DOI: 10.37943/15XNDZ6667

Aigerim Mansurova
Bachelor of Information Technology and Management, Student of Master's Applied Data Analytics
222215@astanait.edu.kz, orcid.org/0009-0003-1978-9574
Astana IT University, Kazakhstan

Aliya Nugumanova
PhD, Head of the Research and Innovation Center "Big Data and Blockchain Technologies"
a.nugumanova@astanait.edu.kz, orcid.org/0000-0001-5522-4421
Astana IT University, Kazakhstan

Zhansaya Makhambetova
MSc Engineering Management, Department of Science and Innovation
zhansaya.makhambetova@astanait.edu.kz, orcid.org/0000-0001-5024-0289
Astana IT University, Kazakhstan

DEVELOPMENT OF A QUESTION ANSWERING CHATBOT FOR BLOCKCHAIN DOMAIN

Abstract. Large Language Models (LLMs), such as ChatGPT, have transformed the field of natural language processing with their capacity for language comprehension and generation of human-like, fluent responses for many downstream tasks. Despite their impressive capabilities, they often fall short in domain-specific and knowledge-intensive domains due to a lack of access to relevant data. Moreover, most state-of-art LLMs lack transparency as they are often accessible only through APIs. Furthermore, their application in critical real-world scenarios is hindered by their proclivity to produce hallucinated information and inability to leverage external knowledge sources. To address these limitations, we propose an innovative system that enhances LLMs by integrating them with an external knowledge management module. The system allows LLMs to utilize data stored in vector databases, providing them with relevant information for their responses. Additionally, it enables them to retrieve information from the Internet, further broadening their knowledge base. The research approach circumvents the need to retrain LLMs, which can be a resource-intensive process. Instead, it focuses on making more efficient use of existing models. Preliminary results indicate that the system holds promise for improving the performance of LLMs in domain-specific and knowledge-intensive tasks. By equipping LLMs with real-time access to external data, it is possible to harness their language generation capabilities more effectively, without the need to continually strive for larger models.

Keywords: Chatbot, LLM, LangChain, RAG, NLP, ChatGPT.

Introduction

In recent years, the fintech industry in Kazakhstan has undergone significant changes due to the introduction of new technologies. For example, the Kaspi application makes it possible to make mobile payments and use banking services for more than 12 million users in 2022 [1]. At the moment, the development of blockchain technologies, in particular the development

Copyright © 2023, Authors. This is an open access article under the Creative Commons CC BY-NC-ND license
Received: 02.08.2023 Accepted: 13.09.2023 Published: 30.09.2023

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

and implementation of digital currency are becoming an integral part of the development strategies of financial systems. In this context, the National Bank of the Republic of Kazakhstan (NBK) is considering the possibility of introducing the digital tenge.

The introduction of the digital tenge aims to improve the financial infrastructure, ensure transparency and efficiency of payment transactions, stimulate innovation and support economic growth in Kazakhstan. Fintech companies and the government are actively exploring the possibilities of using blockchain technology to improve financial services. However, it is important to remember about citizens who are also participants in the digital economy. Since successful digital transformation depends on the level of trust and acceptance of citizens [2],[3].

The problem is that, despite the fact that digital currencies have been around for more than a decade, ordinary citizens still have insufficient awareness of how they work and what consequences may arise when using them. This may lead to an incomplete understanding of the risks and benefits of using digital currencies, as well as the potential impact that blockchain technology can have on the financial industry. In other words, in order to fully realize the potential of digital financial technologies, it becomes an important task to raise awareness of citizens in this aspect.

Thus, developing innovative approaches to raise citizen awareness is key to the successful implementation of projects in blockchain domain. One such solution is the development of a chatbot capable of answering questions related to blockchain technologies.

Literature review

A chatbot is a computer program that uses text to conduct chats over the Internet [4]. It is designed in such a way as to resemble the interaction of a person with an interlocutor as much as possible and help them find the necessary information without requiring human participation [5]. Thus, chatbots offer a natural language interface that allows users to perform actions without contacting people directly through meetings, emails or phone calls [6].

In 1950, Alan Turing introduced the Turing Test, posing the question of whether machines can think, marking the popularization of the concept of chatbots [7]. The earliest known chatbot, Eliza, was created in 1966 with the goal of functioning as a psychotherapist, responding to user inputs with questions [8]. Eliza employed a basic pattern matching system and used predefined templates for its responses. Although its conversational abilities were limited, it inspired further chatbot development [9]. A notable improvement over Eliza was PARRY, a personality-based chatbot developed in 1972 [10]. In 1995, ALICE, a chatbot utilizing a simple pattern-matching algorithm and powered by the Artificial Intelligence Markup Language (AIML) [11], was created. Subsequently, chatbots like SmarterChild [12] emerged in 2001 and became accessible through messenger applications. The evolution continued with the development of virtual personal assistants such as Apple Siri, Microsoft Cortana, and Google Assistant. Figure 1, based on Scopus data [13], illustrates a significant surge in interest in chatbots, particularly after the year 2016.

