M.J. Hornos et al. (Eds.)

doi:10.3233/AISE240018

Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications

Matjaž GAMS [a,1], Maj SMERKOL [a], Primož KOCUVAN [a] and Matic ZADOBOVŠEK [b]
[a] *Jožef Stefan Institute, Slovenia*
[b] *University of Ljubljana, Slovenia*
ORCiD ID: Matjaž Gams https://orcid.org/0000-0002-5747-0711, Maj Smerkol https://orcid.org/0000-0003-4789-2395

Abstract. In this study, we explore the integration of ChatGPT with the Insieme platform, a robust electronic and mobile health system designed as an Italian and Slovenian project. This integration provides a novel way in which users access medical information, offering online support from healthcare professionals and enabling interactions with a sophisticated virtual assistant that utilizes cutting-edge natural language processing technologies. Our paper delves into the specific features of the Insieme platform, presenting a comprehensive explanation of the virtual assistant's implementation. The incorporation of ChatGPT into this medical platform introduces new solutions and challenges stemming from integrating a chatbot and an integral medical platform, potentially transforming the landscape of the Slovenian healthcare system. Furthermore, we examine the broader implications of this technology in enhancing patient care and optimizing healthcare workflows. Our working prototype provides perspectives on the evolution and future prospects of digital health solutions.

Keywords. virtual assistants, vector databases, word embeddings, GPT-4, natural language processing

1. Introduction

According to the WHO report Health and care workforce in Europe: time to act [1], all countries in the European region are facing challenges due to health and care workforce ageing. Many countries are already facing shortages of medical workers. While technology generally cannot replace medical professionals directly, it can help reduce their workload. This work investigates the possibilities of using an LLM integrated with a dedicated e-health platform to inform patients about healthcare related topics.

One of the key features of ChatGPT is its ability to understand and process complex queries in natural language, thus making it more intuitive for users without medical train-

[1]Corresponding Author: Matjaž Gams, matjaz.gams@ijs.si.

ing. This is particularly important in areas where access to healthcare professionals is limited. By providing immediate, AI-driven responses, ChatGPT can effectively bridge the gap in primary healthcare information. Continuing from the initial integrations of ChatGPT with medical knowledge, we delve deeper into the specific applications and benefits of this technology within the Insieme integral medical platform.

Furthermore, we explore the use of ChatGPT in patient education and health literacy improvement. By offering personalized, easy-to-understand explanations of medical conditions, treatments, and health tips, the platform may play a significant role in empowering patients to take charge of their health, providing medical information on the national services. This aspect is crucial in preventive medicine, where informed patients are more likely to engage in health-promoting activities and adhere to treatment plans.

Another aspect covered in this paper is the potential of ChatGPT in assisting healthcare professionals. The AI can serve as a support tool for doctors and nurses, providing quick access to medical literature, drug information, and case studies, thus enhancing the quality of care provided to patients.

Finally, we present the results of preliminary studies conducted to assess the effectiveness of ChatGPT in the Insieme platform. These studies focus on user satisfaction, accuracy of information provided, and the impact on healthcare outcomes. The results indicate a promising future for AI in healthcare, with potential applications extending beyond the current scope of the Insieme platform. We conclude by discussing future developments, including the integration of more advanced AI capabilities and expanding the reach of the platform to more users and healthcare settings. This exploration sets the stage for a new era in digital health, where AI becomes a fundamental component in delivering patient-centered, efficient, and accessible healthcare national services.

2. ChatGPT in Medicine

Our experiment started with several tests with open-source AI chatbots (for example: Bot for waiting queues and JSI assistant both were developed at JSI) integrated with the Insieme platform, but unfortunately they did not provide the desired quality of performance. Comparisons even with the default ChatGPT-4 without additional knowledge from the Insieme platform did not show any advantage. The next step was to integrate the large language model GPT-4 with the knowledge of our Insieme platform. The main reason for the project were reports that when compared to humans, even the default GPT-4 version generated more elaborated responses in terms of the quality of the answers and empathy [2], as demonstrated in Figure 1. Consequently, we decided that our virtual assistant would be based on the GPT-4 model.

GPT-4 was developed in March 2023 and represents a significant advancement in the field of natural language processing, particularly useful for answering questions, generating texts, and translating into other languages. Compared to previous generation models, the reliability and accuracy of responses have increased, and there is improved management according to user commands (for example, specifying the style of the generated response). Numerous tests [3] (various exams and knowledge tests from different fields) have shown that GPT-4 achieves results that are often comparable to those achieved by humans. It also performed well in several medical tasks [4–8].

