DOI: 10.53469/jtpes.2024.04(03).16

Application of Conversational Intelligent Reporting System Based on Artificial Intelligence and Large Language Models

Hong Zhou[1,*], KangMing Xu[2], Qiaozhi Bao[3], Yan Lou[4], Wenpin Qian[5]

[1]Computer Technology, Peking University, Beijing, CN
[2]Computer Science and Engineering, Santa Clara University, CA, USA
[3]Statistics,North Carolina State University, NC, USA
[4]Software Engineering,Illinois Institute of Technology, Va, USA
[5]Information Science ,Trine University,Phoenix, Arizona, USA
[*]*Corresponding author: Hong Zhou, E-mail:zh.robot@pku.edu.cn*

Abstract: ***As large language models gain traction in the financial sector, they are revolutionizing the workflows of financial professionals. From data analysis and market forecasting to risk assessment and customer management, these models demonstrate significant potential and value. By automating data processing tasks, they enhance productivity and empower professionals to derive deeper insights and make more precise decisions. This article explores the application of conversational intelligent reporting systems, leveraging artificial intelligence and large language models, within the financial industry. It examines how these systems streamline reporting processes, facilitate efficient communication, and contribute to informed decision-making, ultimately reshaping the landscape of financial market operations.***

Keywords: Large language models; Financial sector; Conversational intelligent reporting systems; Decision-making.

1. INTRODUCTION

A Large language model (LLM) is an artificial intelligence system based on deep learning techniques to understand, process, and generate human natural language. It is a major breakthrough technology in the field of artificial intelligence. Based on the training of a large number of dense text data, it trains and learns relevant statistical relationships from text documents through self-supervised and semi-supervised learning to achieve the understanding and generation of human natural language.

In terms of technical principles, the large language model is mainly based on the Transformer architecture, which was first proposed by Google at the NeurIPS conference in 2017. It is a neural network architecture designed specifically for processing sequence data, and its core is the self-attention mechanism. The self-attention mechanism enables the model to process a word by taking into account all the words in the input sequence, by assigning different attention weights to different parts of the input sequence, so as to more accurately capture the relationship between words and language context.

Moreover, during the training process, the large language model (LLM) based on the converter architecture converts the text into numerical form through the decoding part of its encoder, thus capturing the similar meanings of words and phrases, the context, the parts of speech and other language features [1-3]. Next, the LLM will use this acquired language knowledge via a decoder to produce a unique text output based on its own parameters. In terms of specific flow, the operation flow of large-scale language model is divided into several steps: accepting input text, encoding text, decoding generation and output prediction. Before the LLM can begin to receive inputs and produce predictive outputs, it goes through a pre-training period during which the LLM learns to perform basic language functions, followed by a fine-tuning phase with specific constraints.

2. RELATED WORK

2.1 Introduction to large language models

On the basis of semantic understanding, large language model has a strong ability of natural language reasoning, decision making and generation, which is embodied in the ability and behavior of text content summary, induction, sorting, contrast, summary, transformation, creation and analysis [4-6]. This comes from the improvement of large language model learning algorithms and the input and "feeding" of massive training data, and the continuous emergence of large models with hundreds of billions or even trillions of parameters, which gives large models the ability to "think" in a certain sense. In short, large models possess a considerable degree of human-like ability to "understand" and "think." Therefore, it is natural to think of such a smart "brain" to collaborate with existing enterprise applications, so as to achieve a leapfrog upgrade in the degree of intelligence of enterprise applications. Some typical landing scenarios of large language models in enterprise applications are as follows [7-8]:

Build applications with natural language as the interactive interface, and combine mature speech/image recognition, speech synthesis and other technologies to improve customer experience in marketing, customer service and other fields. For example: intelligent customer consultation, intelligent call center, intelligent sales assistant and so on

Analyze call center conversation records with the help of text extraction and analysis capabilities of large language model, complete quality check, customer sentiment analysis and classification, hot issue extraction, and sales lead discovery.

Build a natural language-based enterprise content search engine and portal. Compared with traditional search solutions, large models can help better optimize and summarize search results on the basis of semantic understanding and improve search experience. And it can be further integrated with other enterprise applications, for example, in a typical marketing process, product descriptions can be automatically searched and sent to customers based on customer conversations.

