Front. Comput. Sci.
https://doi.org/10.1007/s11704-024-31018-5

LETTER

RA-CFGPT: Chinese financial assistant with retrieval-augmented large language model

Jiangtong LI[1], Yang LEI[1], Yuxuan BIAN[1], Dawei CHENG[1,2], Zhijun DING[1,2], Changjun JIANG (✉)[1,2]

1 Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
2 Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China

1 Introduction

Retrieval-Augmented Generation (RAG) enhances the generative capacity of Large Language Models (LLM) by appending retrieved documents to the current context. This approach has shown success in reading comprehension [1] and language modeling [2]. RAG assumes the intent is in the input query, which can be expanded with a task description. However, in the financial domain, queries often span multiple sectors, challenging the ability of retrieval phase to adequately inform the generation phase.

The complexity of financial texts necessitates large language models (LLMs) to adeptly understand intricate financial terminology and concepts. Several financial LLMs, such as FinGPT [3], XuanYuan [4], and DISC-FinLLM [5] have emerged to meet this demand. However, most current FinLLMs prioritize pre-training or supervised fine-tuning, overlooking the knowledge-intensive nature of the financial domain and underutilizing the RAG during FinLLM training.

In this paper, we introduce an integrated system tailored for Chinese financial tasks, including but not limited to question answering, document analysis, and risk assessment. In detail, we first construct a hybrid knowledge base tailored for different aspects of the financial domain, which aims to provide robust and comprehensive information for the FinLLM. Second, we fine-tune a Chinese LLM across various financial tasks with the retrieved document from our hybrid knowledge base as background knowledge, which serves as the cornerstone in our system. Finally, we establish a system workflow to ensure the generated output not only is accurate but also meets compliance mandates and sufficiently flags associated risks.

2 Framework architecture

In this section, we will introduce our system from three aspects, 1) the hybrid financial knowledge base to provide comprehensive information; 2) the training of large language model to fit for the RAG process; 3) the system pipeline to ensure the accuracy, compliance and risk warning in output.

2.1 Hybrid financial knowledge base

Our hybrid financial knowledge base (Fig. 1) encompasses five facets: concept, company, event, industry, and legislation, ensuring comprehensive coverage in financial domain.

In the concept base, we catalog detailed explanations of various financial, economic, and general concepts as key-value databases. In the company base, we document research reports, financial statements, and basic company data. Research reports and financial statements are stored paragraph-wise in vectorized database. Basic company details, extracted from financial statements, reside in key-value database. In the event base, we archive announcements, news, and social media posts. To maximize coverage in limited storage, we prioritize and update content based on importance and post time in vectorized databases. In the industry base, we log industry reports and indicators. The reports provide macro analyses of industries and sectors and are saved in a vectorized database by paragraph. Industry indicators, extracted from these reports, are stored in a key-value database. Lastly, the legislation base encompasses laws, regulations, and policies essential for compliance mandates in vectorized databases.

During the knowledge database construction, we source documents from various websites. All vectorized databases are built using BGE-large-v1.5 [6] via the faiss [7] library. In the retrieval phase, input queries traverse all knowledge bases. During retrieval, we employ fuzzy searches for key-value databases and vector searches for vectorized ones.

2.2 Financial large language model

The cornerstone of our system is the financial large language model, designed to assist users by processing user input and background information. To specialize the LLM for the financial domain, we utilize the financial supervised fine-tuning dataset, i.e., CFData-sft [8], encompassing six financial tasks: sentiment analysis, event detection, topic

Received December 13, 2023; accepted April 8, 2024

E-mail: cjjiang@tongji.edu.cn

Special Issue—Excellent Young Computer Scientists Vision on Foundation Models

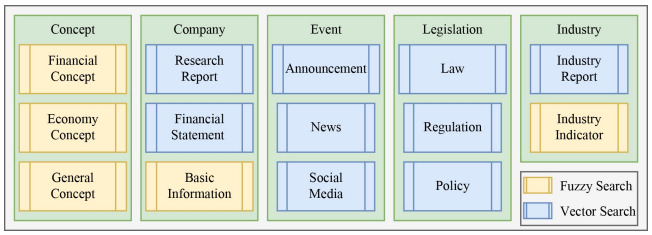

Fig. 1 The component of hybrid financial knowledge base

decomposition, report summary, question-answering, and stock prediction. To augment the FinLLM with retrieved content during training, we adopt a two-step enhancement for CFData: 1) Summarizing input content of each instance using the GPT-3.5 API; 2) Using these summaries as queries to retrieve pertinent background knowledge from our hybrid financial database. We fine-tune CFGPT1-pt [8] on the retrieval-augmented dataset to get our RA-CFGPT. To harmonize the capabilities of LLM across general and financial domains, we incorporate the Moss-03-sft dataset [9] during fine-tuning.

