Automated Central Bank Policy Question and Answer System based on Large Language Models

Shanshan Yan[1†], Fengting Mo[2†], Huixi Li[3], Yinhao Xiao[*]

{jiandanshoule@163.com[1], mof_ting@126.com[2], lihuixi@gdufe.edu.cn[3],20191081@gdufe.edu.cn[*]}

Guangdong University of Finance and Economics, Guangzhou, China
[†]These authors contributed equally to this work.

Abstract. The monetary policy is a crucial tool for guiding national economic regulations, as it upholds economic stability and fosters economic development. Specifically, the policies implemented by the Chinese Central Bank play a vital role in managing inflation and bolstering economic growth. These policies directly influence the living standards of the population and their financial decisions. However, in the past, interpreting these policies required a strong professional background in economics or finance, along with significant manual effort. The emergence of Q&A systems has provided the general public with better access to understanding the measures taken by the central bank and accessing monetary policy information more efficiently. Unfortunately, most of these Q&A systems employ basic and outdated natural language processing (NLP) algorithms and models, resulting in limited abilities to understand and reason about the policies and questions, ultimately leading to poor outcomes. In this paper, we present our approach to creating and training a state-of-the-art Q&A system for Chinese Central Bank policies using large language models (LLMs). Specifically, we demonstrate three of the most popular Chinese-based LLMs: IFlytek, ChatGPT, and LangChain-ChatGLM. Among them, IFlytek and ChatGPT are cloud-service-based LLMs trained through the prompt-tuning method. On the other hand, LangChain-ChatGLM utilizes the Langchain framework, which includes autoregressive gap filling and combines the advantages of autocoding and autoregressive pretraining. The ChatGLM-6B model is used for knowledge learning in this approach. We compare the performance of all these LLMs and publish the comparison results for future study.

Keywords: Central bank policy, Langchain-ChatGLM, LLM

1 Introduction

Central bank policies play a crucial role in guiding national economic and financial activities, with a direct impact on people's livelihoods. However, interpreting these policies often requires checking the official data release channels of the People's Bank of China (PBOC) or relevant financial institutions to obtain a comprehensive understanding of the latest updates on the central bank's monetary policies. In this work, we present a novel implementation of a Q&A system for the central bank's monetary policy, utilizing large language models (LLMs), significantly enhances the efficiency of information acquisition. This system provides accurate, consistent, and timely policy interpretations, facilitating effective information exchange between the central bank and various stakeholders. In the following context of the introduction section, we present some preliminary knowledge on the concepts related to our paper.

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China

DOI 10.4108/eai.27-10-2023.2341977

1.1 Central Bank Monetary Policy

Central bank monetary policy is a variety of guidelines, policies, and measures that the central bank governs and regulates the money supply and the quantity of credit in order to maintain economic stability. The monetary policy is real-time, professional, and confidential. The Monetary Policy Implementation Report contains five sections, namely monetary and credit overview, monetary policy operations, financial market operations, macroeconomic analysis, and monetary policy trends.

1.2 LangChain

LangChain is a versatile framework designed for the development of language model driven applications. Its primary purpose is to assist developers in building end-to-end applications that leverage the power of language models. With LangChain, developers can seamlessly integrate the language model into their applications and incorporate various data sources to facilitate interactions with the language model within the runtime environment. One of the key features of LangChain is its ability to query and extract information from documents. It enables users to ask questions related to the document and utilize the information within the document to construct accurate answers. This functionality enhances the application's capability to provide relevant and contextual responses. By leveraging LangChain, developers can harness the full potential of language models and create robust applications that incorporate natural language understanding and generation capabilities [1].