While several other AI applications such as question answering methods provided reasonable information [9], GPT-4 seemingly outperforms competition particular in its generality and generativity [8, 10, 11].

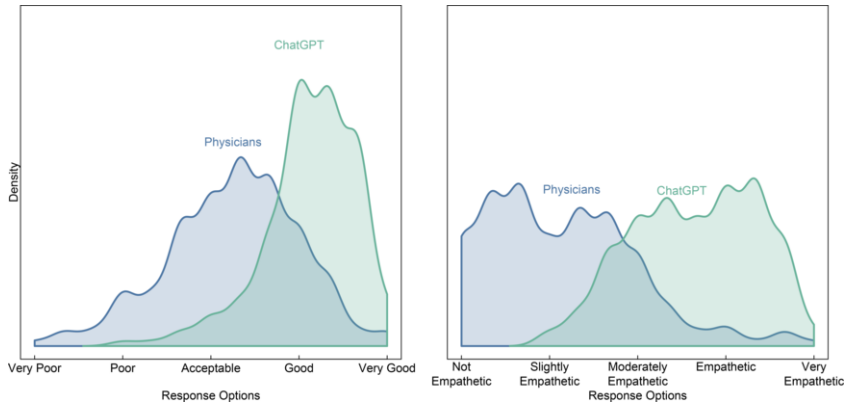

Figure 1. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine, 183, April 2023.

AI vs MD [2]: ChatGPT Outperforms Physicians in Providing High-Quality, Empathetic Healthcare Advice - SciTechDaily: This article reports on a study that compared the quality and empathy of the responses of ChatGPT and physicians to real-world health questions. The study found that healthcare professionals preferred AI responses to those of physicians 79 % of the time, citing higher quality and empathy.

Comparison of GPT-3.5, GPT-4, and human user performance on a clinical decision support system - Nature [12]: This article reports on a study that compared the performance of GPT-3.5, GPT-4, and human users on a clinical decision support system (CDSS) that provides diagnosis and treatment recommendations for eye diseases. The study found that GPT-4 achieved the highest accuracy and efficiency, followed by GPT-3.5 and human users.

Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions - UC San Diego: This article reports on a study that compared the quality and empathy of the responses of ChatGPT and physicians to patient questions. The study found that ChatGPT responses were rated significantly higher in quality than physician responses: good or very good quality responses were 3.6 times higher for ChatGPT than physicians (physicians 22.1 % versus ChatGPT 78.5 %).

The conclusion drawn from these studies suggests that we are progressing towards the development and implementation of artificial intelligence systems equipped with soft skills and empathy, similar to those of humans. This advancement shows promise in enhancing the support provided to medical personnel.

Instead of using the default GPT model or modifying the existing GPT model, we enabled the user to inquire about data obtained from our Insieme platform and other documents that can be supplied in any number. In this way, we separate the language model and the knowledge base, allow the user to communicate with the given documents, and use only information found within the supplied documents to generate the answer, ensuring the most relevant response for the user. With this approach, we can easily add new

sources of information and adapt the model for specific tasks; without training the existing model, which would otherwise be time-consuming and computationally demanding.

3. Insieme

For our test medical platform, we chose Insieme Figure 2., which was recently developed in collaboration with Slovenian and Italian partners as part of the cross-border ISE-EMH project [13]. It is equipped with a user-friendly interface that allows users to easily and elegantly access useful healthcare information from a single website. The core idea is that it provides similar information as "Dr. Google", but concentrated on the needs and possibilities of the local population. It is considered that an average user would find most of the relevant medical basic information in a couple of minutes, including applications, services, institutions and alike.

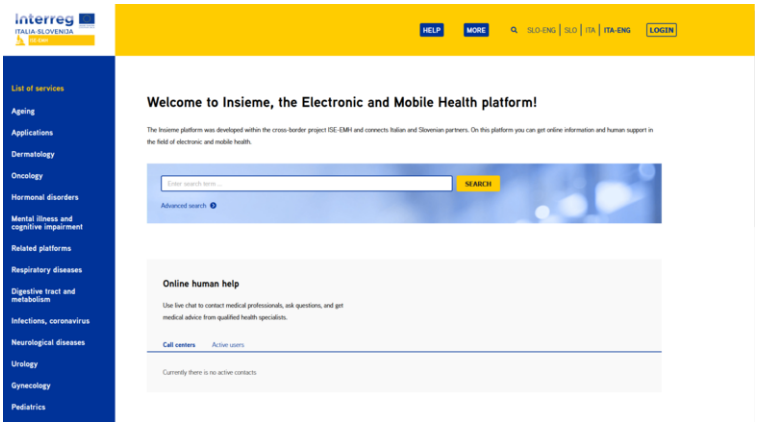

Figure 2. Insieme platform - https//www.ise-emh.com.