DOI: 10.37943/15XNDZ6667
© Aigerim Mansurova, Aliya Nugumanova,
Zhansaya Makhambetova

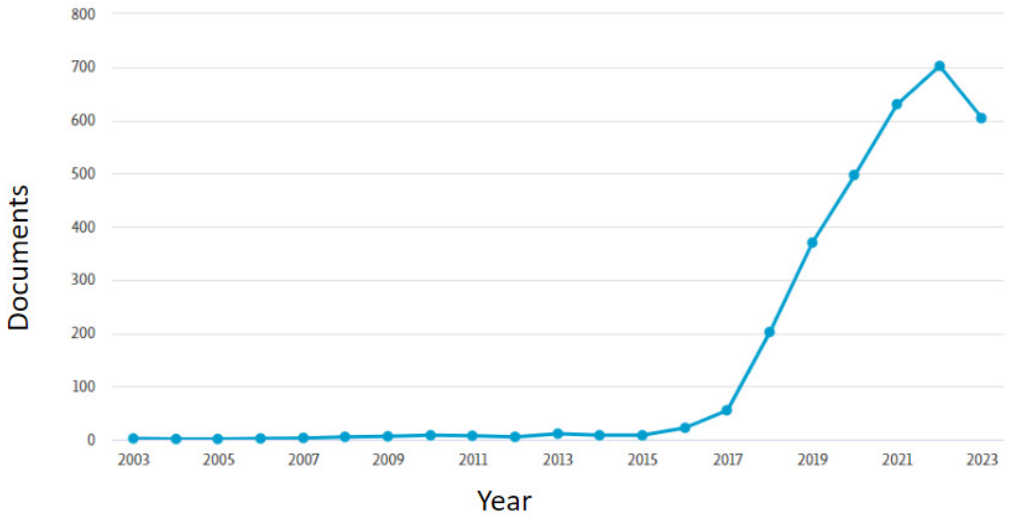

Figure 1. Growth of Chatbot Research (Scopus data, 2003-2023)

The current technological breakthrough is ChatGPT's, a chatbot based on Large Language Models (LLMs). The research community uses the term "Large Language Models" for large-scale Pre-trained Language Models (PLMs) [14-15], like PaLM with a parameter of 540B and GPT-3 with a parameter of 175B. The basic idea of PLM is to pre-train models on large unsupervised data and then fine-tune them to improve performance in supervised tasks. With the advent of higher computational power and Transformer technology [16], the PLMS architecture has evolved to more sophisticated architectures such as BERT [17] and OpenAI GPT [18]. ChatGPT is capable of performing a wide range of complex text queries and tasks such as coding, creating thank-you notes, and guiding people in complex performance discussions [19].

With the introduction of ChatGPT, question-and-answer technologies have become very attractive to users. Although they are very useful as a tool for solving the above tasks, they are not enough as chatbots with which end users communicate to explore and learn more about what they find interesting, or to satisfy the need for reliable information.

During the review of existing literature, several studies were found on the impact of using ChatGPT on the specific domain: education, economics and finance, programming and medicine. The paper [20] investigates the ability to improve academic performance in economics and finance using ChatGPT. The study concludes that ChatGPT has the potential to help researchers in analyzing data, developing scenarios, and disseminating research results. Another study [21] also highlights the language model's ability to perform a variety of programming tasks, including providing guidance on code analysis and correction, as well as answering technical questions. Research in the field of medicine has explored the potential of using ChatGPT to simplify radiological reports intended for patients [22]. The results of the study confirmed that the quality of simplified reports created using ChatGPT was mostly accurate. However, some errors were also identified that could pose a potential danger if medical personnel were not involved.

Although LLM-based chatbots have made significant progress, they face two major challenges:

Hallucination. In factual text generation, a challenging problem is the generation of hallucinations [23] - confident but misleading information. Hallucinations are widespread in existing LLMs, even in the most advanced LLMs such as GPT 4 [24]. When deploying LLMSs in real-world applications, hallucinations can mislead the system, which would likely result in unwanted outputs and significant performance degradation.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

Lack of novelty of knowledge. LLMs are "frozen in time" and do not contain current information, especially in a particular subject area. For example, ChatGPT also doesn't know anything about events occurring after 2021 [25]. Consequently, there are difficulties in solving problems that require up-to-date knowledge that goes beyond the data on which they were trained.

Hence, LLMs do not have the ability to accurately tell about events that occurred after their preliminary training and are much less aware of less popular topics. They do, however, tend to generate hallucinations. Therefore, if a user of chatbots like ChatGPT wants to learn more about fintech news or regulation, they need to carefully and painstakingly check any information they receive against external sources so as not to be misled.

There are three main ways in the literature [26] to overcome the above problems:

a. Creation of foundation model.

Creating a large language model is a significant endeavor. According to CEO of OpenAI, estimated cost of developing the foundational model behind ChatGPT cost more than $100 million [27]. However, beyond costs, recruiting specialized talent, obtaining and preparing datasets, and navigating technical hurdles pose additional obstacles. Even with sufficient resources and expertise, success is not guaranteed, as the AI industry has seen ambitious start-ups both succeed and fail.

b. Fine-tuning.

Fine-tuning involves adapting a foundational model to your specific domain's data or new tasks, and while it may be a cost-effective alternative to building a model from scratch, it still presents challenges. It requires specialized expertise, sufficient data, and entails ongoing costs and technical complexities for model deployment. This approach is outdated and involves recurring, expensive data labeling by subject-matter experts, ongoing quality monitoring, and the risk of accuracy decline as data changes over time, necessitating frequent re-labeling and fine-tuning efforts. In fact, constantly updating a debugged model to maintain accuracy when data changes can be akin to updating the model every time, which makes this process cumbersome and resource-intensive.

c. Prompt engineering.

Prompt engineering is insufficient for reducing hallucinations. Prompt engineering, which involves adjusting the instructions given to the model to guide its behavior, is an economical way to enhance the accuracy of the application. Yet, it has limitations as it cannot provide new or dynamic context to the model.

