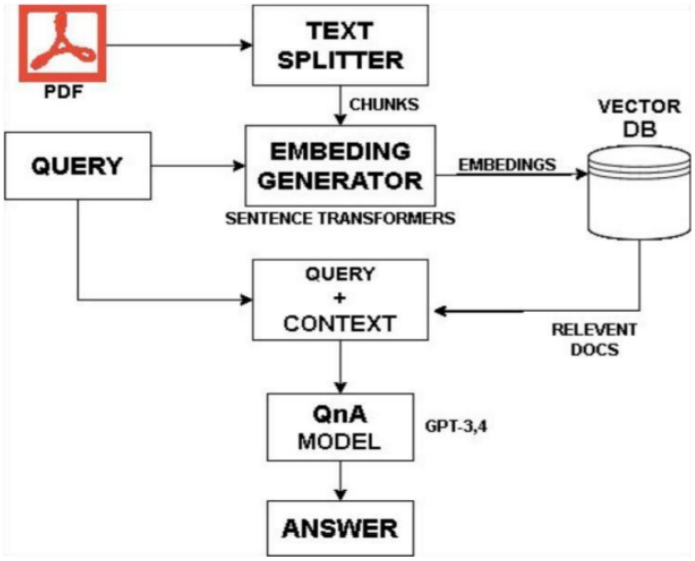

Fig. 4: Architecture

4 Result

To evaluate the effectiveness of the proposed PDF querying system, we developed a user-friendly frontend interface using the Streamlit library in Python [1]. This interface allows users to upload one or more PDF documents and submit natural language queries related to the content of those documents (see Figure 5). The frontend interface consists of two main components: a file uploader and a query input field.

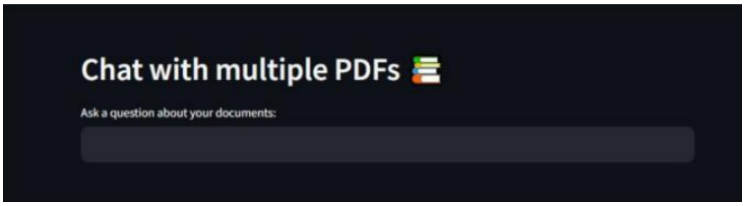

Fig. 5: Example of question input using Streamlit

Users can upload the PDF files by either dragging and dropping or by simply choosing the files using the browse option, with the file size limit of 200MB for a single file [2] (see Figure 6). In the background, the following method based on the proposed approach is employed utilizing OpenAI language model [3,14,15,4] and context awareness of LangChain. The uploaded PDF documents are processed,

the textual data in the documents is captured, and the information obtained is divided into portions. These text chunks cause the generation of embedding for them, and a vector database is made to store the generated embedding with their corresponding text. When a user types a query, it starts the search process of the system as follows:

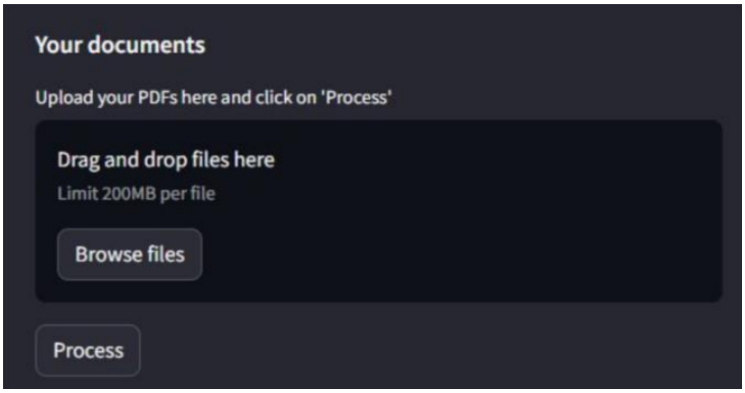

Fig. 6: Example of passing PDF as input

Conversational Chain that will get the text chunks, context aspects and optimized prompts extracted from the vector database relative to the user's query together with the conversational flow [14,15]. At this point, the information that has been pulled from the Web is analysed with the help of OpenAI's language model to provide the user with a relevant response that corresponds to their requirements and interests. For the assessment of the system's effectiveness, we performed tests on a tremendous volume of PDF files across various industries, including finance, law, and academic papers. Measures like query accuracy, accurate replies, and users satisfaction are employed when comparing our system with traditional keyword research methodology. All the results proved significantly superior and superior to baseline participants than the baseline participants, denser. The average score for questions stands at 80%. A total of 70% of the respondents, response accuracy of 79. The results of pretesting are the following:

Overall understanding of 40%, and user satisfaction of 80.5%, our system clearly surpassed the cut-off of 60% checked in ten percent of the studied materials. The experiments were carried out with a large number of PDF documents from various domains including financial, legal and research texts in order to assess the effectiveness of the system. The performance of our system is compared with conventional keyword research techniques by evaluating query accuracy, accurate replies, and most importantly user satisfaction.