The main functionalities include the ability to search for services using a side menu bar or search function, online human assistance (live chat with a call center or healthcare expert), viewing health-related video content, and using a virtual assistant, which is presented as the central theme in the remainder of the article. All this content is available to users in three or four languages.

On the left side of the main menu is a list of services offered by the Insieme platform. This menu bar allows selection among different branches of medicine, after which the choice expands to diseases and medical conditions associated with the chosen branch of medicine, and information services related to the selected medical specialization are also displayed. By clicking on one of the medical conditions, e.g. oncology, the user is redirected to the corresponding subpage. There or a step further, essential basic information about the course of the disease, symptoms, possible prevention, and further action is available. Below, there are more links to external websites, enabling the user to acquire appropriate knowledge about the chosen disease. While that information can be found on the web, it might be even medically misleading in contrast to the information gathered in the platform, carefully evaluated by medical experts.

Besides manually navigating between the subpages of the platform, users also have a search functionality that displays all services on the platform matching the search string.

Online human assistance is also available to users. On the entrance page, there are lists of call centers and active doctors that can be contacted via the live web chat integrated into the platform.

The Insieme platform offers several built-in assistants: queue assistants, IJS assistants, service search, virtual assistant for medicine, and links to other non-integrated assistants.

4. GPT-4 Insieme-Enriched Medical Assistant

4.1. Background

The preexisting virtual assistant on the Insieme platform was designed to answer health-related questions. With the appearance of GPT-4, we enhanced the existing assistants with a ChatGPT-type assistant (or inversely, enriched GPT-4 with the Insieme platform). This assistant possesses a vast amount of its general GPT-4 knowledge from the web, as well as additional local information related to the Insieme project.

The first issue in using the assistant is the large amount of information to be included (perhaps books or even videos). Large language models typically have a limitation on how much text they can accept [14]. Therefore, it is crucial to provide only essential information to the large language model. Key to this are word embeddings and vector databases.

4.2. Word Embeddings and Vector Databases

Embeddings are a way to represent words, sentences, or even entire documents. To calculate them, we need appropriate models that have been trained on a huge amount of data and can find relationships between words by analyzing patterns in the data [15]. In our case, we used a model offered by OpenAI — text-embedding-ada-002. By obtaining a vector for each word, we can represent the meaning of the text. Word embeddings can be represented in multidimensional spaces, where words or sentences with similar meanings are close to each other — we can calculate distances between vectors to find semantically related words.

Vector databases store information in the form of vectors, often referred to as word (vector) embeddings. This allows us to index and search through a huge amount of unstructured data, such as images, raw text, or sensor data. Vector databases organize data using high-dimensional vectors, each dimension describing a specific characteristic of the data object it represents. Vector databases differ from traditional databases that store data in tabular form in that they return results based on similarity (traditional databases return exactly matching objects) [16]. Various measures, such as cosine similarity, are used to measure similarity between vectors in vector space. These measures allow us to compare vectors stored in our vector database and find those most similar to the user's input vector. They thus enable work with complex data and fast searching, which would otherwise cause difficulties for traditional databases. Suppose we have a document that we would like to index. We will use a model that enables the creation of word embed-

dings (we mentioned text-embedding-ada-002 above) [17]. We store them in the selected vector database, and a reference to the document from which the embedding was created is saved. Whenever our user sends a query, the same model is used to create embeddings — we use them to find the most similar word embeddings in the vector database, which are linked to the original document where they were created due to the mentioned reference. The obtained documents can then be provided to the large language model — the documents will be used as context for generating a response. Because of all these features, vector databases are an excellent choice to enrich our generative models.

4.3. Implementation

The overall schema of the system implementation is presented in Figure 3.