1.3 Large Language Model (LLM)

Large Language Modeling (LLM) is an advanced natural language processing technique rooted in machine learning. It involves constructing a model that comprehends human language and generates language automatically by learning from extensive textual datasets. In the area of natural language processing, LLM has many different applications. It can be used to carry out a variety of tasks, such as automatic question answering, machine translation, speech recognition, and text generation. By leveraging LLM, these tasks can be executed efficiently and effectively. ChatGLM-6B, an open-source bilingual conversational language model for Chinese and English, is a remarkable illustration of an LLM model [2]. Built on the GLM (Generative Language Modeling) architecture, ChatGLM-6B comprises an impressive 6.2 billion parameters. This model has been specifically optimized for Chinese Q&A and conversational scenarios, and it has been trained on an extensive dataset consisting of approximately 1 trillion Chinese and English examples.

In this paper, we propose a lightweight LLM-based central bank policy Q&A system by combining LangChain and ChatGLM. The system utilizes the open-sourced model ChatGLM, which can be deployed offline without relying on Internet connectivity. Additionally, the LangChain-based method does not require excessive hardware resources for training or fine-tuning, eliminating a common drawback. Our project leverages the LangChain framework to swiftly access multiple data sources, offering support for various file types such as PDF, TXT, MD, DOCX, and more. Furthermore, it is optimized specifically for Chinese usage scenarios, taking into account clause breaks and document reading. This optimization enables better handling of Chinese text and contexts, leading to more accurate results.

Our contributions are threefold:

• We present a novel LLM-based Central Bank Policy Q&A system that significantly improves the efficiency and professionalism of policy-querying services.

• Additionally, we explore different training approaches, such as LangChain-based training and prompt-tuning, to further enhance the system's performance.

• To the best of our knowledge, we conducted the first comparative study on different LLMs using Central Bank policy data.

1.4 Organization of the paper

The rest of the paper is structured as follows: The principle and training procedure of our method are introduced in Section II. The results of our evaluation of this Q&A system are presented in Section III. The most pertinent works are listed in Section IV. This essay is concluded in Section V.

2 Approach and implement

In this section, we detail the principal and training process of the Langchain-ChatGLM based local knowledge base automated Q&A system. The overall structure of the principle is shown in Figure 1.

2.1 Core principle

The core principle of the Langchain-ChatGLM based local knowledge base automated Q&A, system includes the following steps:

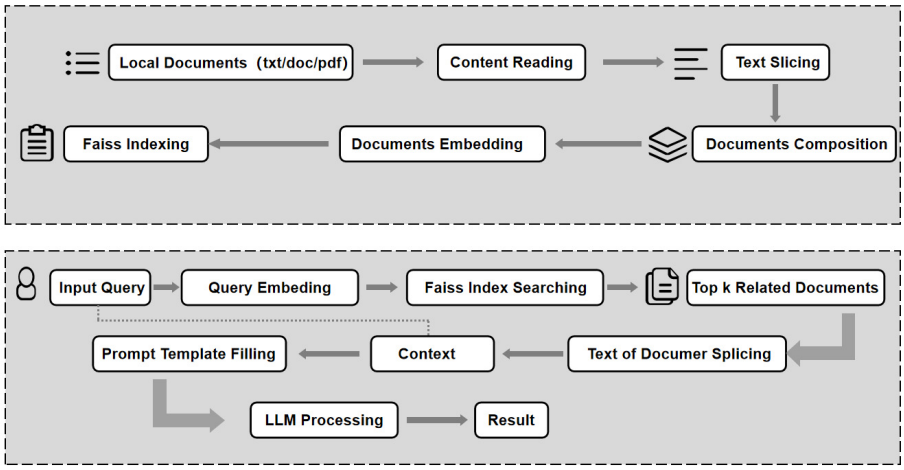

Fig. 1. Langchain-ChatGLM Principal Analysis

1. Document processing and slicing: Langchain-ChatGLM is able to parse and slice multiple academic document formats (e.g. txt, doc, and pdf), converting the content of the documents into individual documents.

2. Document embedding vector generation: Using the embedding model, each document is transformed into an embedding vector with a semantic captures the semantic features of the document.

3. Fast retrieval index construction: Langchain-ChatGLM achieves efficient document retrieval by building faiss index structure. The faiss index can efficiently store an hed organize embedding vectors of a large number of documents and support fast similarity search.

