Effective Tool Augmented Multi-Agent Framework for Data Analysis

Xilin Zhang[1], Zhixin Mao[2], Ziwen Chen[2], Shen Gao[2†]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]School of Computer Science and Engineering (School of Cyber Security), University of Electronic Science and Technology of China, Chengdu 611731, China

Keywords: Exploratory data analysis; Multi-agent; Large language models

Citation: Zhang X.L., Mao Z.X., Chen Z.W., et al.: Effective Tool Augmented Multi-Agent Framework for Data Analysis. Data Intelligence 6(4), 923-945 (2024). doi: https://doi.org/10.3724/2096-7004.di.2024.0013
Submitted: March 24, 2024; Revised: April 10, 2024; Accepted: May 28, 2024

ABSTRACT

Data analysis tasks aim to provide insightful analysis for given data by incorporating background knowledge of the represented phenomenon, which in turn supports decision-making. While existing large language models (LLMs) can describe data trends, they still lag behind human data analysts in terms of integrating external knowledge and in-depth data analysis. Therefore, we propose a multi-agent data analysis framework based on LLMs. The framework decomposes the data analysis task into subtasks by employing three different agents. By empowering agents with the ability to utilize data search tools, the framework enables them to search for arbitrary relevant knowledge during the analysis process, leading to more insightful analysis. Moreover, to enhance the quality of the analysis results, we propose a multi-stage iterative optimization method that iteratively performs data analysis to form more in-depth conclusions. To validate the performance of our framework, we apply it to three real-world problems in the research development of higher education in China data. Experimental results demonstrate that our approach can achieve more insightful data analysis results compared to directly using LLMs alone.

1. INTRODUCTION

In recent years, there has been a notable progression in large language models (LLMs). Starting from GPT-2, advancements have led to larger models like GPT-3, GPT-4 [1], Mistral [2], and GLM [3]. These models are evolving rapidly and have significantly improved natural language processing (NLP) tasks,

† Corresponding author: Shen Gao (E-mail: shengao@pku.edu.cn; ORCID: 0000-0003-1301-3700).

© 2024 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

including translation [4, 5, 6, 7, 8], summarization [9, 10, 11, 12, 13, 14], dialogue [15, 16, 17, 18, 19, 20, 21] etc. This development highlights the growing importance and potential of LLMs in a wide range of applications in general artificial intelligence. LLMs represent cutting-edge artificial intelligence systems that have demonstrated remarkable progress in the field of NLP over the past few years. These models are distinguished by their capacity to learn intricate patterns and structures from extensive datasets comprising diverse textual information. Through pretraining on colossal corpora, LLMs acquire a deep understanding of language, enabling them to generate coherent and contextually relevant text that closely resembles human speech.

However, directly employing LLMs proves challenging in addressing complex tasks, such as creative writing [22] or software development [23]. In real-world scenarios, accomplishing such tasks typically necessitates collaboration among teams comprising multiple domain experts. Consequently, some research endeavors have explored the integration of LLMs roles as multiple domain experts, facilitating their collaboration through natural language communication. Existing works [24, 25, 26] in the realm of multi-agent systems have predominantly adopted a roleplaying approach, assigning LLMs a role name and general task objectives without endowing them with the essential knowledge, tools, and capabilities required to fulfill the tasks.

Data analysis tasks represent one of the most common needs in our lives [27]. To achieve insightful data analysis, it is imperative not only to analyze trends within the data itself but also to combine domain knowledge and relevant experience to gain insights into patterns and identify underlying issues reflected in the data. These data analysis results can better assist users in taking further actions. Therefore, when it comes to combining domain knowledge for data analysis tasks, inspired by the existing works, we propose to utilize the LLM-based multi-agent system to achieve insightful data analyses. However, since these methods lack the necessary domain knowledge and the ability to employ domain-specific tools, conventional LLM-based multi-agent approaches fall short of achieving insightful analyses.

Therefore, we propose the Tool Augmented Multi-Agent Framework, aimed at conducting more in-depth data exploration based on specified databases or data sources. First, we designed the **Reasoner module** for specific data preprocessing. It contains a central LLM responsible for extracting relevant keyword information from additional data. Macro events often have more complex causations, which may go beyond direct surface-level data. To solve this complex problem, we introduce the **Verifier module** which is responsible for exploring more external knowledge sources and combining the results from the Reasoner's output. To integrate more external knowledge, we propose to integrate the search tools for the Verifier module. Subsequently, the LLM analyzes the search data and combines the new information from the search result, and creates a detailed summary.

Since achieving a deeper understanding of information requires continuous data exploration by the Reasoner and the Verifier, a single analysis and search is not enough. Thus, we propose a **multi-stage iterative optimization** method. To enhance the Reasoner's ability for in-depth analysis, we find that more related information is an important factor affecting the quality of the output of our data analysis system. To

this end, we introduce a related information database to record the highly relevant evidence, taken from the output of the Verifier. Compared with only inputting the original data trend, using related information as input will greatly improve the prediction ability of Verifier. Thus, we use the raw data and the related information database as inputs of the Verifier module, helping this module to provide key information that is considered causal for data fluctuations. Through this iterative process, the LLM systematically investigates the problem, explains new keywords, and seeks the discovery of emergent evidence.

After conducting each iteration of analysis, detailed feedback can help enhance the quality of generated data analysis. Therefore, we introduce the **Evaluator module**. The Evaluator module aims at understanding the evidence, evaluating the generated analysis, and creating conclusions. After an iteration, the LLM combines the original data with the feedback of the Evaluator module as the input of the next iteration. Finally, the LLM selects the top-k evidence based on scores and generates the data analysis based on the retrieved evidence.

We evaluate the performance of our proposed framework by addressing several representative questions derived from several real-world data analysis scenarios. The first question focuses on analyzing changes in the number of students in higher education institutions in Beijing. The second question explores variations in China's research capabilities, while the third question examines fluctuations in the number of research and development institutions. Since these three questions not only require in-depth analysis of data trends but also need to be combined with other relevant background knowledge to draw in-depth insights, we employ these three questions to evaluate the performance of our proposed framework. Compared with directly using strong LLMs, extensive experiments conducted on several real-world datasets demonstrate the effectiveness of our proposed multi-agent data analysis framework.

Our contributions are as follows:

1. We propose a tool-augmented multi-agent framework for data analysis generation.
2. We introduce a tool-augmented LLM-based agent method to incorporate external knowledge for data analysis.
3. We propose a multi-agent framework with parallel optimization to generate insightful analysis of the given data.
4. Experimental results on several real-world data analysis questions illustrate the advantage of our proposed method in terms of the accuracy of analysis.