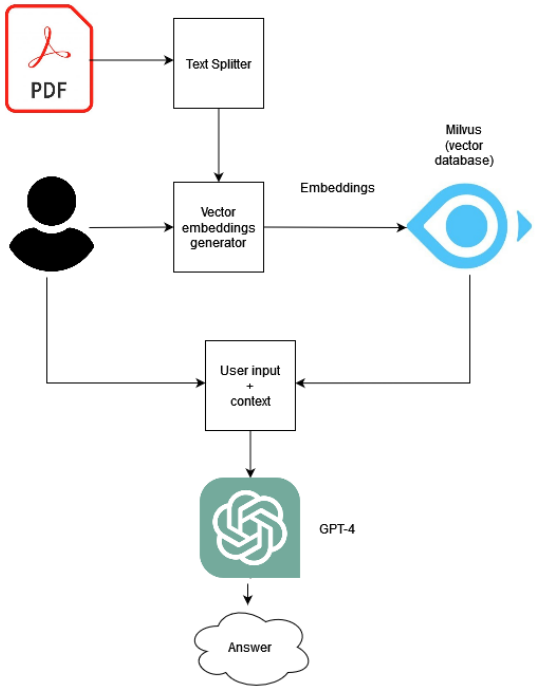

Figure 3. Diagram of key components integrated into the system.

One of the more important tools for implementing ChatGPT in Insieme is the LangChain library, which facilitates working with large language models (LLM). LLMs can efficiently perform a large number of different tasks, but there is a likelihood that they will not be able to correctly answer questions from specialized fields, such as medicine. LangChain helps us upgrade our models with knowledge of specific fields and enables them to be aware of data and conversation context. LangChain is a powerful tool that fills the gap between language models and domain knowledge, which is also why LangChain is increasingly used in applications that perform tasks related to natural language processing. LangChain includes numerous modules that help in development [18]. The next essential component of the system is vector database, presented in section 4.2. We de-

cided to use the open-source vector database Milvus, which allows efficient storage of vectors, their indexing, and also offers an API (application programming interface) that enables easy integration with different programming languages. The connection between the Milvus vector database and the GPT-4 language model is straightforward, as there is no need to specially label the data or retrain the model. The data needs to be converted into vector form and stored in Milvus. The final response generated by the model is thus created by referencing content in our document collection, ensuring that the virtual assistant obtains the right data and consequently reduces the likelihood of errors.

The first step in this development is uploading data into 'Documents', which are actually pieces of text. The Document Loader module in the LangChain tool simplifies this task and allows for easy uploading and preprocessing of our data — we can use DirectoryLoader, which allows us to store all used documents in a common directory. This is followed by dividing the documents into smaller pieces — the text splitter allows for breaking long text parts into smaller, semantically meaningful chunks [19]. This task may seem simple, but it involves some complexity. The goal is to divide the text in a way that keeps semantically connected parts together, where 'semantic connectivity' depends on the type of text being processed. Text splitters divide the text into small pieces, often based on sentence boundaries. These small pieces are combined into larger pieces until they reach a certain size determined by a pre-defined function for measuring the size of the piece — when a piece reaches the desired size, it becomes an independent piece of text. Then a new piece is created with some overlap (chunk overlap) to maintain context between individual pieces.

This is followed by the generation of word embeddings, which play a key role in representing textual information. The Embedding class in the LangChain tool serves as a standardized interface for various embedding providers, including OpenAI. Through the generation of word embeddings, the text is converted into a vector representation, enabling semantic analysis and tasks such as semantic searching. All this is stored in our vector database as a new index using built-in methods enabling semantic searches over this object and retrieval of documents relevant to the user's input. The obtained documents are then forwarded to the language model, which treats the documents as context for generating a response.

In Figure 4, one can see the process of responding to user questions. ChatGPT fluently answers health questions by considering general knowledge, medical knowledge from the entire web, and specific knowledge from the Insieme platform. The user can fluently change language from Slovenian to English and Italian. All users are anonymous in order to prevent identification. The communication is through the platform calling for GPT-4 enriched by the Insieme platform information and knowledge.

This reply at first glance resembles replies from GPT-4, but in reality they are tailored for an average local user in terms of understandability and in addition they are checked by medical experts so no hallucination is possible - whenever the reply can be obtained directly from the Insieme platform.

The next phase is typically a discussion about which measures to take for prevention and cure, and where professional help can be found.

The verification and assurance of informative and appropriate responses were carried out by the authors of the paper and members of the department, and medical experts, but no formal evaluation was yet performed. ChatGPT was tested with questions that covered the knowledge contained in the Insieme platform with the purpose to improve the system.

