XBRL Agent: Leveraging Large Language Models for Financial Report Analysis

Shijie Han
Columbia University
US
sh4460@columbia.edu

Haoqiang Kang*
University of California San Diego
US
haoqik88@gmail.com

Bo Jin
Rensselaer Polytechnic Institute
US
jinb2@rpi.edu

Xiao-Yang Liu†
Rensselaer Polytechnic Institute
US
Columbia University
USA
xl2427@columbia.edu

Steve Y Yang
Stevens Institute of Technology
US
syang14@stevens.edu

Abstract

eXtensible Business Reporting Language (XBRL) has attained the status of the global *de facto* standard for business reporting. However, its complexity poses significant barriers to interpretation and accessibility. In this paper, we present the first evaluation of large language models' (LLMs) performance in analyzing XBRL reports. Our study identifies LLMs' limitations in the comprehension of financial domain knowledge and mathematical calculation in the context of XBRL reports. To address these issues, we propose enhancement methods using external tools under the agent framework, referred to as *XBRL-Agent*, which invokes retrievers and calculators. Extensive experiments on two tasks - the Domain Query Task (which involved testing 500 XBRL term explanations and 50 domain questions) and the Numeric Type Query Task (tested 1,000 financial math tests and 50 numeric queries) - demonstrate substantial performance improvements, with accuracy increasing by up to 17% for the domain task and 42% for the numeric type task. This work not only explores the potential of LLMs for analyzing XBRL reports but also augments the reliability and robustness of such analysis, although there is still much room for improvement in mathematical calculations.

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Large language models (LLM), XBRL reports, Semantic-augmented generation

*Haoqiang Kang worked as a RA at Columbia University during summer 2024.
†Corresponding author. Email: XL2427@columbia.edu

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698614

ACM Reference Format:
Shijie Han, Haoqiang Kang, Bo Jin, Xiao-Yang Liu, and Steve Y Yang. 2024. XBRL Agent: Leveraging Large Language Models for Financial Report Analysis. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698614

1 Introduction

In recent years, finance, business and accounting industries have paid more attention to advanced technologies to tackle market complexities and overcome conventional analytical limitations. The recent breakthroughs in large language models (LLMs) such as ChatGPT [39] and LLaMA [47] have demonstrated strong abilities in comprehending complex text files and generating human-like texts. These LLMs are adept at processing vast amounts of textual data and leveraging domain knowledge to extract critical insights [49]. Their abilities to interpret market trends [41], assess risks [8], and provide strategic guidance [10] underscore their extensive application potential prospects in the finance domain [37].

eXtensible Business Reporting Language (XBRL) [44, 56, 57], as shown in Fig. 1, is an open international standard for digital business reporting by numerous regulators globally [36]. It streamlines the creation and dissemination of financial data, thereby facilitating exchanges among investors, financial regulators, and market participants. Over the past two decades, XBRL has become the *de facto* standard for business reporting worldwide, with its adoption by most global economies for financial information sharing [24, 38]. XBRL leverages XML (eXtensible Markup Language) to tag data, providing a standardized format that links numerical data to its semantic context [15]. This allows for precise identification and contextualization of each data point—such as revenue, expenses, or assets—enabling interoperability, accuracy, and transparency. This tagging protocol significantly increases the explanatory power of financial text, making precise interpretation and comparison of financial information possible, while reducing the need for manual re-entry and enhancing automated analysis [12].

However, XBRL's complexity demands specialized knowledge for accurate understanding and insight generation, presenting a steep learning curve for both businesses and the general public [21]. LLMs hold the potential for transforming XBRL analysis by

Total long-term debt	1,149	1,165	1,170
Less current portion	15	13	15
Total long-term debt, less current portion	$ 1,134	$ 1,152	$ 1,155

Fair Value and Future Maturities

See Note 4, *Fair Value Measurements*, for the fair value of long-term debt. Other than the $500 million of principal amount of notes due October 1, 2028, we do not have any future maturities of long-term debt within the next five fiscal years.