During fine-tuning, we follow the similar operation and hyper-parameter setting in Li et al. [8], leading to RA-CFGPT-LoRA and RA-CFGPT-Full, where the learning rate is 2e-4 with a global batch size of 256. In the LoRA setup, the LoRA rank and alpha is set as 64 and 16. The target modules for LoRA are set to all weight matrices except for the token embeddings. During fine-tuning, we first train on general instructions and then train on financial instructions.

2.3 System pipeline

Upon receiving user input, our system first retrieves relevant background knowledge from a hybrid financial database. This input, along with conversation history and background information, is organized using a predefined prompt template. The organized input is then fed into our Retrieval-Augmented CFGPT (RA-CFGPT) to generate a response. This response subsequently undergoes a series of checks for evidence, compliance, and risk. If the response successfully passes all these checks, it is then displayed on the conversation interface. In cases where the response fails to meet the required standards, RA-CFGPT is prompted to re-generate the response, incorporating the feedback received. This re-generation process can occur up to three times before the system defaults to displaying an "I don't know" message on the conversation page.

The evidence checker is designed to verify the accuracy of the response against the background. The compliance checker ensures that the response adheres to relevant legislative items in background. Meanwhile, the risk checker evaluates whether risk warnings are marked in response. To develop these checkers, we first collect data using our RA-CFGPT. Subsequently, we transform these tasks into different binary classification tasks for data annotation with GPT-3.5 API. We then split the annotated data with the ratio 8:2 to train and validate these three checking modules as three binary classifiers, providing robust layers of validation for the responses of system.

3 Experiment

In Table 1, we present the results for the effectiveness of retrieval-augmented training and system design, as assessed on CFBenchmark [10]. This benchmark evaluates LLMs in financial domain across three key aspects: entity recognition, text classification, and content generation. The experimental results indicate a significant enhancement to the original CFGPT across all tasks due to the retrieval-augmented training. This training approach yields notable improvements in text classification and content generation. This is attributed to the retrieval step, which provides additional information. Furthermore, RA-CFGPT system enhances performance in financial content generation. The inclusion of checkers ensures that the outputs are not only accurate and compliant but highlight potential risks, which is crucial for the quality and reliability of responses in financial contexts.

4 Conclusion

In this paper, we present the RA-CFGPT system, designed for Chinese financial tasks such as question answering, document analysis, investment advising, and risk assessment. Our system combines a hybrid knowledge base, a fine-tuned Chinese FinLLM, an information organizer, and response checkers to ensure outputs are accurate, compliant, and highlight potential risks. Experimental result reveals the

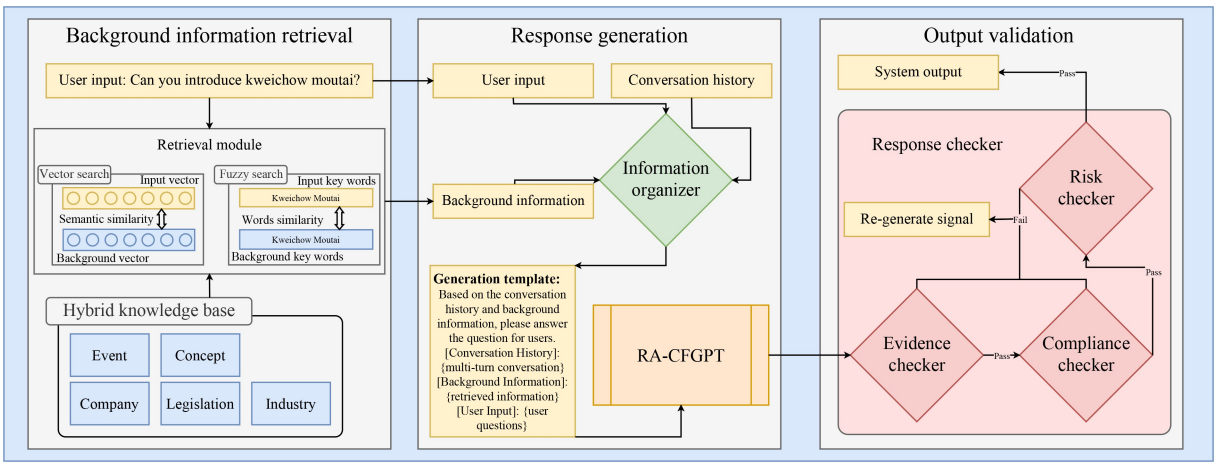

Fig. 2 The overall framework of RA-CFGPT system

Table 1 The results on CFBenchmark. The experiment results of baseline methods are from CFBenchmark [10]. R.avg, C.avg, and A.avg are the average results of entity recognition, text classification, and content summarize task. RA-CFGPT-Full + Sys indicates that we equipped the RA-CFGPT-Full model with three checking modules. Best results are highlighted in boldface