4. Query embedding vector generation: The user-supplied query also generates the corresponding embedding vectors through the embedding model, which represent the semantic information of the query.

5. Related document retrieval: With the help of the previously constructed faiss index, Langchain-ChatGLM is able to efficiently retrieve the top k documents with the highest relevance to the query, which are considered to be the most relevant documents to the query.

6. Context construction: By stitching together the text of related documents to form a context, more comprehensive information is provided to better meet the information needs of users.

7. Prompt generation and result acquisition: The context and query are populated into a predefined prompt template to generate a prompt for the language model (LLM). The prompt is passed to the LLM to obtain the results.

Based on these principles, LangChain-ChatGLM automates document processing, embedding-based index construction and retrieval of relevant documents, as well as prompt generation and result retrieval using the language model, providing efficient tools and intelligent support for academic research.

2.2 The training process

The training process of this system requires data collection and preparation, collecting data related to central bank policies, such as central bank reports and documents, from official central bank websites, public databases, government release channels and other websites to ensure that the data sources are reliable, accurate and cover a wide range of central bank policy information. The next step is to clean and preprocess the collected data, including removing noise, dealing with missing values, and unifying the format to ensure data quality and consistency.

We perform Docker deployment on the GitHub OpenI community platform[1][2] and create the project for debugging purposes. We prepare the zip files for the text2vec-chinese-base and chatglm-6b-int8 datasets and then run the app.py file. These zip files should be moved to the same folder where the master.zip file is located. Unzip the master.zip file and move it to the "langchain-chatglm-6b-webui" folder. Then, unzip the prepared dataset and modify the "share"

[1] GitHub Link: https://openi.pcl.ac.cn/mof/LangChain-ChatGLM-Webui.git
[2] GitHub Link: https://github.com/thomas-yanxin/LangChain-ChatGLM-Webui

parameter in the app.py file from “false” to “true”. Also, update the paths for text2vec-english-base and chatglm-6b-int8 in the configuration file to the respective paths where they have been unzipped. Next, downgrade gradio to version 3.10. Once the above preparations are complete, run the app.py file and obtain the URL address in the browser. Open the browser and import our preprocessed central bank policy Q&A data text document. With this, the document vectorization process will be initiated, and the initial LangChain-ChatGLM-based central bank policy automated Q&A system will be ready for operation.

2.3 Transformer

The Transformer (Figure 2) is an influential deep learning model that has significantly impacted the field of natural language processing (NLP) since its introduction by Google in 2017 [3]. It has demonstrated remarkable achievements, particularly in the domain of machine translation.

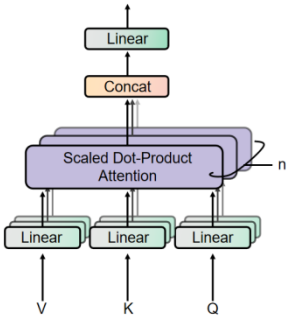

Fig. 2. Transformer

Transformers' success is driven by the self-attention mechanism, a crucial innovation that allows the model to effectively capture long-term dependencies in sequences [3]. This mechanism is critical in the encoding and decoding of textual information.

Self-attention, a fundamental building block of Transformers, replaces traditional recurrent neural networks (RNNs) and excels at capturing interdependencies across the input sequence. It allows each position within the sequence to attend to all other positions, enabling the model to consider the context and relationships among words during the encoding and decoding processes [3].

The Transformer’s innovative design, combining the self-attention mechanism and the concept of Query, Key, and Value representations [3], has revolutionized the NLP landscape, elevating the performance of a variety of tasks, including sentiment analysis, text summarization, and machine translation.

3 EVALUATION

The performance of the Langchain-ChatGLM based central bank policy Q&A system was evaluated in this section. First, we demonstrate the evaluation's experimental configuration. Then describe the performance comparison between other models and ours.

3.1 Experimental Setup

We primarily present the environment configuration and performance metrics datasets.