The generated responses were critically evaluated, and corrections were introduced in case of any deviations from expectations. The accuracy of the obtained answers is of crucial importance to the user, as misleading and inaccurate answers, if considered by the user, can in severe cases even lead to a deterioration of health conditions. The content of the Insieme platform, which provides the context for the virtual assistant to generate responses, was previously prepared by healthcare experts, allowing the user to obtain verified and useful information. Testing the system was performed by medical and AI experts. We have inserted one PDF document, which includes the text from Insieme platform of size around 1 MB.

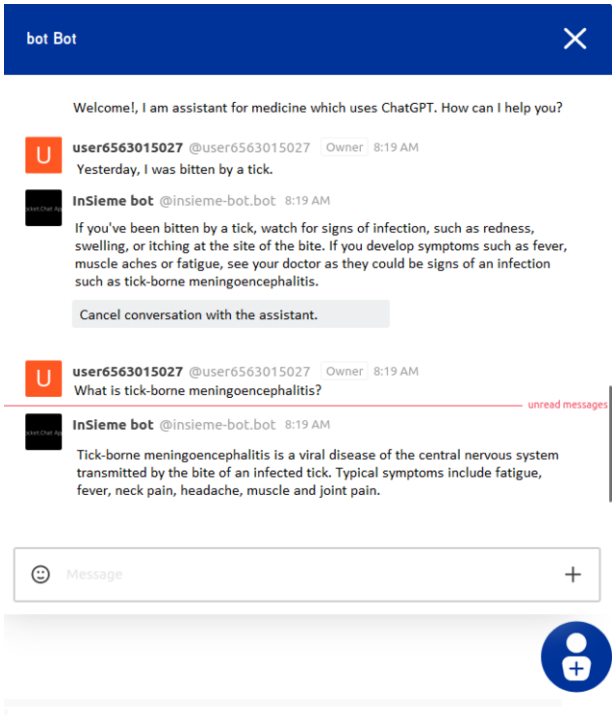

Figure 4. An example of questions and answers from ChatGPT in the multilingual platform Insieme.

5. Conclusion

This paper presents the implementation of the conversational virtual assistant ChatGPT into an electronic and mobile health platform called Insieme. The core idea is to provide information as GPT4, "Dr. Google" and the Insieme platform in a way that the core medical world-wide and local knowledge (e.g. local E-health mobile services, medical institutions) is integrated for most common issues, providing multimodal multimedia information, e.g. QA, video information. The rare medical cases are left to GPT-4 to figure them out.

In general, in the sense of reading from a vector database, the experiments demonstrate that the prototype system is capable of providing both standard ChatGPT responses

and additional information based on data accessible from the platform. The implementation was successful for testing in terms of local evaluation.

As part of the introduction of the virtual assistant, other language models from the Llama 2 family were also tested and locally installed. A problem arose as most of the training data used was in English, making the model unsuitable for use in the Slovenian language. The possibility of testing more advanced models from the Llama 2 family still remains open in the future, primarily because of local installation. This brings a new aspect of use, as in this way all data would be locally accessible, eliminating the need for external data access, as is currently necessary with the use of ChatGPT. Sending formal medical data out of Slovenia without the consent of the user is legally not permitted.

A problem with all communication with GPT-4 is that the quality of reply relies on the quality of input. If a user inputs wrong information, GPT-4 will not provide sensible replies. It is also a problem with training data [19].

Our main aim is for the platform to present foundational ideas on how the healthcare system could be modernized, while aiming to relieve the burden on professionals in this field and simultaneously provide all users with access to an effective and constantly available source of information based on the latest research findings.

Experiments in this study show that generative artificial intelligence is indeed useful and promises radical improvements if successfully implemented in Slovenian healthcare, in particular if enriched by general basic and local information and knowledge.

Acknowledgment

The Insieme platform was developed as part of the cross-border Interreg ISE-EMH project, which is funded by the Italy-Slovenia Cooperation Program from the European Regional Development Fund. Support was also provided by ARIS - Slovenian Research and Innovation Agency. We also thank members of the Department of intelligent systems and medical experts providing info and testing the system.

References

[1] World Health Organization, “Health and care workforce in Europe: time to act.”, Health and care workforce in Europe: time to act. 2022.

[2] Ayers, John, Adam Poliak, Mark Dredze, Eric Leas, Zechariah Zhu, Jessica Kelley, Dennis Faix, Aaron Goodman, Christopher Longhurst, Michael Hogarth, and David Smith, “Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum”, *JAMA Internal Medicine*, vol. 183, April 2023, doi:10.1001/jamainternmed.2023.1838.