Therefore, there is a need to create efficient and effective approaches that can incorporate up to date knowledge into current LLMs. But this strategy is still at a superficial level.

Recent advances have focused on jointly finetuning the retriever and generation components of retrieval-augmented text generation systems [28],[29]. However, these approaches are not applicable to proprietary LLMs, which inner workings are not interpretable. The study [30] has explored how to integrate external knowledge sources, such as search engines, to complement LLMs. The approach made GPT-3 more faithful by utilizing the exact question as a query, initiating a search operation through the Google Search API. Other research [31] has incorporated the information from a dynamic memory of user's corrective feedback into the context, so that LLMs could perform better on relevant tasks. In contrast, we retrieve external relevant information from vector database and incorporate it into prompt. Another difference of this work is in the proposed architecture, which opens up the possibility of using machine translation instead of following the trend in the field of natural language processing to create models in the native language even for low-resource languages. The authors of the paper [32] have empirically demonstrated the validity of this alternative approach. The combination of a large language model of English with modern machine translation outperformed many models of low-resource Scandinavian languages.

DOI: 10.37943/15XNDZ6667
© Aigerim Mansurova, Aliya Nugumanova,
Zhansaya Makhambetova

Methodology

Our research work was carried out in accordance with the design process named double diamond [33].

Problem phase:

1. At the first stage, our research included an analysis of question-and-answer chatbots and the current state of their functioning, as well as the possibility of their usage in a certain domain knowledge. We tried to understand what type of chatbots exist and how they can be developed.
2. At the second stage, we formulated the problem statement and defined the purpose of this study. The purpose of this work is to create a user application that would allow providing information from local and verified sources to users interested in the topic of blockchain technology using artificial intelligence resources used by OpenAI.

Solution phase:

3. Then we explored possible ways to implement a chatbot, trying different solutions for each research goal and each problem we encountered.
4. In the next phase of our research, the solutions we implemented began to merge into a workable prototype. This happened after we determined which approaches are most effective for the chatbot and best match the goals we set in this study.

This design process gave us the opportunity to experiment with research. In the first part of the research on the topic of chatbots, we acquired a lot of new knowledge about the current state of this technology and its application for QnA purposes. In the second part, we were able to introduce new components into the functionality of the chatbot, adding new features that have not been used before.

The overall architecture of our proposed solution is shown in Figure 2. It consists of a frontend layer that implements the user interface and a backend layer. The backend layer, in turn, consists of a database, a module that loads data into a database, a module that implements application logic and includes Python code, access to the LangChain framework, Open AI. An agent is a flexible component based on LLM that makes tool selection decisions.

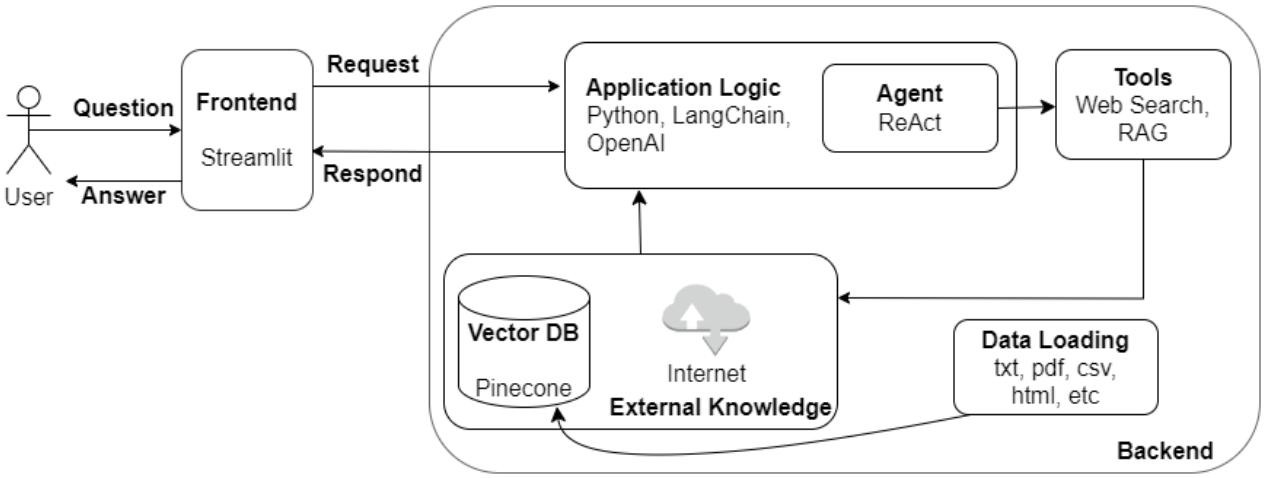

Figure 2. General architecture of the proposed solution

The use of Python as the selected programming language for this project provides developers with versatility and robust capabilities, catering to a wide range of skill levels. Python is extensively employed in diverse domains, including web development, data analytics, information science, artificial intelligence, and machine learning, among numerous others [34].

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

Additionally, Python's seamless compatibility with various operating systems and platforms positions it as an outstanding option for software development.

The choice to employ OpenAI's GPT-3.5 Large Language Model is a strategic decision informed by its demonstrated capabilities. GPT-3.5 represents a culmination of advancements in large language models, boasting an extensive knowledge base and natural language processing proficiency. Its core strength lies in its ability to generate coherent and contextually relevant text, making it a valuable asset in numerous applications, from chatbots to content generation. Furthermore, GPT-3.5's adaptability through fine-tuning allows it to be tailored to specific tasks and domains. Its widespread adoption and support from the developer community also contribute to its appeal. In essence, the selection of GPT-3.5 stems from its proven track record in delivering human-like text generation and its potential to enhance a wide array of language-related tasks.