2. RELATED WORK

The paper presents a tool-augmented multi-agent framework designed to tackle the challenges inherent in Exploratory Data Analysis (EDA). Our primary objective is to leverage the outstanding problem-solving capabilities of LLMs to refine and optimize the EDA process, especially for complex and multifaceted problems. In the following section, we will briefly describe several directions closely associated with our work.

2.1 Large Language Model

Emerging as powerful tools in natural language processing, Large Language Models (LLMs) [17, 28, 29, 30] have been increasingly prompted to solve real-world tasks such as decision making, reasoning, and tool usage. Their ability to excel in completing complex tasks has garnered considerable attention from scholars, prompting a concerted effort to explore and expand upon their potential capabilities. This pursuit involves delving deeper into the underlying mechanisms of LLMs, identifying areas for enhancement, and devising novel methodologies to leverage their strengths effectively.

2.1.1 LLMs for Reasoning

Reasoning, a crucial component of human intelligence, has gained great concern in LLMs. Chain-of-Thought (CoT) [31], an insightful work along this direction, focuses on prompting LLMs to generate a series of intermediate reasoning procedures for multi-step problems. Many follow-up works have been proposed consecutively, including zero-shot-CoT [32], least-to-most prompting [33], and reasoning with self-consistency [34]. SC samples multiple reasoning chains and selects the majority answer among them. However, MCR [35] adopts an approach named meta-reasoning instead of aggregating through a voting mechanism, which is closely tied to self-reflection. Considering interaction with the external environment, ReAct [36] combines reasoning and acting with language models to enhance the trustworthiness of answers. We leverage the reasoning capabilities of LLMs for the analysis and distillation of key information from data, thereby facilitating subsequent information retrieval processes.

2.1.2 Retrieval-Augmented LLMs

Although LLMs have achieved great success in various NLP tasks, the knowledge generated by these models may inevitably be incomplete, incorrect, outdated, or even fabricated. This considerably facilitates the development of retrieval-augmented language models. There has been a line of work exploring how to utilize external evidences to assist LLMs. RR [37] is designed, searching relevant knowledge based on the decomposed reasoning steps obtained from CoT, for more faithful explanations. IRCoT [38] interleaves CoT generation with knowledge retrieval steps, while in-context RALM [39] enhances outputs by adding retrieved content to the input. Gao [40] proposed an approach named RARR, which utilizes retrieval to revise the correctness of input text and preserve the original input as much as possible through post-editing. Although large language models are credible in discussing a wide range of topics, their repository of specialized knowledge in specific domains remains not fully comprehensive. Consequently, we employ external interfaces to retrieve pertinent content, serving as a supplement to the internal knowledge base of large language models.

2.2 LLM-Based Multi-Agent Framework

Recently, LLM-based multi-agent frameworks, e.g., Generative Agents [41], CAMEL [42], Multi-Agent Debate [43], have attracted tremendous attention for achieving human-level intelligence in diverse

domains. MetaGPT [23] is a representative work for automatic software development, incorporating LLMs into a multi-agent conversation system. Each agent, like GPT, is assigned a specific task in streamlined workflows and collaborates to accomplish the final goal. In addition to practical applications, LLM-based autonomous agents can also promote social sciences by leveraging their impressive abilities [44]. For example, LLM-based agents can be utilized for providing mental health support in psychology [45], for predicting voting patterns in economy [46], and for simulating the decision-making processes of multiple judges in jurisprudence [47]. In our work, we utilize a multi-agent system to analyze sociological data in a novel way. Given that large language models operate as black-box models, utilizing a single large language model for data analysis lacks interpretability in the conclusions. To address this, we employed several task-specific agents that iteratively collaborate to obtain persuasive conclusions with enhanced interpretability.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) [48] is identified as a pivotal phase within any knowledge discovery process. This critical stage involves the interactive implementation of a myriad analytical techniques, including data filtering, aggregation, and sophisticated visualization strategies, all aimed at extracting valuable insights from raw datasets. Traditionally recognized as a labor-intensive and time-consuming process, EDA demands deep analytical skills, comprehensive experience, and robust domain-specific knowledge from researchers and data practitioners. Notably, the landscape of EDA is undergoing a transformative shift, thanks to significant advancements in computer technology. These developments have not only streamlined the facilitation of EDA but have also opened the door to the potential of fully automating the entire process. This paradigm shift offers a promising horizon for enhancing efficiency and accuracy in data analysis, thereby revolutionizing the way insights are extracted and applied across various domains.

In recent years, scientists have made advanced progress in automating EDA processes, whether partially or fully. There are three main approaches to EDA automation. One of them is a guided/interactive data exploration system, which assists the user by recommending relevant data components based on the user's engagement with the system. AIDE [49] is an automated interactive data exploration framework that combines classification algorithms and data management optimization techniques to efficiently identify novel data patterns, learning from user relevance feedback on selected samples. Studies have been conducted to develop practical EDA recommendation systems. Such systems typically reside between the dataset and the EDA interface, and are used to suggest high-quality EDA operations without relying on any user input or specifications [50]. Previous work mostly focuses on one specific aspect of the process, such as Snipsuggest [51] for data retrieval, Seedb [52] for data visualization, and Dbexplorer [53] for data summaries. To solve the above problem, REACT [54] provides a holistic method that exploits diverse user experience to generate personalized and context sensitive recommendations. End-End automated EDA systems are also a common application in the domain of exploratory data analysis, which produces whole EDA sessions utilizing techniques like Reinforcement Learning based on prior EDA sessions and dataset characteristics [55, 56, 57, 58, 59, 60]. Considering that the knowledge conveyed by EDA is usually

disjointed or fragmented, MetaInsight [61] proposes a new formulation of basic data patterns to gain structured knowledge for users from raw datasets.

3. METHOD

In this paper, we propose a tool-augmented LLM-based multi-agent framework to tackle the data analysis task, as shown in Figure 1. The framework, when provided with specific databases or data sources as inputs, leverages the intellectual capabilities of LLMs and cross-validates their outputs with external knowledge bases in a cohesive manner. This process ensures the responses generated are not only reliable but also encompassing a broad spectrum of information.

3.1 Data Analysis Multi-Agent Framework

Solving complex tasks or problems often needs the cooperation of agents possessing diverse skills and expertise, each contributing specialized outputs to specific aspects of the problem. We have modularized the problem by proposing three distinct modules:

1. **Reasoner**, represented by M_r, which generates data patterns and key terms;
2. **Verifier**, denoted as M_v, which retrieves keywords generated by M_r and integrates information obtained from the Internet;
3. **Evaluator**, symbolized as M_e, is responsible for scoring and assessing the evidence produced by M_v.