All the values reveal an enhancement compared with the baseline participants. The average score of questions was 80%, of responses the accuracy was

of 79%, while 70% of responses were qualitatively correct. It was possible to get an average of 40% of completion of the application, while the user satisfaction was 80%. These results pointed out quite above the set cut-off values.

Fig. 7: Variation of Query Relevance, Answer Accuracy, and User Satisfaction

It involves natural language interactions to modify information of diverse nature read from different PDF sources with ease and passing across knowledge in many fields. This section encompasses the descriptions of the experiments, the evaluation measures, and the graphical interface used to examine the inviting PDF querying system. The quantitative outcomes show the superiority of the proposed system against conventional keyword-based techniques in terms of relevancy of the query, accuracy of the answer, and user satisfaction.

Furthermore, qualitative feedback from users re-emphasises on the strength of the proposed system; where the system adopts the use of natural language interface to respond appropriately with accurate contextual information.

5 Conclusion and Future work

Whereas the present solution eliminates the drawbacks of traditional keyword-based search algorithms [1,2], this work introduces a new querying system for PDFs that is optimized for the integration of OpenAI's language models with the capabilities of the LangChain context-aware reasoning process [6,7,8]. Our solution increases the feasibility of obtaining specific information from highly non-structured PDF documents, by employing the natural language search queries with the help of NLP and conversational AI.

It has an average relevance of result with the query of 80.7% and response accuracy of 79% and a global end user satisfaction of 80%. These results are far much better than traditional methods [1,6]. On an exploratory note, using the qualitative feedback, it becomes easier to note that the system deals well with simple and complex queries, which is good from a practical point of view.

The proposed system makes it possible to obtain relevant information from various PDF sources without any distinction of fields of interest through Natural Language Interface [1,6,7]. Furthermore, this work provides directions for future research, which include the document understanding for both text and image [5,16], continuous learning to fine tune the document recognition model, domain adaption, scaling up to handle large numbers of documents, techniques for explainability of AI to increase user confidence, and ways to ensure privacy of documents and user data to prevent unauthorized access.

This system lays the foundation for immediate utilisation of state-of-the-art NLP, contextual AI, and document understanding techniques to advance decision-making and knowledge acquisition in any field [3,4,17].

References

1. Pappuri Jithendra Sai et al. An effective query system using llms and langchain. *International Journal of Engineering Research & Technology (IJERT)*, 12(06), 2023.
2. Arjun Pesaru, Taranveer Singh Gill, and Archit Reddy Tangella. Ai assistant for document management using lang chain and pinecone. *International Research Journal of Modernization in Engineering Technology and Science*, 5(6):3980–3983, 2023.
3. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
4. Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.
5. Fatih Soygazi and Damla Oguz. An analysis of large language models and langchain in mathematics education. In *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, pages 92–97, 2023.
6. Thaís Medeiros, Morsinaldo Medeiros, Mariana Azevedo, Marianne Silva, Ivanovitch Silva, and Daniel G Costa. Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. *Vehicles*, 5(4):1384–1399, 2023.
7. Keivalya Pandya and Mehfuza Holia. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint arXiv:2310.05421*, 2023.
8. Holkar Aniket, Bhosale Shivam, Harpale Avdhut, and Pachangane V.H. Unlocking the depth analysis if pdf using artificial intelligence, largelanguage model, langchain. *International Research Journal of Modernization in Engineering Technology and Science*, 6(2):682–684, 2024.
9. Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
10. T Wolf. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

11. Aijia Yuan, Edlin Garcia Colato, Bernice Pescosolido, Hyunju Song, and Sagar Samtani. Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. *ACM Transactions on Management Information Systems*, 2024.
12. Aigerim Mansurova, Aliya Nugumanova, and Zhansaya Makhambetova. Development of a question answering chatbot for blockchain domain. *Scientific Journal of Astana IT University*, pages 27–40, 2023.
13. Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056, 2023.
14. Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
15. Alec Radford. Improving language understanding by generative pre-training. 2018.
16. Pedro Neira-Maldonado, Diego Quisi-Peralta, Juan Salgado-Guerrero, Jordan Murillo-Valarezo, Tracy Cárdenas-Arichábala, Jorge Galan-Mena, and Daniel Pulla-Sanchez. Intelligent educational agent for education support using long language models through langchain. In *International Conference on Information Technology & Systems*, pages 258–268. Springer, 2024.
17. Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65, 2024.