Moreover, convenient use of various data sources, integration with OpenAI and other components of the system was achieved through LangChain [35]. It is a framework for developing applications that utilize large language models. LangChain can provide such an opportunity with the help of its components such as Prompt Templates, Chains, Memory, Models and Agents.

A prompt template is a predefined structure or format used in LangChain for generating prompts (input to a language model). These templates help in providing consistency and structure to the prompts that users interact with when using the LLM. Here are some key components of a Prompt template:

Table 1. Main components of Prompt Template

Component	**Description**	**Example**
Text String (Template)	Base structure for prompts with placeholders	"Tell me about [topic]."
Parameters	Variables filled with specific information	[topic] = " Distributed Ledger Technology"
User Input (Question)	Represents user's query or request	"What can you tell me about Distributed Ledger Technology?"
Few-Shot Examples	Short input examples for context	"query": " What is a cryptocurrency?", "answer": " A cryptocurrency is a virtual form of currency that uses cryptography for security and operates independently of a central authority. "
Instruction	Guidance for processing user input	" Below are excerpts from conversations with an artificial intelligence assistant. The assistant should only give correct answers to users' questions, and if it doesn't know the answer, it says 'I don't know'."

The primary foundational element within LangChain is the Chain, which typically combines an LLM with a prompt. A basic chain operates with a single input prompt and generates an output, and we can run multiple chains consecutively, where the output of one chain serves as the input for the next.

Memory in the context of a conversational system refers to the system's capability to store and recall information from prior interactions, as depicted in Figure 3. While a LLM is inherently stateless and does not hold onto previous interactions, the introduction of an external memory component allows for the implementation of memory in a chatbot system. The following is an explanation of how the memory mechanism illustrated in Figure 3 works. This component stores the conversation history separately from the LLM. Upon receiving a user

DOI: 10.37943/15XNDZ6667
© Aigerim Mansurova, Aliya Nugumanova,
Zhansaya Makhambetova

query, the chatbot pulls the relevant context from the external memory, combines it with the present prompt, and passes it to the LLM. This process helps in generating responses that are informed by the context of previous conversations.

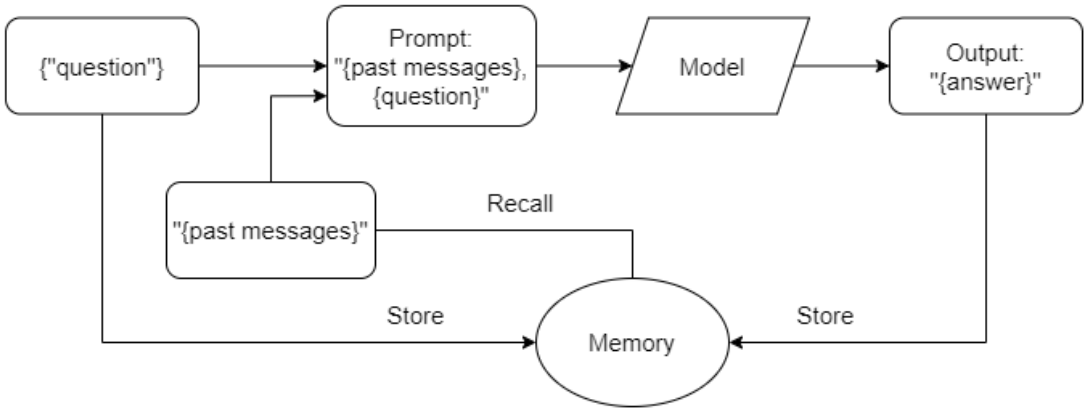

Figure 3. Representation of how memory works

In LangChain, there are various types of models used, including LLMs (which have text string as input, and a text string as output), Chat Models (which have a structured API for processing chat messages), and Text Embedding Models (which take text as input and return its corresponding embedding as a list of floating-point numbers).

An agent in LangChain is a flexible component, often based on LLMs, equipped with multiple tools like search engines and calculators. It serves to make tool selection decisions based on user input, enabling dynamic sequences of interactions with LLMs and other tools in applications. In this work, we will utilize two tools: RAG (Retrieval-Augmented Generation) and Web Search.

Pinecone is the next tool that was used. It is a cloud-based vector database designed specifically for use in machine learning applications. A vector database works differently than traditional databases because it works with vector data (embeddings) instead of scalar data. Pinecone uses Approximate Nearest Neighbor Search (ANN) algorithms that optimize similarity search. It simplifies data management tasks such as insertion, deletion and updates, which can be cumbersome when using standalone indexes such as FAISS.

Finally, in order to create desktop applications with an interactive graphical user interface, we employed Streamlit, a free and open-source framework. This Python-based library enables the rapid development and sharing of visually appealing machine learning and data science web applications.

The key stages of application development encompass the following:

1. Design Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) is a technique in natural language processing that combines retrieval models and generative models to improve the quality and relevance of generated text [28]. Retrieval models retrieve specific information from a knowledge base using semantic search.

Semantic search is a search technique that aims to understand and provide search results based on the meaning and context of a user's query rather than just matching keywords. While generative models create text based on prompts. Retrieval-augmented generation merges these approaches, using retrieved information to enhance the accuracy and relevance of text generated by a generative model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

To create a RAG system the following steps must be performed:

1.1. Document Decomposition and Loading. The documents, which can be in various formats such as txt, pdf, html, etc., must be deconstructed into chunks and loaded into LangChain. It is possible to control the granularity of text splitting by adjusting two key parameters:

Chunk Size: This parameter determines the maximum number of characters that each chunk can contain.