Subsequently, we provide a detailed exposition of the responsibilities undertaken by each agent.

3.1.1 Reasoner

The Reasoner is built upon a large language model, specifically prompted to generate insights into trends and particular temporal nodes within the data. It excels at extracting keywords from the provided data or information. These keywords, produced by the model M_r, act as a bridge to M_v (the next phase in our process), laying the groundwork for subsequent retrieval results of high relevance and precision. To enhance the Reasoner's capacity for detailed analysis, we have identified the availability of related information as a crucial factor influencing the output effectiveness of LLMs. Following the initial analysis, the Reasoner's capability is further augmented through the integration of a specially designed database. This database is a dynamic tool that enriches the model with a wealth of factual evidence. Introduced after the first iteration, its role is to deepen the model's understanding, allowing for a deeper exploration of data patterns. By providing this enriched context, the database significantly improves the model's ability to make informed and accurate judgments.

3.1.2 Verifier

The Verifier, functioning as a vital component of our framework, searches for professional knowledge and factual evidence related to specific topics, in order to provide clear and accurate explanations for

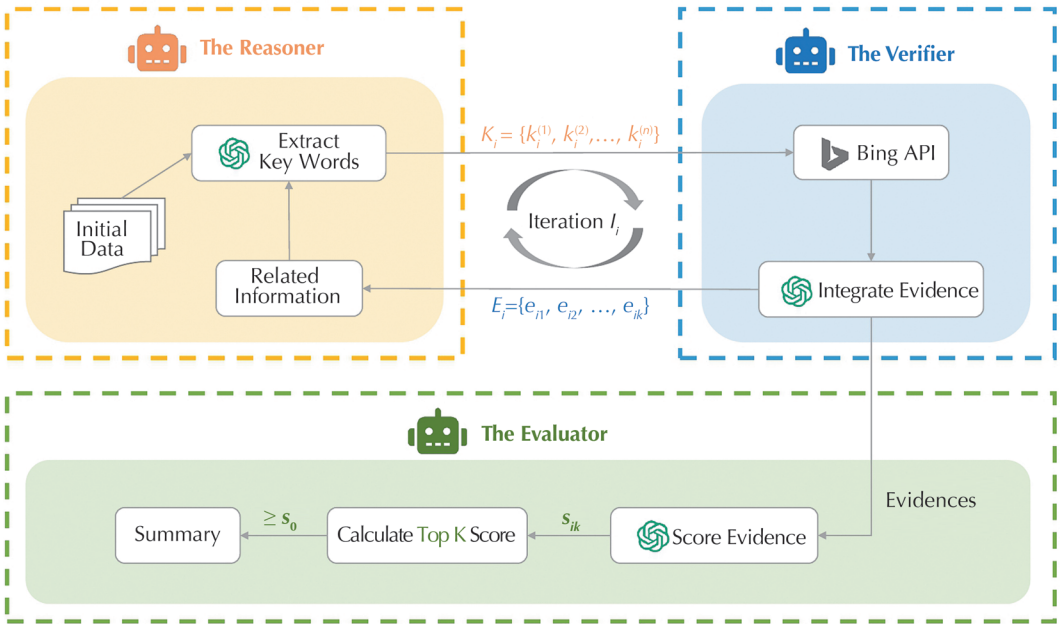


Figure 1. Architecture of our proposed tool-augmented multi-agent data analysis framework.

patterns observed in the data. Utilizing keywords generated by M_r as inputs, it invokes search engine interfaces to acquire a broader and more in-depth range of professional information. However, the Verifier's job doesn't stop at data collection. Aware of the abundance of information available, it employs the advanced capabilities of LLMs to sift through the search results. The Verifier leverages LLMs to filter and eliminate irrelevant or low-quality information from the search results. Through this approach, we achieve the integration and processing of information, ensuring the efficiency and standardization of information transmission within the entire framework.

3.1.3 Evaluator

Evaluating the impact of agents on the quality of generated results is crucial. Due to the inherent uncertainty in the strength of the correlation between the knowledge or arguments retrieved and the original data, information with lower relevance may adversely affect the analysis. To navigate this, the Evaluator utilizes the LLM itself as a means of evaluation. This innovative approach involves the LLM assessing a piece of evidence generated during the analysis and assigning it a relevance score. This scoring is on a scale from one to five, with higher scores denoting a stronger connection to the initial dataset. Upon completing the scoring phase, the Evaluator, denoted as M_e, computes the average score for the top-k arguments. Achieving an average score above a certain predefined threshold serves as an indicator that the arguments are coherent and solidly grounded. This method introduces a systematic and quantitative way to gauge the precision and dependability of the analytical results. Through such a

mechanism, we can ensure a more rigorous evaluation of how well the arguments generated align with the original data, bolstering the overall integrity and validity of the analysis conducted.

3.2 Multi-Round Iterative Optimization Algorithm

It is undeniable that LLMs (such as the GPT series) exhibit commendable performance in responding to general queries. However, when confronted with more complex issues, these models often fall short in providing in-depth and comprehensive information in a single response. Notably, for identical questions, increasing the frequency of queries cannot achieve the desired effect, as the models tend to produce highly uniform content. Furthermore, in scenarios requiring extensive knowledge in specific professional fields, the accuracy of their responses can be less than satisfactory. These limitations pose significant challenges in meeting the high standards of completeness and precision demanded in scientific research.

Our work proposes an innovative multi-round iteration optimization algorithm that amalgamates the collaborative efforts of agents with the advanced capabilities of LLMs. Specifically, the algorithm encompasses two specialized agents. The first agent named the Reasoner utilizes the model's internal knowledge base for initial analysis and reasoning. Importantly, it also integrates relevant information gathered with each subsequent iteration, significantly enhancing the depth and quality of analysis. The second agent, known as the Verifier, is tasked with retrieving external knowledge pertinent to the problem, optimizing information retrieval based on new keywords suggested by the Reasoner in each iteration. This dynamic process benefits immensely from the LLM's ability to adapt contextually and utilize its short-term memory effectively. As the algorithm progresses through iterations, it systematically incorporates and builds upon external knowledge, allowing the model to tackle increasingly complex problems with greater efficiency and accuracy. Through this innovative approach, our work offers a more nuanced and effective method for problem-solving in complex scenarios.