1. Performance Metrics: In studies pertaining to learning, we employ accurate measurements, i.e. accuracy for evaluation. In the following context, we provide definitions for this metric. Set scores for each section in the source file. The answers given by the question and answer system are scored, and the scores are used to determine accuracy.

2. Datasets: We use the Monetary Policy Implementation Report. It contains five sections, namely monetary and credit overview, monetary policy operations, financial market operations, macroeconomic analysis and monetary policy trends.

3. Environment Setup: Our experiment's hardware is set up on a multi-core server with a 16-core 2.10 GHz Intel Xeon CPU and an NVIDIA 3090 GPU for training and testing. The server has 256 GB of RAM and 24 GB of VRAM.

3.2 Methods

We compared ChatGPT, IFlytek and ChatGLM-LangChain shown in Table 1. Parameters include accuracy, response time, whether the input supports structured data, sentence logic and intelligibility, ease of use, expertise. These parameters are used to evaluate the performance of the model.

3.3 Experimental Result

The experimental results show that the central bank policy Q&A system system based on Langchain-Chatglm achieves better performance in terms of accuracy, response time, sentence logic and intelligibility. At the same time, the system also shows good robustness to answer different types of questions efficiently and can respond in a short period of time. In addition, our system has better adaptability compared to traditional rule-based or template-based Q&A systems.

From Table 1, it can be concluded that in terms of response time this quiz system performs better and is faster compared to ChatGPT. In terms of professionalism and accuracy it performs satisfactorily compared to other models and supports unstructured data. Overall, the implementation of the central bank policy based Q&A system using Langchain-ChatGLM has some practicality.

Table 1. Comparison Result

	IFlytek	ChatGPT	LangChain-ChatGLM
Response time(30)	29(20s)	23(40s)	24(30s)
accuracy(30)	25	25	24
Sentence logic and intelligibility(10)	8	9	8
ease of use (10)	10	9	9
Whether the input supports unstructured data(10)	0	10	10
Expertise (10)	9	9	8
score(100)	81	85	83

4 Related Works

4.1 Central bank monetary policy

Using the open source package R and related packages, Jonathan Benchimol et al. demonstrate how text analysis may be used to reveal the information concealed in monetary policy communication and enable consistent organization of it [4]. Hamza Bennani et al. develop a communication measure that evaluates its predisposition toward a monetary policy stance. The findings imply that clear communication is crucial for comprehending developments in monetary policy [5]. Magorzata Walerych et al. reveal that conventional monetary policies of the Fed and ECB have worldwide spillover effects on emerging market economies (EMEs) [6]. By dissecting stock market returns, Tim D. Maurer et al. uncover the economic causes of the stock market reactions of 40 nations to US monetary policy surprises [7]. In their study of the quantity and price regulations for China's monetary policy, Xiangfa Li et al. derive an open economic DSGE model [8]. Jiufeng Zhao et al. present a large-scale framework (Weak-PMLC) for multi-label policy categorization that is based on incredibly weak supervision to lessen the load of human specialists annotating numerous policies [9].

4.2 Language Model

Haifeng Wang et al. introduce the taxonomy of pre-trained models and provide a thorough analysis of current advancements and exemplary work in the field of NLP [10].

Forney, G.D conclude that the N-gram model suffers from the inability to quantify the similarity between words and the difficulty of modeling the long-distance dependency problem [11]. Peters et al. conclude Bidirectional language models are implemented with simple splicing in both directions, without deeper fusion [12]. Cho et al. conclude Encoder-Decoder models are generally large in size and require significant computing power [13].

Long Ouyang et al. consider that LLMs can do some natural language processing tasks, but these models may produce text that is different from what the user means [14]. Sandeep Reddy et al. believe that while LLMs have exhibited serious problems, such as creating false information, fabricating data, and assisting in plagiarism, they have also shown substantial potential in performing human-capable activities [15]. Meng-Lin Tsai et al. propose an LLMs-assisted problem-solving procedure and indicate that incorporating LLM into chemical

engineering education can help students deepen their understanding of core subjects [16]. JianchengYang et al. present an analytical framework for outlining the relationships between LLMs and stakeholders in medical imaging [17]. Peter Organisciak et al. suggest the Ocsai system, which fine-tunes deep neural network-based LLMs depending on human-judged answers [18].