Chunk Overlap: This parameter defines the number of characters that should overlap between two adjacent chunks.

By modifying these parameters, we can tailor the text splitting process to match our specific needs, whether it requires fine-grained segmentation with many smaller chunks or a coarser segmentation with larger chunks.

1.2. Embedding Creation. Embeddings of document chunks are generated in preparation for semantic search. Embeddings are mathematical representations that encapsulate the semantic meaning of text, thereby enabling efficient and accurate identification of similar text chunks.

1.3. Vector Database Maintenance. A Pinecone vector database must be created and maintained, which will store data that has been transformed into vectors. The advantage of this database lies in the fact that embeddings can be stored permanently, obviating the need to regenerate them each time the application is initiated.

In the developed system a procedure has been established wherein a query is subjected to an embedding process, subsequently enabling the retrieval of documents that exhibit similarity to the query. Following this retrieval, the question along with the retrieved documents can be forwarded to a LLM in order to obtain an output. This approach ensures efficient processing of requests for which there are answers in external knowledge, however, if there is no answer there, the system can start generating answers that do not correspond to the truth. And the system can be further improved by integrating these models with the globally renowned search engine, Google Search.

2. Web Search

By combining OpenAI models with Google Search, our chatbot can provide accurate and relevant information in real-time. This synergy not only makes the conversation smoother and more efficient but also ensures that the information provided is trustworthy and valuable.

SerpApi, a Google Search results API, will be utilized to create customized Search tool. The interaction is facilitated through Python's google-search-results library. It simplifies access to Google search results. By installing this library, we gain the capability to utilize the API for retrieving search results such as web pages, images, news, and more, and integrate them into our application.

When a user request is understood, actions are performed, and information is retrieved. Chatbot performs the requested actions by retrieving the data it is interested in from the information sources accessed via API call.

Thus, the integration of semantic search and external knowledge retrieval forms the foundation of our chatbot system. Versatile chatbot agent that leverages the power of both generative and retrieval-based techniques, delivers high-quality responses to user queries. This approach not only ensures the accuracy and relevance of responses when there is existing knowledge but also allows the system to generate informed answers when necessary. The integration of a versatile chatbot agent that leverages the power of both generative and retrieval-based techniques results in the delivery of high-quality responses to user queries.

DOI: 10.37943/15XNDZ6667
© Aigerim Mansurova, Aliya Nugumanova,
Zhansaya Makhambetova

Evaluation

For this experiment, in order to test proposed architecture of the chatbot system, we narrowed the scope of our documents on the topic of the Digital Tenge project. We also conducted a survey in order to assess the effectiveness of the chatbot in enriching users' knowledge about the Digital Tenge project and to assess their level of satisfaction with the interface of the system. The total number of participants in the survey was 20. Respondents were asked the following questions:

1. "Have you gained new knowledge or learned more about the 'Digital Tenge' project after using the chatbot?" (A 'Yes' or 'No' question).
2. "Do you consider the presence of similar systems such as chatbot necessary for informational support of the population?" (A 'Yes' or 'No' question).
3. "Are you satisfied with the interface of the website where you interact with the chatbot?" (Assessed on a five-point scale, where 1 signifies 'Strongly Dislike,' and 5 signifies 'Strongly Like').

Result

A total of 1,143 vectors were obtained from 381 pages of text about Digital Tenge project, using the "text-embedding-ada-002" embedding model. And the application is ready to receive questions and to provide answers to them using OpenAI's "gpt-3.5-turbo" language model in the background.

Figure 4. Chatbot Memory

After entering the question, there are two possible courses of events: first when information is found in external knowledge and an answer is generated based on it (Figure 4), or the other when there is no information to formulate an answer. In this case, the application provides information that was obtained from the search services, allowing services to give reliable an-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

swers to a user (Figure 5). Figure 4 also represents the memory of the chatbot. It responded to a question using the information it had in its memory about previous questions and answers.