3.3 Overall Process

We present a tool-augmented multi-agent framework for complicated data analysis tasks, which is formalized as an iterative optimization process. Our framework initiates with the Reasoner M_r, where the input is an initial dataset q, and a related information database R designed to store relevant evidence. It is noteworthy that R is empty at the beginning. In the first trial, M_r produces a series of trends and keywords $K_1 = \{k_1^{(1)}, k_1^{(2)}, \cdots, k_1^{(n)}\}$ by analyzing q, leveraging the internal knowledge of LLMs. For each keyword $k_1^{(i)}$, the Verifier M_v then retrieves web pages and selects the relevant evidence snippets $\{e_{i1}, e_{i2}, \ldots\}$, which will be stored in the corpus R after undergoing data cleansing. After the first attempt, to enhance the depth and breadth of the analysis, M_r utilizes associated information from R and the original data q for a thorough reexploration. Upon completing each iteration I_j, the Evaluator M_e also assigns a score s_{ik} to each piece of evidence e_{ik}, and ranks the evidence based on the correlation between the evidence e_{ik} and the trends present in q. Then M_e selects the top K evidence and computes an average score S_j. When the score S_j reaches the threshold S_0 we have set, we can consider the information mining process as sufficiently comprehensive.

The core of our framework lies the strategic integration of advanced LLMs, which leverages their deep understanding and analytical capabilities to sift through and interpret vast datasets. By doing so, it extracts nuanced insights that might otherwise remain obscured. Concurrently, our framework implements a thorough validation process, wherein the insights generated by the LLMs are continuously integrated with external knowledge. This dual-pronged approach facilitates a robust verification process, ensuring the accuracy and trustworthiness of the information being provided.

3.4 Parallel Optimization

To address the inefficiency caused by the sequential operation of the Reasoner and Verifier, where each is in a waiting state while the other works, we have introduced a parallel optimization strategy. This new approach allows both the Reasoner and Verifier to operate concurrently, significantly reducing the overall processing time. In this parallelized system, the Reasoner can continue analyzing new segments of data or refining its previous analysis based on feedback, while the Verifier simultaneously searches for and validates external information. This parallel optimization process is shown in Figure 2.

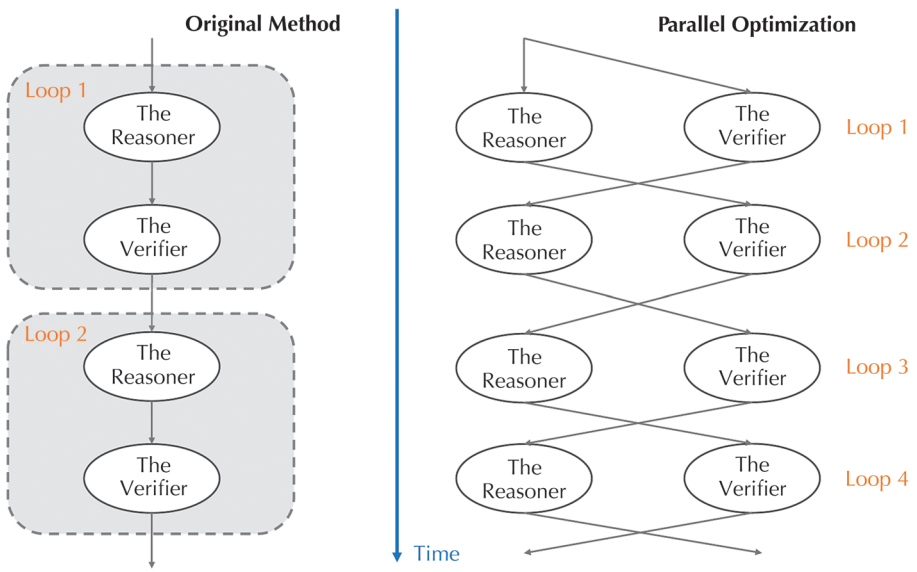


Figure 2. Our proposed parallel optimization mechanism.

We optimize the original model by using misaligned output. The main idea is to utilize the original waiting time of both the Reasoner and the Verifier to process the output from the previous iteration. In the first round, in the lack of prefix information, the verifier will directly analyze the information in the original table, such as title and column name.

To demonstrate the efficiency of parallel optimization, we denote the processing times for the Reasoner and the Verifier to complete their work as T_r and T_v, respectively. The time required for information exchange is negligible compared to the processing time. Therefore, in the original case, the total time $T_{original}$ of each iteration is $T_r + T_v$. However, in the optimization case, this time $T_{parallel}$ becomes $max(T_r, T_v)$.

The misalignment of outputs theoretically does not affect the final output effect. This is because our output is not continuous, and the LLM process itself incorporates an element of randomness. For instance, if you were to swap the outputs of the related information from the second and third rounds in the original model, the model would still achieve a statistically similar result.

In the optimization process, the work of the Evaluator remains unchanged. It continues to use the Verifier's output as a marker to identify the completion of an iteration and to score it.

4. EXPERIMENTAL SETTING

4.1 Dataset

In our experiment, we employ multiple authentic statistical datasets for problem analysis, all sourced from the website of the National Bureau of Statistics of China. We conduct the pruning and cleansing procedures on the datasets, restricting the temporal scope to the most recent two decades. The utilized datasets encompass: "Beijing Higher Education Institutions and Student Annual Data", "Annual Data on the Number of Students per 100,000 People in Beijing", "Annual Data on Higher Education Science and Technology Activities in China", "Annual Data on China's Gross Domestic Product (GDP)", "Annual Data on the Consumer Spending Levels of Chinese Residents" and "Annual Data on the Contribution Rates of the Three Industries in China". Due to missing data in both earlier and later years within some datasets, we truncated the analysis to the years spanning from 2005 to 2021 to ensure data alignment.

We illustrate the detailed content of each dataset used in our experiments:

1. The "Beijing Higher Education Institutions and Student Data" provides a comprehensive overview of pertinent statistics pertaining to higher education in Beijing over the specified two-decade period.
2. The "Annual Data on the Number of Students per 100,000 People in Beijing" dataset reflects the enrollment figures from kindergarten through primary, junior high, and high school, to higher education.
3. The "Annual Data on Higher Education Science and Technology Activities in China" dataset encompasses data on research funding, research personnel, and research outcomes across major educational and research institutions.
4. The "Annual Data on China's Gross Domestic Product (GDP)" dataset portrays the development of the primary, secondary, and tertiary industries over the past 20 years, along with the growth of the national GDP.
5. The "Annual Data on the Consumer Spending Levels of Chinese Residents" dataset documents changes in consumer spending levels in both urban and rural areas.

6. Lastly, the "Annual Data on the Contribution Rates of the Three Industries in China" dataset records the contribution rates of the primary, secondary, and tertiary industries to the overall GDP.