Issues with ChatGLM-6B include: limited model capacity leading to biased answers; the potential for harmful or offensive responses; inadequate proficiency in English; susceptibility to being misled.

To address these concerns, additional knowledge can be incorporated into LLM to enable it to answer tasks based on the associated knowledge. The implementation of LangChain-ChatGLM for a central bank monetary policy question-answering system serves as an example of this approach.

5 CONCLUSION

In this paper, we address the limitations of traditional language models by developing an implementation-based Q&A system specifically designed for central bank monetary policy. Our approach involves integrating the best open-source pretrained language model in China with the LangChain framework, which leverages the capabilities of language models to provide improved answers based on a comprehensive knowledge base. The knowledge base utilized in this study is the central bank's monetary policy implementation report. We employ a process that includes loading the text, segmenting and vectorizing it, matching the most similar vectors to the question, and utilizing the language model to generate the answer. To evaluate the performance of our Q&A system, we compare it with existing models such as ChatGPT and IFlytek. The results demonstrate that our Q&A system performs satisfactorily in terms of accuracy and reasoning, matching the performance of the compared models.

Acknowledgment: This study was supported by the National Natural Science Foundation of China (62002067) and the Guangzhou Youth Talent of Science (OT20220101174)

References

[1] Yue, T., Au, C.C.: Gptquant's conversational ai: Simplifying investment research for all. Available at SSRN 4380516 (2023)

[2] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: Glm: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 320–335. Association for Computational Linguistics, Dublin, Ireland (2022)

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)

[4] Benchimol, J., Kazinnik, S., Saadon, Y.: Text mining methodologies with r: An application to central bank texts. Machine Learning with Applications 8 (2020)

[5] Fanta, N.: Does central bank communication signal future monetary policy in a (post)-crisis era? the case of the ecb (2020)

[6] Walerych, M., Wesoowski, G.: Fed and ecb monetary policy spillovers to emerging market economies. Journal of Macroeconomics 70(C), 103345 (2021)
[7] Maurer, T.D., Nitschka, T.: Stock market evidence on the international transmission channels of us monetary policy surprises. Working Papers (2020)
[8] Li, X., Wang, H.: The effective of china's monetary policy: Quantity versus price rules. The North American Journal of Economics and Finance 54 (2020)
[9] Zhao, J., Song, R., Yue, C., Wang, Z., Xu, H.: Weak-pmlc: A large-scale framework for multi-label policy classification based on extremely weak supervision.
Information Processing & Management 60, 103442 (2023)
[10] Wang, H., Li, J., Wu, H., Hovy, E., Sun, Y.: Pre-trained language models and their applications (2022)
[11] Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE 61(3), 268–278(1973) https://doi.org/10.1109/PROC.1973.9030
[12] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (2018)
[13] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder decoder for statistical machine translation. Computer Science (2014)
[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A.: Training language models to follow instructions with human feedback. arXiv e-prints (2022)
[15] Reddy, S.: Evaluating large language models for use in healthcare: A framework for translational value assessment. Informatics in Medicine Unlocked 41, 101304 (2023)
[16] Tsai, M.-L., Ong, C.W., Chen, C.-L.: Exploring the use of large language models (llms) in chemical engineering education: Building core course problem models with chat-gpt. Education for Chemical Engineers 44, 77–95 (2023)
[17] Yang, J., Li, H.B., Wei, D.: The impact of chatgpt and llms on medical imaging stakeholders: Perspectives and use cases. Meta-Radiology, 100007 (2023)
[18] Organisciak, P., Acar, S., Dumas, D., Berthiaume, K.: Beyond semantic distance: automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity 49, 101356 (2023)