[3] GPT-4 Technical Report, OpenAI, 2023, 2303.08774, arXiv, cs.CL

[4] Tirth Dave, Sai Anirudh Athaluri, Satyam Singh, “ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations”, *Frontiers in Artificial Intelligence*, vol. 22, no. 37, May 2022, doi:10.3389/frai.2023.1169595.

[5] Tina Reed, “Can ChatGPT be used in healthcare”, Online, Available at: https://www.axios.com/2023/11/29/chat-gpt-health-care-medicine-clinical-diagnosis, Last accessed: January 10, 2024.

[6] Partha Pratim Ray, Poulami Majumder, “The Potential of ChatGPT to Transform Healthcare and Address Ethical Challenges in Artificial Intelligence-Driven Medicine”, *Journal of Clinical Neurology*, vol. 19, no. 9, September 2023, pages 509-511, doi:10.3988/jcn.2023.0158.

[7] Linda Rosencrance, “9 Uses of Generative AI in Healthcare”, Online, Available at: https://www.techopedia.com/9-uses-of-generative-ai-in-healthcare, Last accessed: January 10, 2024.

[8] Harvey Castro, "ChatGPT and healthcare", Independently published, February 2023.
[9] Anuradha Welivita, "A survey of consumer health question answering systems", *AI Magazine*, November 2023, doi:https://doi.org/10.1002/aaai.12140.
[10] Bertalan Mesko, Generative AI in Healthcare, Pearson: 1st Edition, 2023,
[11] Peter Lee, Carey Goldberg, Isaac Kohane, The AI Revolution in Medicine: GPT-4 and Beyond, The Medical Futurist, 2023,
[12] Ghadiri, N., Comparison of GPT-3.5, GPT-4, and human user performance on a practice ophthalmology written examination' and 'ChatGPT in ophthalmology: the dawn of a new era?'., Eye 2023, https://doi.org/10.1038/s41433-023-02773-9
[13] Insieme Platform, "Insieme Platform", September 2023, Available at: https://ise-emh.eu, Last accessed: September 3, 2023.
[14] Naveed, Humza and Khan, Asad Ullah and Qiu, Shi and Saqib, Muhammad and Anwar, Saeed and Usman, Muhammad and Akhtar, Naveed and Barnes, Nick and Mian, Ajmal, A Comprehensive Overview of Large Language Models, ACM Computing Surveys, 2023, 55, 3, 1–30, 2307.06435, arXiv, cs.CL
[15] Qilu Jiao and Shunyao Zhang, A Brief Survey of Word Embedding and Its Recent Development, 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, 5, 1697-1701, 10.1109/IAEAC50856.2021.9390956
[16] Wang, Jianguo and Yi, Xiaomeng and Guo, Rentong and Jin, Hai and Xu, Peng and Li, Shengjun and Wang, Xiangyu and Guo, Xiangzhou and Li, Chengming and Xu, Xiaohai and Yu, Kun and Yuan, Yuxing and Zou, Yinghao and Long, Jiquan and Cai, Yudong and Li, Zhenxiang and Zhang, Zhifeng and Mo, Yihua and Gu, Jun and Jiang, Ruiyi and Wei, Yi and Xie, Charles, Milvus: A Purpose-Built Vector Data Management System, 2021, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3448016.3457550 10.1145/3448016.3457550, Proceedings of the 2021 International Conference on Management of Data, 2614–2627, 14, vector database, data science, machine learning, high-dimensional similarity search, heterogeneous computing, Virtual Event, China, SIGMOD '21
[17] Arvind Neelakantan and Tao Xu and Raul Puri and Alec Radford and Jesse Michael Han and Jerry Tworek and Qiming Yuan and Nikolas Tezak and Jong Wook Kim and Chris Hallacy and Johannes Heidecke and Pranav Shyam and Boris Power and Tyna Eloundou Nekoul and Girish Sastry and Gretchen Krueger and David Schnurr and Felipe Petroski Such and Kenny Hsu and Madeleine Thompson and Tabarak Khan and Toki Sherbakov and Joanne Jang and Peter Welinder and Lilian Weng, Text and Code Embeddings by Contrastive Pre-Training, 2022, 2201.10005, arXiv, cs.CL
[18] Amogh Agastya, "Harnessing Retrieval Augmented Generation With Langchain", September 2023, Available at: https://betterprogramming.pub/harnessing-retrieval-augmented-generation-with-langchain-2eae65926e82, Last accessed: September 5, 2023.
[19] Sreeram A S, Adith and Sai, Pappuri Jithendra, An Effective Query System Using LLMs and LangChain, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), 12, 06, June, 2023