2.2 LLM in user interaction applications

The application of large language model in the user interaction stage mainly focuses on conversational recommendation, aiming to break the traditional paradigm of single-round recommendation item display, accurately capture and model the user's conversational intention by using text as a bridge, and provide users with a more intelligent and accurate interactive recommendation process [9]. The application of LLM in the user interaction stage can be roughly divided into the following two categories:

Task-oriented user interaction. In this scenario, we will assume that the user has a clear goal of "seeking recommendations", and the recommendation system needs to support the user's thinking and decision-making process, and ultimately assist the user in finding relevant items. Specifically, the large language model mostly exists as a sub-module of the recommendation system, which is used to analyze the user's goal intention and build the user portrait within the conversation [10-11].

Open conversational user interaction. Compared with the previous scenario, this kind of scenario mostly assumes that the user's behavioral intention is diverse, changing and vague, and the system needs to gradually acquire the user's interest or guide the user through interaction (including topic dialogue, small talk, question and answer, etc.) to finally achieve the purpose of recommendation. Therefore, the large language model is usually used as the core module of this kind of recommendation system to complete a series of behaviors from dialogue, question and answer to recommendation at the same time with a unified text medium [12-14].

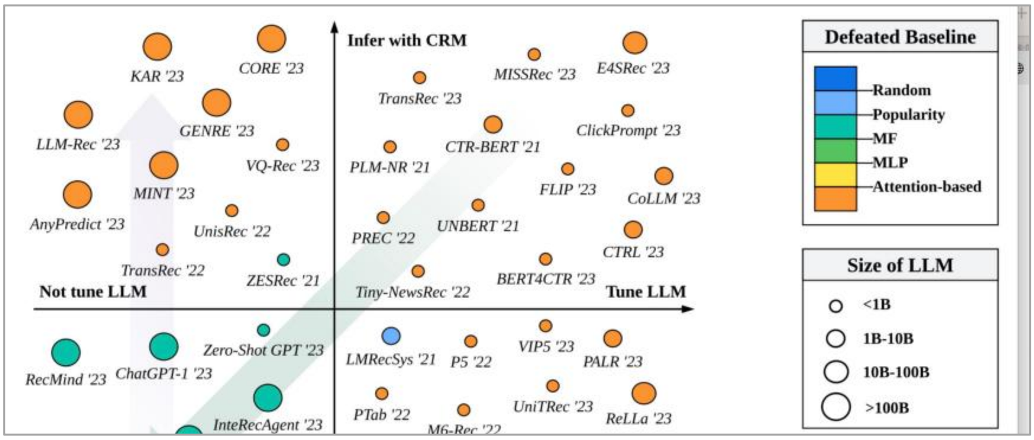

Figure 1: Interactive application of large language models

2.3 Large language model dialog system

After OpenAI launched the ChatGPT dialog system, it exceeded one million users in five days, and reached 100 million users in two months, becoming the fastest growing consumer application in history [15]. The emergence of ChatGPT demonstrates the powerful potential of large language models for human-computer dialogue. However, as a commercial product, OpenAI does not disclose key details of ChatGPT, which makes it difficult for researchers to fully understand how it works [16]. Replicating ChatGPT has become a target for many researchers and Internet companies. The researchers focused on InstructGPT proposed by OpenAI before ChatGPT appeared, believing that InstructGPT contained the key technology of ChatGPT. InstructGPT, on the other hand, uses the SFT and RLHF mentioned in the previous chapter to respond to human instructions. To train a large language model that can follow human instructions like InstructGPT, there are three core steps:

(1) Fine-tune a good SFT dataset with a very strong base model

(2) Get feedback scores based on multiple outputs of the fine-tuned model and train an RW model

(3) Use the RW model to fine-tune the language model

It can be seen that the dialogue language model first needs a strong base model, and the key difficulty of the subsequent RW model and RLHF fine-tuning is how to obtain a high enough quality data set.

3. METHODOLOGY

3.1 DiagGPT Framework

DiagGPT is a multi-agent and collaborative system consisting of several modules, of which the Topic Manager is particularly important, which tracks the conversation state and automatically manages the conversation topics. Each module is an LLM with specific prompts that guide its functions and responsibilities.

DiagGPT's workflow consists of four phases: (1) Thinking about topic development: The topic manager takes user queries and analyzes and predicts topic development for the current conversation round; (2) Maintain the topic stack: Maintain the topic stack of the entire dialogue according to the operation command of the topic manager; (3) Enrich the topic: Retrieve the current topic according to the context of the conversation and enrich it; (4) Generate replies: Generate replies for users according to specific guidance prompts, combined with rich topics and context. Defining the topic as the main topic of the conversation identifies the main focus of the communication. Define a task as a specific goal that needs to be accomplished in a task-oriented conversation. This particular task should be completed after passing through all the predefined topics in the conversation.