To validate the performance of our framework, we apply it to three real-world problems in the Research Development of Higher Education in China scenario:

- **Question 1**: Changes in the number of higher education institutions (Data source: the "Beijing Higher Education Institutions and Student Data").
- **Question 2**: Changes in scientific research data (Data source: "Annual Data on Higher Education Science and Technology Activities in China", "Annual data on the basic situation of scientific research and development institutions", "Annual data on basic scientific and technological activities").
- **Question 3**: Changes in China's industrial structure data (Data source: "Annual data of the composition of the three industries", "Annual data of the contribution rate of the three industries", "Annual data of the GDP").

4.2 Baseline

In this paper, we employ the ChatGPT4 as the baseline to compare with our proposed method. We prompt the ChatGPT4 using the following prompt to conduct the data analysis task:

You're a senior data analyst.Now we will provide you with some real statistical data. Please analyze the changing trend and the special points of these data.Then, provide us the relevant evidence and analysis, according to the website of the National Government of China (www.gov.cn). You should give both the source and the analyze.

4.3 Evaluation Metrics

To quantitatively evaluate the performance of our proposed model, we use the correlation between the output and the original data as the evaluation metric. In our model, all evaluations are conducted by the evaluator, which is also based on the LLM. To determine whether our model aligns with human understanding, we invited four data science participants to score the generated analysis as a reference.

Specifically, the LLM-based evaluator and participants are provided with three datasets and asked to score results based on the raw data. We apply the same scoring standards for both experiment participants and the LLM-based evaluator. To be specific, a score of 1 indicates no correlation, that is, the rater believes that the generated result has no causal relationship with the source data trend. A score of 5 signifies a strong correlation, suggesting the generated results have a strong causal link to the original data. A score of 3 means moderate relevance, implying there is some likelihood that the generated results could cause the trend of original data, or just have a moderate impact on the original data. Scores of 2 and 4 represent intermediate levels of correlation. Finally, we average the scores to represent the final score for a single outcome.

To address the integrity of our model's output, we introduce another evaluation metric–keyword coverage. For each question, we invite data analysis experts to formulate the "correct answer". Keyword coverage measures the extent to which the model's output covers the correct answer, also known as semantic repetition.

5. EXPERIMENTAL RESULTS

5.1 Overall Performance

The experiment collected a total of 272 (68*4) manually assessed scores. Concurrently, we developed a standard answer for each question based on the consensus of three data analyses. We process both the generated results and standard answers using an LLM to extract their keywords. Then we determine their similarity using synonyms, setting the threshold for "covered" as a similarity greater than 0.5. This process allows us to calculate the keyword coverage of each model against the settled standard answer. Results are reported in Table 1, Table 2, and Table 3.

Table 1. Evaluation analysis results of using Top-4 retrieval results.

	Question 1	Question 2	Question 3
ChatGPT4	4.68	4.56	4.58
Our Model	4.96	4.95	4.89

Table 2. Evaluation analysis results of using Top-8 retrieval results, where ChatGPT4 failed to get enough retrieval results in Question 2.

	Question 1	Question 2	Question 3
ChatGPT4	4.36	–	4.01
Our Model	4.92	4.93	4.79

Table 3. Keyword coverage score on the results based on top-8 retrieval results.

	Question 1	Question 2	Question 3
ChatGPT4	0.16	0.36	0.31
Our Model (iteration number = 3)	0.42	0.68	0.56
Our Model(iteration number = 5)	0.8	0.91	0.88

Table 1 and 2 show the manual evaluation scores for the results generated by each model, with our model utilizing four iterations to mine the data. Table 1 presents the results of the top-4 scores, while Table 2 focuses on the results of the top-8 scores. This approach evaluates whether the model can identify the most important reasons for the given data trend, and it also assesses the depth and breadth of each model's data exploration.

Table 3 shows the keyword coverage rate, which ranges from 0 to 1. Obviously, rates closer to 1 indicate that the results are more in line with human understanding.

As shown in Tables 1-3, our model consistently outperforms others across all three questions. Specifically, our model has significantly improvement over the results produced by ChatGPT4. Our model is more adept at identifying high-scoring results, making it easier to pinpoint the most relevant findings. An interesting observation from the table is that, while ChatGPT4 achieves an average level of scoring, it struggles to accurately cover the "correct answer". We believe this is because ChatGPT4 tends to identify direct connections and is less adept at recognizing indirect connections. However, our model, after several iterations of the iteration, showcases strong mining capabilities as anticipated. It excels at uncovering indirect connections. These results show that after the number of iterations reaches 5, the coverage rate exceeds 80% on all three real data problems, illustrating our model's superior ability to delve into and elucidate the underlying complexities of the data.

5.2 Ablation Study

In our paper, we perform ablation experiments to evaluate the effect of certain components. In our model, we update the information we get through each iteration. Within this process, both the Reasoner agent and the Verifier agent are pivotal; they supply each other with new keywords and new related information, respectively. Consequently, to verify the efficiency of each agent, we remove the output of related information by the Verifier, and the output of keywords by the Reasoner, to form two ablation models. We perform an experiment on question 2, with iteration number 5. In this experiment, we utilized two variations of the original model to assess the impact of specific components on its performance. The model *w/o Verifier* involved using a modified version of the model without providing the output of related information to the Reasoner, effectively testing the importance of this feedback iteration. The model *w/o Reasoner*, on the other hand, involved using a version of the model without supplying the output of keywords to the Verifier, thereby evaluating the significance of keyword guidance in the model's ability to identify and verify relevant information. Results are shown in Table 4.

From Table 4, it's evident that the model's performance declines when specific functionalities of components are removed. This decline in performance can be attributed to the lack of association, which hampers the model's ability to uncover deeper layers of information. Notably, the performance degradation is more pronounced when the related information feedback to the Reasoner is eliminated. This observation underscores the critical importance of introducing new information into the system, highlighting its pivotal role in enhancing the model's predictive capabilities. Another significant finding

Table 4. Performance of each ablation model.

	w/o Verifier	w/o Reasoner	Original Model
Top-4 score	4.62	4.87	4.95
Top-8 score	4.25	4.56	4.93
Keyword Coverage	0.63	0.78	0.91

is the profound impact of these removed functionalities on the breadth of the results, as evidenced by the keywords coverage and the top-8 average score. This suggests that the ability to generate new keywords and to incorporate related information is integral not only to the depth but also to the breadth of the model's exploration and analysis.

5.3 Analysis of Hyper-parameters

5.3.1 Iterative Number

The core of our model revolves around multi-round iterations, making the selection of iteration number a critical issue. Multiple iterations can effectively explore indirect connections within the data, but they also significantly increase computational costs. To achieve a balance between effectiveness and computing resource utilization, we record the scores obtained by the LLM-based evaluator for each iteration. We aim to identify patterns within these scores to optimize the number of iterations needed without compromising the quality of the analysis. The results are presented in Figure 3.