7. Revenue

We generate substantially all of our revenue from contracts with customers from the sale of products and services. Contract balances primarily consist of receivables and liabilities related to unfulfilled membership benefits and services not yet completed, product merchandise not yet delivered to customers, deferred revenue from our private label and co-branded credit card arrangement and unredeemed gift cards. Contract balances were as follows ($ in millions):

	May 4, 2024	February 3, 2024	April 29, 2023
Receivables, net(1)	$ 453	$ 512	$ 523
Short-term contract liabilities included in:			
Unredeemed gift card liabilities	242	253	256
Deferred revenue	923	1,000	1,015
Accrued liabilities	57	53	68
Long-term contract liabilities included in:			
Long-term liabilities	239	245	260

(1) Receivables are recorded net of allowances for expected credit losses of $17 million, $23 million and $18 million as of May 4, 2024, February 3, 2024, and April 29, 2023, respectively.

During the first three months of fiscal 2025 and fiscal 2024, $642 million and $747 million of revenue was recognized, respectively, that was included in the contract liabilities at the beginning of the respective periods.

Figure 1: Example of an XBRL-based report from an SEC-10Q file of the company Best Buy.

automating the extraction and interpretation of financial data, identifying key metrics, and generating comprehensible summaries for various applications such as auditing, valuation analysis, forecasting, investment decisions, and more.

LLMs are increasingly integrated into many financial analyses [29, 37, 52–55], such as sentiment analysis [19, 59, 60], financial reports summarization [29], and financial database querying [61]. While existing research has demonstrated the effectiveness of LLMs in handling various financial tasks [53], their application to XBRL data analysis remains under-explored. Given the critical importance of XBRL reports in the financial sector, leveraging LLMs for their analysis could significantly enrich the efficiency of insight generation and reduce barriers to information access for the general public. Therefore, this study empirically investigates the performance of LLMs in the context of XBRL reports.

In this paper, we introduce the first LLM-based agent specifically designed to analyze XBRL reports, called *XBRL-Agent*. Our research empirically evaluates the performance of LLMs in analyzing XBRL reports and identifies crucial limitations in their current capabilities. Notably, existing LLMs exhibit significant deficiencies in expertise in the financial domain and mathematical capabilities when analyzing XBRL reports. To address these limitations, we propose two enhancement methods: 1) incorporating supplementary financial knowledge using Retrieval-Augmented Generation (RAG) technology [26], and 2) incorporating specialized tools such as a financial calculator. Our findings demonstrate that these enhancements substantially improve the performance of LLMs in XBRL report analysis.

Our contributions are summarized as follows:

- **Identifying LLMs' limitations for analyzing XBRL reports**: We conduct the first assessment of LLMs' capabilities in analyzing XBRL reports. It is found that LLM has limited financial domain knowledge and insufficient mathematical capabilities.
- **Implementation of enhancing methods**: To mitigate the identified shortcomings, we propose and implement specific enhancements. These include integrating RAG technology to bolster LLMs with specialized financial knowledge and incorporating a dedicated financial calculator to improve accuracy in complex numeric computations.
- **Improvement in analysis reliability**: Through extensive experiments, we demonstrate substantial improvements. Our enhancements lead to FactScore increases of up to 17% in the XBRL domain query task and up to 42% in the numeric type query task, respectively, effectively boosting the trustworthiness and accuracy of our XBRL-agent.

However, there is still much room for improvement in mathematical calculations. The associated code repository is https://github.com/KirkHan0920/XBRL-Agent

The structure of the rest of this paper is organized as follows. Section 2 provides a comprehensive review of related works. Section 3 presents our motivating experiments. Section 4 details our proposed enhancements to address these limitations. Section 5 describes our experimental setup and presents the results. Finally, Section 6 concludes the paper, summarizing our key findings and broader implications.

2 Related Works

2.1 Application of LLMs in Finance Domain

Recently, the applications of LLMs have improved the efficiency of financial tasks [37]. Several LLMs have been specifically developed to address financial challenges, such as FinGPT [29, 30, 55, 59], BloombergGPT [52] and FinMA [54]. Furthermore, Meyer et al. [33] discusses the opportunities of general LLMs in investment decision-making. In addition, some LLMs have undergone fine-tuning tailored for the financial tasks. As for sentiment analysis, Zhang et al. [59] propose the Instruct-FinGPT model that can effectively explain numerical values and comprehend financial background, thereby assisting users in gaining a deeper understanding of market trends. In the field of decision-making, Aguda et al. [2] develop an LLM framework for predicting and analyzing financial time-series data. Kang and Liu [22] also demonstrate the deficiency of LLMs in answering financial questions through an empirical examination. Specifically, they highlight the severe issue of hallucinations. Following this work, we conduct an in-depth analysis of the causes of hallucinations in LLMs when handling financial queries and propose mitigation strategies for the analysis of XBRL financial reports.