3.2 Thinking Topic Development

DiagGPT's main module is the Topic Manager, which is responsible for determining topic development based on user queries. Before each round of conversation, the system needs to adjust the current topic. Therefore, the user's query is first entered into the Topic Manager.

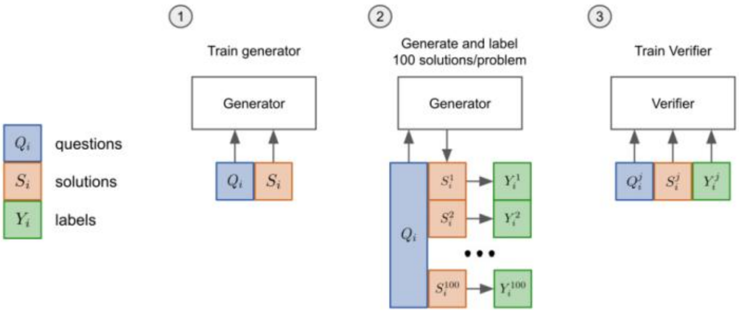

Figure 2: DiagGPT model architecture diagram

Topic Manager is an AI agent that analyzes chat history and predicts topic development. It has an action list with various tools to perform the task. Each action corresponds to a program function, implemented using Python's decorator function [17]. When the Topic Manager receives a user query, it analyzes all the available information and decides which action to perform based on the prompts for each action.

The LLM has strong understanding and reasoning capabilities to accurately understand user intent and effectively communicate with users.

3.3 Maintaining Topic Stack

After the system takes the action output from the theme manager, it executes the corresponding commands to process and control theme changes, including maintaining the theme stack.

The topic stack in an AI system is a data structure used to store and track the state of a conversation. The progression of a conversation is considered to have multiple phases, or states, and these states occur in a first-in-first-out (FIFO) order that can be effectively modeled using the stack [18-19].
Consultants often have a list in mind when diagnosing, and if the user doesn't ask a new question, the conversation stalls.

The consultant will check the checklist and provide a report and comprehensive analysis. You can simplify this process by loading topics for predefined tasks. When the function prompted for this action is executed, the topic in the list is loaded into the topic stack.

In addition, there are other common operations to manipulate the theme stack. These actions include creating a new topic, completing the current topic, and keeping the current topic [20]. The Create New theme action adds a new theme to the stack when the user wants to start a new theme. Removes the top topic from the stack when the user no longer wishes to discuss the topic or when the system considers the topic closed. The Keep Current topic action indicates that the system has determined that information is still needed and that it needs to continue discussing the current topic, so no changes will occur to the topic stack. These three basic actions cover most theme change scenarios.

3.4 New evaluation method iEvaLM of conversational recommendation system

Considering the problems of existing evaluation methods, we propose a new evaluation method-iEvaLM, which adopts interactive evaluation based on LLM user simulation. Our measurement methods integrate seamlessly with existing CRS datasets, and the system-user interaction extends to every human-labeled conversation. The key idea of our approach is to simulate real users based on LLMs' [21-24] excellent "role playing" capabilities. We treat labeled items as user preferences and let LLM simulate the role of the user by way of instruction Settings. Through the interaction, we can not only measure the accuracy of the recommendation by comparing the predicted item to the label, but also measure the interpretability of the generated interpretation

3.5 through the LLM-based rater

(1) Attribute based question and answer

The behavior of the system is limited to choosing one of k predefined attributes to ask or recommend to the user. In each round, we first ask the system to select one of these k+1 options, and then the user responds with a template-based response [25]: answer the question with the properties of the target item or provide feedback on the system's recommendations. An example of an interactive round might look like this:

Which genre do you like?

Which genre do you like?

(2) Free talk

This type has no restrictions on interaction, and both the system and the user are free to initiate. An example of an interactive round might look like this:

Do you have any specific genre in mind?

Do you have any specific genre in mind?

(3) To support interaction with the system, we use LLMs for user simulation. A simulated user can adopt one of three behaviors:

Talk about preferences. When the system clarifies or asks about the user's preferences, the simulated user answers with information about the target item.