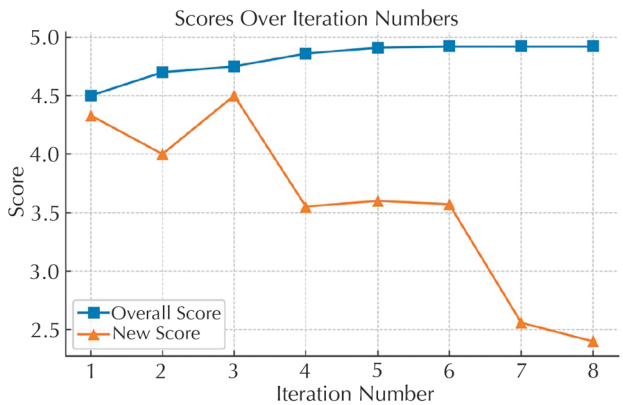


Figure 3. Performance of the multi-agent system using different iterative numbers.

Figure 3 illustrates the variation in the total score after each iteration of the iteration, as well as the change in the average score of the newly obtained results. In this context, the total score refers to the top-8

scores. From Figure 3, we can find that the overall score is constantly improving. This is intuitive since we only use the high score results from each new iteration. For this experimental problem, we can obviously see that the total score growth slows down significantly after the number of iterations reaches 5. And then, we look at the line of newly got scores in each iteration. It can be found that there is a short rise at first, but the overall trend shows a fluctuation decline. The potential reason for this phenomenon lies in direct or short-range indirect correlations that are easy to find. The fluctuation shows that our model successfully finds deep information, which means long-range indirect correlations.

5.3.2 Analysis of Top-K

The choice of the K value has a certain influence on the judgment of iteration end condition. Generally speaking, the larger K is, the higher the number of accurate results we will require. But as the same time, also cause an increase in computing power. Here we calculate the change in top-k rating during an execution in Figure 4.

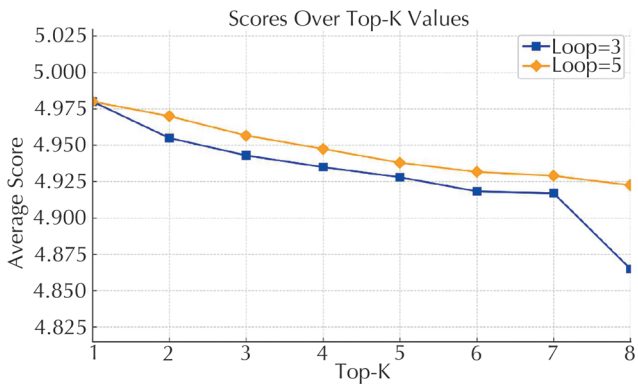


Figure 4. The average score for using different top-k results.

5.3.3 Analysis of S_0

In § 3.3 we mentioned that when the score S_j reaches the threshold S_0 we have set, we can consider the information mining process as sufficiently comprehensive. The choice of S_0 is crucial. A low S_0 will result in inaccurate final results. But an increase in S_0 can increase the number of iteration exponentially and even make it difficult to end. Although we set an upper limit on the number of loops, too many loops are obviously meaningless. So we test the number of iterations with settled S_0

From Table 5, we find that when the default value of s0 is higher than a certain threshold, the iteration number increases significantly. In order to keep both iteration time and result accuracy within a reasonable limit, we believe that the optimal S_0 in top-8 may be around 4.80. An increase in K of top-K should be accompanied by a decrease in S_0. Since a higher K makes it harder for the average score S_j to meet S_0.

Table 5. The number of iteration with settled S_0 with K=8.

S_0	4.50	4.60	4.70	4.80	4.90	4.925	4.95
Iteration time	1	2	3	5	6	9	11

5.4 Token Consumption Analysis

Multi-agent frameworks tend to have high token consumption, mainly reflected in token consumption. Here, we counted the token consumption in all three problems on our framework and baseline. Results are shown in Table 6.

Table 6. Average token consumption.

	ChatGPT 4	The Reasoner	The Verifier	The Evaluator
token (input)	74	161	197	829
token (output)	2908	2326	2522	2451
total	2982		8486	

Due to the addition of related information, the input and output token consumption of our model have increased. But the manual evaluation scores show the effectiveness of our model. In particular, there has been an improvement in keywords coverage. So, we think that only a triple increase in token consumption is worth it.

5.5 Performance of Different Backbone LLM

A multi-agent structure is used in our framework as mentioned in § 3.1. The LLM is the core of every agent. In our framework, users are allowed to choose different LLMS according to their needs. This choice may have a significant impact on overall performance. We use Problem 2 with iteration number 4 as an example to investigate the effects of different LLM. The results are as follows in Table 7. Here, in addition to the original gpt-4-turbo, we replace the backbone of the framework with Claude-3-Sonnet, Tongyi (developed by Alibaba Cloud) and Yiyan (developed by Baidu), respectively.

Table 7. Performance of using different LLM backbone.

	gpt-4	Claude	Tongyi	Yiyan
top-4 score	4.95	4.93	4.91	4.92
top-8 score	4.93	4.78	4.89	4.91
Keyword Coverage	0.91	0.94	0.84	0.81

We are quite surprised to find that, with the use of our framework, almost all LLMs can explore deeper results after sufficient iterations, since the top-4 score difference is very small. The difference in keyword coverage is not so much as well. But some LLMS stand out–gpt-4 and Claude. They were able to mine a larger breadth of results in a shorter number of iterations, possibly due to the amount of pre-training data and parameters.

5.6 Case Study

In this section, we demonstrate one of our model's outputs of question 2 (Research data change in China) in Table 8. The user wants to know the reasons behind changes in Chinese research data. The model finds the trend of the data and speculates that the reasons might be related to policies. Consequently, the Reasoner provides "Keywords" such as "government policy" and "funding", which help the Verifier in locating relevant events in the database. These events are marked as "Data source" and used as evidence. Then, the reasoner agent summarizes the information from the data source, which is shown in "Evidence". Finally, the results are evaluated and scored by the LLM-based evaluator, which is shown in "Machine Score".

Table 8. The first four rows show an example of the evidence with the title of the knowledge data source used by our framework. The evidence is extracted by the Verifier agent from the knowledge data source. This evidence is scored 5 by the LLM-based evaluator. The last row shows an example of the generated final analysis by our model.