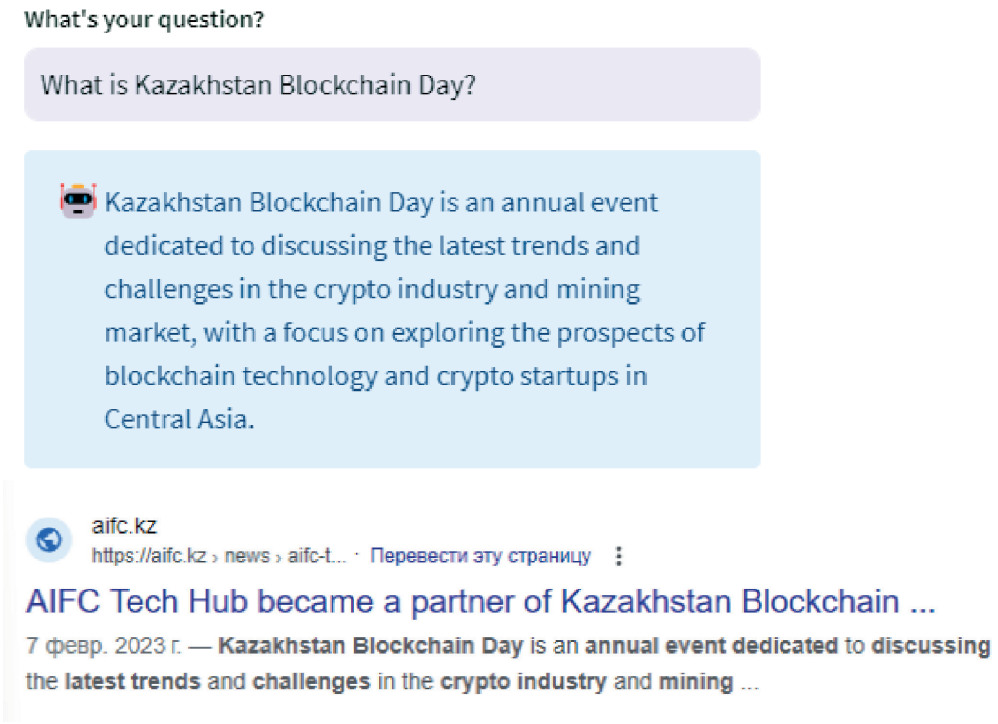

Figure 5. Delivering Reliable Answers via Search Services

It's noticeable that our chatbot provides accurate answers, while ChatGPT 3.5 may provide questionable information (Figure 6) if it hasn't learned about a specific topic or event from its training data.

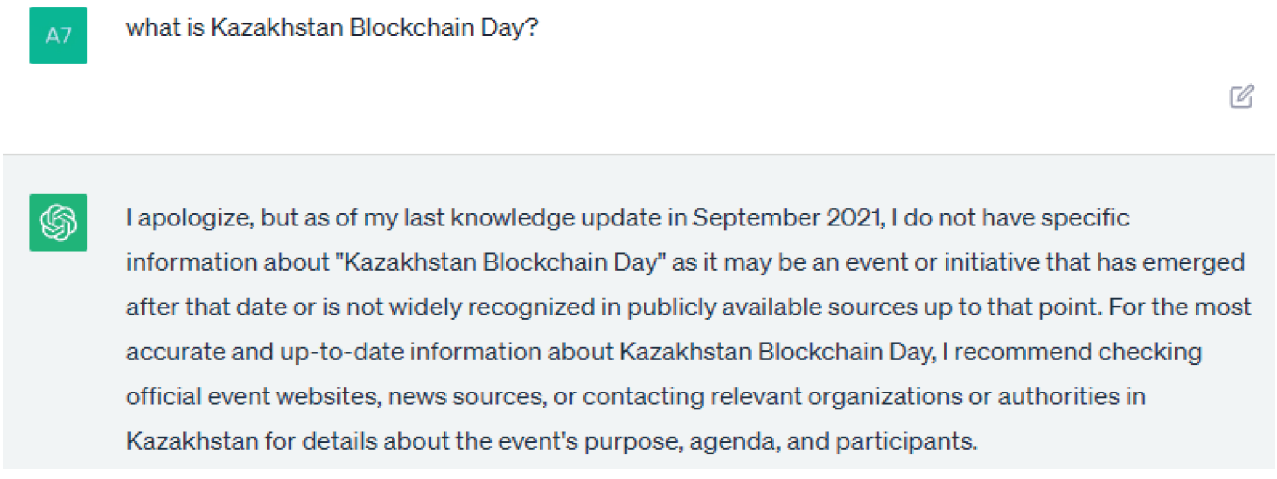

Figure 6. Illustrating Limitations in ChatGPT's Knowledge

All the components of the system were successfully realized. There is a screenshot showing how the conversational chatbot looks like, with all the implementation mentioned above (Figure 7).

DOI: 10.37943/15XNDZ6667
© Aigerim Mansurova, Aliya Nugumanova,
Zhansaya Makhambetova

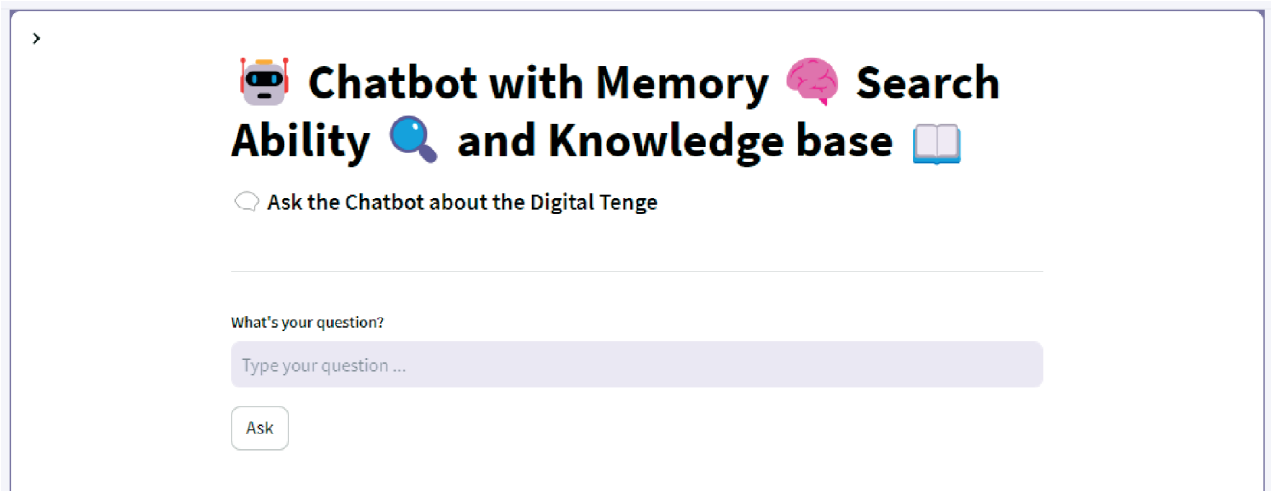

Figure 7. Screenshot of the Fully Implemented Chatbot

As a result of the survey, the following key aspects were revealed:

- 80% of the respondents answered affirmatively to the question of whether they gained new knowledge or deepened their understanding of the Digital Tenge project after interacting with the chatbot (Figure 8). This confirms the effectiveness of the chatbot as a means to increase knowledge about the project.
- 70% of the respondents agreed on the need for such systems, including the chatbot, to provide information support to the public (Figure 9). This indicates a broad recognition of the potential of such technological solutions for information services.
- The average rating of the website interface where users interact with the chatbot was 3.45 out of 5 points (Figure 10). This result indicates a satisfactory perception of the user interface by the majority of respondents. However, it also notes that there is potential for further improvement of the interface to better meet the needs of the users.

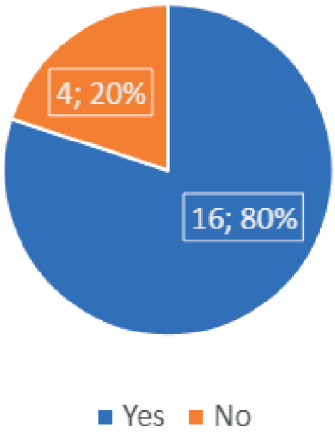

Figure 8. Survey response: Knowledge enhancement

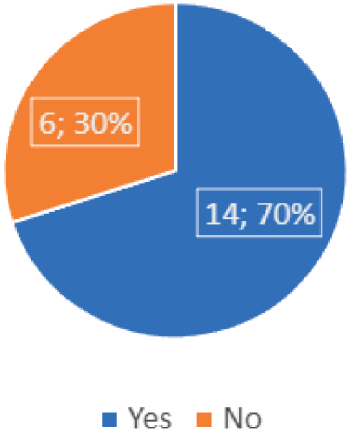

Figure 9. Survey response: Necessity of similar systems

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

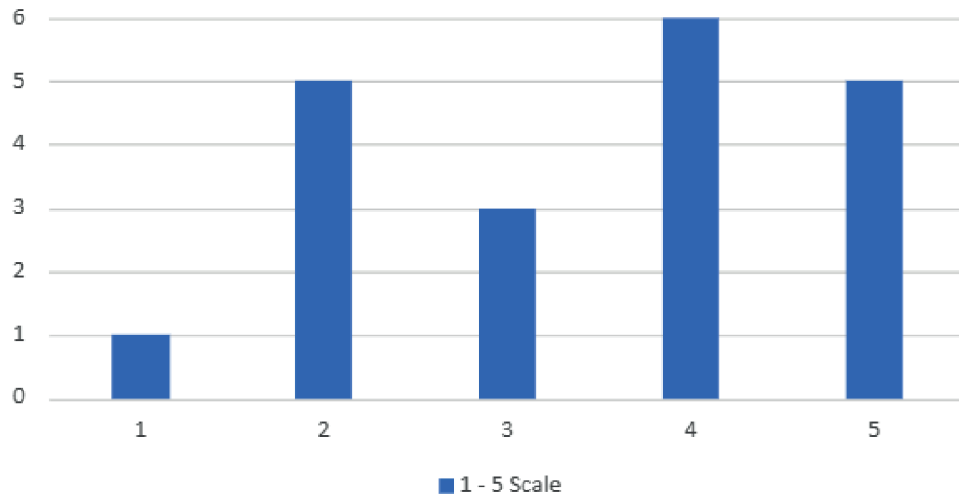

Figure 10. Survey results: Perceptions of the interface
(5 – point scale, from 1 – 'strongly dislike' to 5 – 'strongly like')

Overall, the results indicate a positive attitude towards the chatbot as a means of increasing awareness about the Digital Tenge project and providing information support to the population. Nevertheless, taking into account the noted potential for improvement of the interface, it is recommended to continue working on improving this aspect in order to ensure higher user satisfaction.

Conclusion

Overall, the research introduces a promising and novel approach to the development of a chatbot, grounded in the utilization of Language Model (LLM) technology. The architectural framework we have devised seamlessly integrates advanced technologies aimed at enhancing response factuality and relevance. The empirical evidence derived from improvements in user knowledge after interactions with the chatbot underscores the substantial positive impact of this system on the acquisition of knowledge pertaining to the Digital Tenge project. This finding serves to underscore the pivotal role of such chatbot systems in providing information support. The future research endeavors will focus on adaptation of this architecture for the Kazakh language.

Acknowledgement

This research has been funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP19677756 «Unsupervised term extraction: a set of models and datasets for high-tech domains and low-resource languages».