Provide feedback. When the system recommends a list of items, the simulated user checks each item, providing positive feedback if the target item is found, and negative feedback if not.

Complete the conversation [26]. If the system recommends exactly one target item, or if the interaction reaches a certain number of rounds, the simulated user ends the conversation.

Specifically, we use the text-davinci-003 model and hand-constructed instructions to build a realistic user image. In these instructions, we first fill the template with real objects and then define their behavior using a set of hand-constructed rules.

4. CONCLUSION

In conclusion, the integration of large language models (LLMs) in the financial sector marks a significant advancement in leveraging artificial intelligence for improved decision-making processes. The transformative potential of LLMs extends across various domains within the financial industry, from data analysis and market forecasting to risk assessment and customer management [27-28]. By automating complex data processing tasks, LLMs not only enhance productivity but also enable financial professionals to gain deeper insights and make more precise decisions. Moreover, the application of conversational intelligent reporting systems, driven by LLMs, streamlines reporting processes and facilitates efficient communication, ultimately reshaping the landscape of financial market operations [29-30].

Moving forward, it is imperative for financial institutions to continue exploring and harnessing the capabilities of LLMs to stay competitive in an increasingly digitized and data-driven environment. Further research and development efforts should focus on optimizing LLM-based systems for specific financial applications, improving their real-time performance, and ensuring robustness and reliability in decision-making processes. By embracing the potential of LLMs and conversational intelligent reporting systems, the financial industry can unlock new opportunities for innovation, efficiency, and strategic growth in the years to come.

5. ACKNOWLEDGEMENT

In the course of this research, we are deeply inspired and guided by many literatures, which provide us with valuable knowledge base and research direction. We are especially grateful to the authors of all these studies, whose work not only provided a solid foundation for our research, but also led us to explore new ways to integrate AI and machine learning in DevOps to enhance security. We recognize that our work would not stand without the research and discoveries of those who came before us. Therefore, we have deep gratitude and respect for all the researchers cited in this article and their contributions. This literature not only enriches our research, but also guides us to avoid duplicating the work of others, allowing us to focus on the cutting edge of innovation and research.

In particular, we thank the researchers who have done pioneering work in artificial intelligence, machine learning, and DevOps security. Their research points us in the right direction and allows us to contribute to this important interdisciplinary field.

Finally, we hope that through our research, we can make our contribution to this field and make progress together with our colleagues in the academic community. We look forward to continuing to explore and extend the boundaries of knowledge in this area in future work.

REFERENCES

[1] Liang, Penghao, et al. "Enhancing Security in DevOps by Integrating Artificial Intelligence and Machine Learning." Journal of Theory and Practice of Engineering Science 4.02 (2024): 31-37.

[2] Chen, Jianhang, et al. "One-stage object referring with gaze estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[3] "Unveiling the Future Navigating Next-Generation AI Frontiers and Innovations in Application". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 147-56, https://doi.org/10.62051/ijcsit.v1n1.20.

[4] He, Yuhang, et al. "Intelligent Fault Analysis With AIOps Technology". Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Feb. 2024, pp. 94-100, doi:10.53469/jtpes.2024.04(01).13.

[5] Su, Jing, et al. "Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review." arXiv preprint arXiv:2402.10350 (2024).

[6] Wang, Yong, et al. "Construction and application of artificial intelligence crowdsourcing map based on multi-track GPS data." arXiv preprint arXiv:2402.15796 (2024).

[7] Chen, J. (2022). The Reform of School Education and Teaching Under the "Double Reduction" Policy. Scientific and Social Research, 4(2), 42-45. (Feb 2022)

[8] Zhou, Y., Osman, A., Willms, M., Kunz, A., Philipp, S., Blatt, J., & Eul, S. (2023). Semantic Wireframe Detection.

[9] Ji, Huan, et al. "Utilizing Machine Learning for Precise Audience Targeting in Data Science and Targeted Advertising." Academic Journal of Science and Technology 9.2 (2024): 215-220.

[10] Qian, Wenpin, et al. "Clinical Medical Detection and Diagnosis Technology Based on the AlexNet Network Model." Academic Journal of Science and Technology 9.2 (2024): 207-211.

[11] Zhu, Mingwei, et al. "Enhancing Collaborative Machine Learning for Security and Privacy in Federated Learning." Journal of Theory and Practice of Engineering Science 4.02 (2024): 74-82.