Keywords	Government Policy; Funding
Evidence	China has been increasingly focusing on education, science, and technology as key drivers of its economic growth and global influence. Government policies and funding have been instrumental in shaping these trends.
Data Source	Notice of The State Council on Measures to Optimize Scientific Research Management and Improve Scientific Research Performance.Science.Gov.cn (www.gov.cn)
Machine Score	5
Analysis	the expansion of higher education institutions, driven by government policies, economic development, and the societal value placed on education. Additionally, there is a focus on the growth of research and development (R&D) institutions and personnel, reflecting a national emphasis on innovation and technological advancement. This trend is tied to strategic goals for enhancing technological competitiveness and addressing complex societal challenges. The balanced growth in both basic and applied research, underlying the importance of scientific progress and practical solutions to real-world problems. Variations in the number of R&D institutions are attributed to funding changes, policy shifts, and the evolving scientific landscape, while the steady increase in R&D personnel underscores the importance of investing in human capital for research and innovation. These trends underscore the pivotal role of education, research, and innovation in driving economic growth, technological leadership, and the ability to tackle global challenges.

Here, we set S_0 as 4.80 and K as 5, which means the process ends when the average score of the top-5 results exceeds 4.80. Based on the top-5 results, the Evaluator then provided us with a summary, which is displayed in "Analysis".

6. CONCLUSION

In this paper, we propose a multi-agent data analysis framework based on LLMs. The framework decomposes the data analysis task into subtasks and empowers agents with the ability to search for relevant knowledge. Additionally, a multi-stage iterative optimization method is introduced to improve the quality of the analysis results. We evaluate our framework on three real-world problems and demonstrate its effectiveness in outperforming baseline methods. Our framework achieves more accurate and insightful analysis results by leveraging the strengths of LLMs and knowledge search. Future research directions include exploring other data analysis tasks, improving the efficiency of the framework, and investigating the use of other types of LLMs.

AUTHOR CONTRIBUTIONS

- Xilin Zhang: methodology, writing review and editing;
- Ziwen Chen: methodology, writing review and editing;
- Zhixin Mao: methodology, writing review and editing;
- Shen Gao: methodology.

REFERENCES

[1] Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2024)
[2] Jiang, A. Q., Sablayrolles, A., Mensch, A., et al.: arXiv preprint arXiv:2310.06825 (2023)
[3] Du, Z., Qian, Y., Liu, X., et al.: Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360 (2022)
[4] McCann, B., Bradbury, J., Xiong, C., et al.: Learned in translation: Contextualized word vectors. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pp. 6297–6308 (2017)
[5] Zheng, K., Zheng, Y., Yuan, N. J., et al.: On discovery of gathering patterns from trajectories. In: The 29th International Conference on Data Engineering (ICDE). pp. 242–253. IEEE (2013)
[6] Zheng, K., Shang, S., Yuan, N. J., et al.: Towards efficient search for activity trajectories. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE). pp. 230–241. IEEE (2013)
[7] Shang, S., Ding, R., Yuan, B., et al.: User oriented trajectory search for trip recommendation. In: Proceedings of the 15th International Conference on Extending Database Technology. pp. 156–167 (2012)
[8] Khan, N. S., Abid, A., and Abid, K.: A novel natural language processing (NLP)–based machine translation model for English to Pakistan sign language translation. Cognitive Computation 12, 748–765 (2020)

[9] Silva, G., Ferreira, R., Lins, R. D., et al.: Automatic text document summarization based on machine learning. In: Proceedings of the 2015 ACM Symposium on Document Engineering. pp. 191–194 (2015)
[10] Zheng, K., Zheng, Y., Yuan, N. J., et al.: Online discovery of gathering patterns over trajectories. IEEE Transactions on Knowledge and Data Engineering 26(8), 1974–1988 (2013)
[11] Shang, S., Ding, R., Zheng, K., et al.: Personalized trajectory matching in spatial networks. The VLDB Journal 23, 449–468 (2014)
[12] Shang, S., Chen, L., Wei, Z., et al.: Trajectory similarity join in spatial networks. In: Proceedings of the VLDB Endowment. 10(11), 1178–1189 (2017)
[13] Han, P., Yang, P., Zhao, P., et al.: GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 705–713 (2019)
[14] Sarda, A. and Kulkarni, A.: Text summarization using neural networks and rhetorical structure theory. International Journal of Advanced Research in Computer and Communication Engineering 4(6), 49–52 (2015)
[15] Kann, K., Ebrahimi, A., Koh, J., et al.: Open-domain dialogue generation: What we can do, cannot do, and should do next. In: Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 148–165 (2022)
[16] Yang, B., Guo, C., Jensen, C. S., et al.: Stochastic skyline route planning under time-varying uncertainty. In: 2014 IEEE 30th International Conference on Data Engineering. pp. 136–147 (2014)
[17] Shang, S., Chen, L., Wei, Z., et al.: Parallel trajectory similarity joins in spatial networks. The VLDB Journal 27(3), 395–420 (2018)
[18] Shang, S., Chen, L., Jensen, C. S., et al.: Searching trajectories by regions of interest. IEEE Transactions on Knowledge and Data Engineering 29(7), 1549–1562 (2017)
[19] Shang, S., Chen, L., Wei, Z., et al.: Collective travel planning in spatial networks. IEEE Transactions on Knowledge and Data Engineering 28(5), 1132–1146 (2015)
[20] Zheng, K., Su, H., Zheng, B., et al.: Interactive top-k spatial keyword queries. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 423–434. IEEE (2015)
[21] Gao, J., Galley, M., and Li, L.: Neural approaches to conversational AI. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR'18. (New York, NY, USA). pp. 1371–1374. Association for Computing Machinery (2018)
[22] Yuan, A., Coenen, A., Reif, E., et al.: Wordcraft: story writing with large language models. In: 27th International Conference on Intelligent User Interfaces. pp. 841–852 (2022)
[23] Hong, S., Zheng, X., Chen, J., et al.: Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv:2308.00352 (2023)
[24] Tweedale, J., Ichalkaranje, N., Sioutis, C., et al.: Innovations in multi-agent systems. Journal of Network and Computer Applications 30(3), 1089–1115 (2007)
[25] Shamshirband, S., Anuar, N. B., Kiah, M. L. M., et al.: An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique. Engineering Applications of Artificial Intelligence 26(9), 2105–2127 (2013)
[26] Zidan, A., Khairalla, M., Abdrabou, A. M., et al.: Fault detection, isolation, and service restoration in distribution systems: State-of-the-art and future trends. IEEE Transactions on Smart Grid 8(5), 2170–2185 (2016)
[27] Milo, T. and Somech, A.: Automating exploratory data analysis via machine learning: An overview. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. SIGMOD'20. (New York, NY, USA). pp. 2617–2622. Association for Computing Machinery (2020)