References

[1] Okasov, B. (2022, September 30). Dlja cifrovizacii vseh gosuslug v Kazahstane ispol'zujut opyt Kaspi.kz [The experience of Kaspi.kz is used to digitalize all public services in Kazakhstan]. *KTK*. https://www.ktk.kz/ru/news/video/2022/09/30/223965/

[2] Alaklabi, S., & Kang, K. (2021). Perceptions towards cryptocurrency adoption: A case of Saudi Arabian citizens. *Journal of Electronic Banking Systems*, 1–17. https://doi.org/10.5171/2021.110411

[3] Mensah, I.K., & Mwakapesa, D.S. (2022). The drivers of the behavioral adoption intention of bitcoin payment from the perspective of Chinese citizens. *Security and Communication Networks*, *2022*, 1–17. https://doi.org/10.1155/2022/7373658

[4] King, M.R. (2022). The future of AI in medicine: A perspective from a chatbot. *Annals of Biomedical Engineering*, *51*(2), 291–295. https://doi.org/10.1007/s10439-022-03121-w
[5] Carayannopoulos, S. (2018). Using chatbots to aid transition. *International Journal of Information and Learning Technology*, 35(2), 118–129. https://doi:10.1108/ijilt-10-2017-0097
[6] Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(05), 811–817. https://doi:10.1017/s1351324916000243
[7] Zemčík, M. T. (2019). *A brief history of chatbots*. https://www.semanticscholar.org/paper/A-Brief-History-of-Chatbots-Zem%C4%8D%C3%ADk/b72c89500dd57f1a4ceadb97f3dbf5015948a5e7
[8] Shum, H., He, X., & Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26. https://doi.org/10.1631/fitee.1700826
[9] Klopfenstein, L.C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of Bots. *Proceedings of the 2017 Conference on Designing Interactive Systems*. https://doi.org/10.1145/3064663.3064672
[10] Zemčík, T. (2019). A brief history of chatbots. *DEStech Transactions on Computer Science and Engineering*, *aicae*. https://doi.org/10.12783/dtcse/aicae2019/31439
[11] Marietto, M.D.G.B., De Aguiar, R.V., De Oliveira Barbosa, G., Botelho, W. T., Pimentel, E.P., França, R.D.S., & Da Silva, V.L. (2013). Artificial Intelligence Markup Language: A brief tutorial. *International Journal of Computer Science & Engineering Survey*, *4*(3), 1–20. https://doi.org/10.5121/ijcses.2013.4301
[12] Molnar, G., & Szuts, Z. (2018). The role of Chatbots in formal education. *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. https://doi.org/10.1109/sisy.2018.8524609
[13] *Scopus preview – Scopus – Welcome to Scopus*. (n.d.). https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=de60450aae1e6d2749d1e0a05c7dacdc&sot=a&sdt=a&sl=22&s=TITLE-ABS-KEY%28chatbot%29&origin=resultslist&count=10&analyzeResults=Analyze+results
[14] Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2023, May 1). *A comprehensive survey on pretrained foundation models: A history from Bert to chatgpt*. arXiv.org. https://arxiv.org/abs/2302.09419
[15] Y. Fu, H. Peng, and T. Khot. (2022). *How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources*. Notion. https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e-36fa1dc1
[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762
[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805
[18] OpenAI. (2022, November 30). Introducing ChatGPT. *OpenAI*. https://openai.com/blog/chatgpt#OpenAI
[19] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2021). GPT understands, too. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2103.10385
[20] Alshater, M. (2022). Exploring the role of artificial intelligence in enhancing academic performance: A case study of chatgpt. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4312358
[21] Biswas, S. (2022). Role of CHATGPT in computer programming. *Mesopotamian Journal of Computer Science*, 20–28. https://doi.org/10.58496/mjcsc/2022/004
[22] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., & Ingrisch, M. (2022, December 30). *Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports*. arXiv.org. https://arxiv.org/abs/2212.14882
[23] Bang, Y. (2023, February 8). *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity*. arXiv.org. https://arxiv.org/abs/2302.04023
[24] OpenAI. (2023, March 15). *GPT-4 Technical Report*. arXiv.org. https://arxiv.org/abs/2303.08774
[25] Natalie. (2023, October). *What is ChatGPT?* OpenAI Help Center. https://help.openai.com/en/articles/6783457-what-is-chatgpt

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 15, SEPTEMBER 2023

[26] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023, September 11). *A survey of large language models*. arXiv.org. https://arxiv.org/abs/2303.18223

[27] Knight, W. (2023, April 17). OpenAI's CEO says the age of giant AI models is already over. *WIRED*. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[28] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021, April 12). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv.org. https://arxiv.org/abs/2005.11401

[29] Zhang, Y., Sun, S., Gao, X., Fang, Y., Brockett, C., Galley, M., Gao, J., & Dolan, B. (2022). RetGen: A joint framework for retrieval and grounded text generation modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 11739–11747. https://doi.org/10.1609/aaai.v36i10.21429

[30] Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. (2022, May 23). *Internet-augmented language models through few-shot prompting for open-domain question answering*. arXiv.org. https://arxiv.org/abs/2203.05115

[31] Madaan, A., Tandon, N., Clark, P., & Yang, Y. (2023, February 18). *Memory-assisted prompt editing to improve GPT-3 after deployment*. arXiv.org. https://arxiv.org/abs/2201.06009

[32] Isbister, T., Carlsson, F., & Sahlgren, M. (2021). *Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?* ACL Anthology. https://aclanthology.org/2021.nodalida-main.42/

[33]Liu, D., & Hsu, H. (2009). An international comparison of empirical generalized double diamond model approaches to Taiwan and Korea. *Competitiveness Review: An International Business Journal*, *19*(3), 160–174. https://doi.org/10.1108/10595420910962043

[34] Saabith, A. S., Fareez, M. M. M., & Vinothraj, T. (2019). Python current trend applications-an overview. *International Journal of Advance Engineering and Research Development*, *6*(10). https://www.ijaerd.com/index.php/IJAERD/article/view/4419

[35] Topsakal, O., & Akinci, T. C. (2023). Creating large language model applications utilizing LangChain: A primer on developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*, *1*(1), 1050–1056. https://doi.org/10.59287/icaens.1127