[12] Yang, Le, et al. "Research and Application of Visual Object Recognition System Based on Deep Learning and Neural Morphological Computation." International Journal of Computer Science and Information Technology 2.1 (2024): 10-17.

[13] Zhang, Y., & Zhang, H. (2023). Enhancing robot path planning through a twin-reinforced chimp optimization algorithm and evolutionary programming algorithm. IEEE Access.

[14] Qian, Jili, et al. "A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2." International Journal of Computer Science and Information Technology 2.1 (2024): 28-35.

[15] Bao, Qiaozhi, et al. "Exploring ICU Mortality Risk Prediction and Interpretability Analysis Using Machine Learning." (2024).

[16] Zhang, Y., Gono, R., & Jasiński, M. (2023). An Improvement in Dynamic Behavior of Single Phase PM Brushless DC Motor Using Deep Neural Network and Mixture of Experts. IEEE Access.

[17] Zhu, Mengran, et al. "THE APPLICATION OF DEEP LEARNING IN FINANCIAL PAYMENT SECURITY AND THE CHALLENGE OF GENERATING ADVERSARIAL NETWORK MODELS." The 8th International scientific and practical conference "Priority areas of research in the scientific activity of teachers"(February 27–March 01, 2024) Zagreb, Croatia. International Science Group. 2024. 298 p.. 2024.

[18] Duan, Shiheng, et al. "Prediction of Atmospheric Carbon Dioxide Radiative Transfer Model Based on Machine Learning". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 132-6, https://doi.org/10.54097/ObMPjw5n.

[19] Chen , Jianfeng, et al. "Implementation of an AI-Based MRD Evaluation and Prediction Model for Multiple Myeloma". Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 127-31, https://doi.org/10.54097/zJ4MnbWW.

[20] "Machine Learning Model Training and Practice: A Study on Constructing a Novel Drug Detection System". International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 139-46, https://doi.org/10.62051/ijcsit.v1n1.19.

[21] Duan, Shiheng, et al. "THE INNOVATIVE MODEL OF ARTIFICIAL INTELLIGENCE COMPUTER EDUCATION UNDER THE BACKGROUND OF EDUCATIONAL INNOVATION." The 2nd International scientific and practical conference "Innovations in education: prospects and challenges of today"(January 16-19, 2024) Sofia, Bulgaria. International Science Group. 2024. 389 p.. 2024.

[22] Gong, Yulu, et al. "RESEARCH ON A MULTILEVEL PRACTICAL TEACHING SYSTEM FOR THE COURSE'DIGITAL IMAGE PROCESSING." OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 272.

[23] W. Sun, W. Wan, L. Pan, J. Xu, and Q. Zeng, “The Integration of Large-Scale Language Models Into Intelligent Adjudication: Justification Rules and Implementation Pathways”, Journal of Industrial Engineering & Applied Science, vol. 2, no. 1, pp. 13–20, Feb. 2024.

[24] Zhou, Yanlin, et al. "Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functional Phenotypes." Journal of Theory and Practice of Engineering Science 4.02 (2024): 45-51.

[25] Shen, Zepeng, et al. "EDUCATIONAL INNOVATION IN THE DIGITAL AGE: THE ROLE AND IMPACT OF NLP TECHNOLOGY." OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 281.

[26] Zhang, Y., Abdullah, S., Ullah, I., & Ghani, F. (2024). A new approach to neural network via double hierarchy linguistic information: Application in robot selection. Engineering Applications of Artificial Intelligence, 129, 107581.

[27] Zhang, Chenwei, et al. "SegNet Network Architecture for Deep Learning Image Segmentation and Its Integrated Applications and Prospects." Academic Journal of Science and Technology 9.2 (2024): 224-229.

[28] Wang, Yong, et al. "Autonomous Driving System Driven by Artificial Intelligence Perception Fusion." Academic Journal of Science and Technology 9.2 (2024): 193-198.

[29] Zhang, Quan, et al. "Application of the AlphaFold2 Protein Prediction Algorithm Based on Artificial Intelligence." Journal of Theory and Practice of Engineering Science 4.02 (2024): 58-65.

[30] Qian, Wenpin, et al. "NEXT-GENERATION ARTIFICIAL INTELLIGENCE INNOVATIVE APPLICATIONS OF LARGE LANGUAGE MODELS AND NEW METHODS." OLD AND NEW TECHNOLOGIES OF LEARNING DEVELOPMENT IN MODERN CONDITIONS (2024): 262.