[28] Han, P., Li, Z., Liu, Y., et al.: Contextualized point-of-interest recommendation. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. Article No.: 344, pp. 2484–2490 (2020)
[29] Shang, S., Zheng, K., Jensen, C. S., et al.: Discovery of path nearby clusters in spatial networks. IEEE Transactions on Knowledge and Data Engineering 27(6), 1505–1518 (2014)
[30] Rao, R., Meier, J., Sercu, T., et al.: Transformer protein language models are unsupervised structure learners. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria (2021)
[31] Wei, J., Wang, X., Schuurmans, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837 (2022)
[32] Kojima, T., Gu, S. S., Reid, M., et al.: Large language models are zero-shot reasoners. Advances in Neural Information Processing System 35, 22199–22213 (2022)
[33] Zhou, D., Schärli, N., Hou, L., et al.: Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625 (2022)
[34] Wang, X., Wei, J., Schuurmans, D., et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
[35] Yoran, O., Wolfson, T., Bogin, B., et al.: Answering questions by meta-reasoning over multiple chains of thought. arXiv preprint arXiv:2304.13007 (2023)
[36] Yao, S., Zhao, J., Yu, D., et al.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
[37] He, H., Zhang, H., and Roth, D.: Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv:2301.00303 (2022)
[38] Trivedi, H., Balasubramanian, N., Khot, T., et al.: Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2212.10509 (2022)
[39] Ram, O., Levine, Y., Dalmedigos, I., et al.: In-context retrieval-augmented language models. arXiv preprint arXiv:2302.00083 (2023)
[40] Gao, L., Dai, Z., Pasupat, P., et al.: Rarr: Researching and revising what language models say, using language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16477–16508 (2023)
[41] Park, J. S., O'Brien, J., Cai, C. J., et al.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–22 (2023)
[42] Li, G., Hammoud, H. A. A. K., Itani, H., et al.: Camel: Communicative agents for "mind" exploration of large scale language model society. arXiv preprint arXiv:2303.17760 (2023)
[43] Liang, T., He, Z., Jiao, W., et al.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118 (2023)
[44] Wang, L., Ma, C., Feng, X., et al.: A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432 (2023)
[45] Ma, Z., Mei, Y., and Su, Z.: Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. arXiv preprint arXiv:2307.15810 (2023)
[46] Argyle, L. P., Busby, E. C., Fulda, N., et al.: Out of one, many: Using language models to simulate human samples. Political Analysis 31(3), 337–351 (2023)
[47] Hamilton, S.: Blind judgement: Agent-based supreme court modelling with gpt. arXiv preprint arXiv:2301.05327 (2023)

[48] Myatt, G. J.. Making sense of data: a practical guide to exploratory data analysis and data mining. John Wiley & Sons (2007)
[49] Dimitriadou, K., Papaemmanouil, O., and Diao, Y.: Aide: an active learning-based approach for interactive data exploration. IEEE Transactions on Knowledge and Data Engineering 28(11), 2842–2856 (2016)
[50] Milo, T. and Somech, A.: Automating exploratory data analysis via machine learning: An overview. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. pp. 2617–2622 (2020)
[51] Khoussainova, N., Kwon, Y., Balazinska, M., et al.: Snipsuggest: Context-aware autocompletion for SQL. Proceedings of the VLDB Endowment 4(1), 22–33 (2010)
[52] Vartak, M., Rahman, S., Madden, S., et al.: Seedb: Efficient data-driven visualization recommendations to support visual analytics. In: Proceedings of the VLDB Endowment International Conference on Very Large Data Bases 8(13), 2182–2193 (2015)
[53] Singh, M., Cafarella, M. J., and Jagadish, H.: Dbexplorer: Exploratory search in databases. In: 19th International Conference on Extending Database Technology. pp. 89–100 (2016)
[54] Milo, T. and Somech, A.: React: Context-sensitive recommendations for data analysis. In: Proceedings of the 2016 International Conference on Management of Data. pp. 2137–2140 (2016)
[55] Bar El, O., Milo, T., and Somech, A.: Atena: An autonomous system for data exploration based on deep reinforcement learning. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 2873–2876 (2019)
[56] Liu, A., Wang, W., Shang, S., et al.: Efficient task assignment in spatial crowdsourcing with worker and task privacy protection. GeoInformatica 22, 335–362 (2018)
[57] Chen, L., Shang, S., Yang, C., et al.: Spatial keyword search: a survey. GeoInformatica 24, 85–106 (2020)
[58] Wei, Z., He, X., Xiao, X., et al.: Topppr: top-k personalized pagerank queries with precision guarantees on large graphs. In: Proceedings of the 2018 International Conference on Management of Data. pp. 441–456 (2018)
[59] Liu, J., Zhao, K., Sommer, P., et al.: Bounded quadrant system: Error-bounded trajectory compression on the go. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 987–998 (2015)
[60] Rao, X., Chen, L., Liu, Y., et al.: Graph-flashback network for next location recommendation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1463–1471 (2022)
[61] Ma, P., Ding, R., Han, S., et al.: Metainsight: Automatic discovery of structured knowledge for exploratory data analysis. In: Proceedings of the 2021 International Conference on Management of Data. pp. 1262–1274 (2021)

AUTHOR BIOGRAPHY



Xilin Zhang received her M.Eng in 2008 and Ph.D. degree in 2015, both from University of Electronic Science and Technology of China (Chengdu, Sichuan, China). She is currently an Associate Professor with the School of Automation Engineering, University of Electronic Science and Technology of China. She has been a Senior Member of Chinese Institute of Electronics (CIE) since 2020. Since 2005, she has been working on the research fields of multi-agent system, multi-objective optimization, and nondestructive testing. Her research interests in recent years also include Big Data Management and Analysis, and Foundation Model.
ORCID: 0000-0002-5204-4086



Ziwen Chen is studying at the School of Computer Science and Technology, University of Electronic Science and Technology of China. She graduated from the School of Software, University of Electronic Science and Technology of China. She has published papers in the World Wide Web journal and so on.
ORCID: 0009-0001-8809-1618



Zhixin Mao is studying at the School of Computer Science and Engineering, University of Electronic Science and Technology of China. She gratuated from School of Computer Science and Engineering, University of Electronic Science and Technology of China.
ORCID: 0009-0009-3554-7018



Shen Gao is now a tenure-track Assistant Professor with the School of Computer Science and Engineering (School of Cyber Security) of University of Electronic Science and Technology of China. Before, he was an Assistant Professor at Shandong University. For the past years, Shen Gao has been working on Natural Language Processing (NLP) and other related research fields such as Information Retrieval (IR) and Machine Learning (ML).
ORCID: 0000-0003-1301-3700