Model	Size	Company	Product	R.Avg	Sector	Event	Sentiment	C.Avg	Summary	Risk	Suggestion	A.Avg	Avg
Human	–	0.931	0.744	0.838	0.975	0.939	0.912	0.942	1.000	1.000	1.000	1.000	0.927
ChatGPT	175B	0.797	0.198	0.498	0.453	0.458	0.425	0.455	0.593	0.541	0.771	0.635	0.529
ERNIE-Bot-4	–	0.819	0.417	0.618	0.418	0.358	0.375	0.384	0.721	0.629	0.718	0.689	0.564
Qwen-Chat-7B	7B	0.763	0.360	0.562	0.400	0.367	0.265	0.344	0.548	0.307	0.379	0.411	0.439
ChatGLM2-6B	6B	0.747	0.313	0.530	0.285	0.300	0.357	0.314	0.657	0.454	0.671	0.594	0.479
Baichuan2-7B-Chat	7B	0.757	0.402	0.579	0.425	0.475	0.323	0.408	0.725	0.648	0.732	0.702	0.563
DISC-FinLLM	13B	0.801	0.357	0.579	0.481	0.512	0.482	0.492	0.728	0.611	0.702	0.680	0.583
CFGPT-stf-LoRA	7B	0.820	0.414	0.617	0.569	0.729	0.769	0.689	0.745	0.584	0.609	0.646	0.650
CFGPT-sft-Full	7B	0.836	0.476	0.656	0.700	0.808	0.829	0.779	0.798	0.669	0.808	0.758	0.731
RA-CFGPT-LoRA	7B	0.828	0.421	0.624	0.602	0.763	0.801	0.722	0.762	0.608	0.693	0.688	0.678
RA-CFGPT-Full	7B	**0.853**	**0.492**	**0.672**	**0.731**	**0.841**	**0.851**	**0.808**	0.821	0.692	0.829	0.781	**0.754**
RA-CFGPT-Full+Sys	7B	–	–	–	–	–	–	–	**0.838**	**0.721**	**0.841**	**0.800**	–

effectiveness of our retrieval-augment training and RA-CFGPT system.

Acknowledgements The work was supported by the National Key R&D Program of China (2022YFB4501704) and the Shanghai Science and Technology Innovation Action Plan Project (22YS1400600 and 22511100700)

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Lee K, Chang M W, Toutanova K. Latent retrieval for weakly supervised open domain question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 6086−6096
2. Izacard G, Lewis P, Lomeli M, Hosseini L, Petroni F, Schick T, Dwivedi-Yu J, Joulin A, Riedel S, Grave E. Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 2023, 24(251): 1−43
3. Yang H, Liu X Y, Wang C D. FinGPT: Open-source financial large language models. 2023, arXiv preprint arXiv: 2306.06031
4. Zhang X, Yang Q. XuanYuan 2.0: A large Chinese financial chat model with hundreds of billions parameters. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, 4435−4439
5. Chen W, Wang Q, Long Z, Zhang X, Lu Z, Li B, Wang S, Xu J, Bai X, Huang X, Wei Z. DISC-FinLLM: A Chinese financial large language model based on multiple experts fine-tuning. 2023, arXiv preprint arXiv: 2310.15205
6. Xiao S, Liu Z, Zhang P, Muennighoff N. C-pack: Packaged resources to advance general Chinese embedding. 2023, arXiv preprint arXiv: 2309.07597
7. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 2021, 7(3): 535−547
8. Li J, Bian Y, Wang G, Lei Y, Cheng D, Ding Z, Jiang C. CFGPT: Chinese financial assistant with large language model. 2023, arXiv preprint arXiv: 2309.10654
9. Sun T, Zhang X, He Z, Li P, Cheng Q, Yan H, Liu X, Shao Y, Tang Q, Zhao X, Chen K, Zheng Y, Zhou Z, Li R, Zhan J, Zhou Y, Li L, Yang X, Wu L, Yin Z, Huang X, Qiu X. Moss: Training conversational language models from synthetic data. 2023 **(查阅网上资料,未找到本条文献信息,请确认)**
10. Lei Y, Li J, Jiang M, Hu J, Cheng D, Ding Z, Jiang C. CFBenchmark: Chinese financial assistant benchmark for large language model. 2023, arXiv preprint arXiv: 2311.05812