2.2 Challenges of XBRL Analysis

Prior to LLMs, extracting and analyzing XBRL filings poses significant challenges due to complex technical language and required financial expertise. Debreceny et al. [13] note that excessive XBRL extensions hinder cost reduction and cross-firm comparability. Janvrin and No [21] identify technology complexity and learning costs as major adoption barriers. Loughran and McDonald [32] highlight technical vocabulary, industry jargon, and figurative language as critical factors. Other studies have emphasized how complex language in XBRL filings obscures important financial information [5, 16, 31]. Additional challenges include extensive custom tag use [11], fragmented taxonomy adoption [6], data quality issues [14], and high implementation costs [40]. These factors contribute to a significant learning curve and limited accessibility.

While LLMs offer promising solutions, they also face challenges in XBRL analysis. Li and Zhang [27] point out that LLMs may struggle with the highly structured nature of XBRL data and the need for domain-specific financial knowledge. Chen and Liu [9] highlight potential issues with LLMs' interpretation of numerical data and complex financial calculations in XBRL reports. Moreover, Wang and Johnson [50] raise concerns about the ethical implications and potential biases in LLM-based financial analysis. Acknowledging these challenges, this work proposes an enhanced approach that combines LLMs with additional tools to mitigate these challenges.

2.3 Mathematics and Financial Mathematics

In XBRL report analysis, financial mathematics provides essential tools for calculating critical financial metrics. XBRL data contains financial statements that require accurate computations to derive key insights [20]. Core financial formulas like Net Present Value (NPV), Future Value (FV), and Present Value (PV) play an integral role. NPV evaluates investment profitability by calculating the difference between discounted cash inflows and outflows [18]. FV and PV formulas determine investment values over time and at present, respectively [28]. By integrating these formulas into XBRL-Agents, systems can perform sophisticated financial analysis, providing reliable outputs that assist in making informed decisions based on real-time financial data [4].

3 Motivating Experiment

In this section, we utilize popular open-source LLMs' strengths and weaknesses in processing and interpreting XBRL reports, as illustrated in Fig. 1.

3.1 Experiment Setup

LLMs: We utilize widely adopted open-source LLMs. Specifically, we download the weights of instruction tuned *Llama3-8B* [3], *Qwen2-7B* [1], and *Gemma2-9B* [46] from HuggingFace.

Tasks: We collect and create four datasets for the following tasks, while the sample questions illustrated in Fig. 2:

(1) XBRL Domain Query Task (financial domain knowledge):
- XBRL Term [1]: We have extracted over 6,000 XBRL terminology entries and their explanations from the official website XBRL International [2] and XBRL document sources [3] and randomly selected 500 of them to form the evaluation dataset.
- Domain Query to XBRL Reports: We utilize 50 domain questions extracted from XBRL reports in the FinanceBench [20] for evaluation. The relevant XBRL reports provided by the dataset are selectively provided to LLMs according to experimental needs.

(2) Numeric Type Query Task (financial calculation):
- Financial Math [1]: We use ChatGPT to generate 100 typical financial formulas for XBRL reports. For 50 of these formulas, we create 20 unique questions each, producing a total of 1,000 test cases. Additionally, we utilize ChatGPT to generate Python code which provides precise calculations and establishes a standard for accuracy verification for each question, serving as our ground truth.
- Numeric Query to XBRL Reports: We employ 50 numeric queries derived from the same FinanceBench [20]. These questions focus on mathematical aspects of XBRL reports, such as financial calculations and metric analysis. As with the domain query task, relevant XBRL reports are provided to the LLMs as needed for the experiments.

Table 1: Details of tasks.

Type	Name	Number	Metrics
Domain Query	XBRL Term	500	FActScore
	Domain Query to XBRL Reports	50	FActScore
Numeric Query	Financial Math	1000	Accuracy
	Numeric Query to XBRL Reports	50	FActScore

Evaluation Metrics: To evaluate LLMs' performance in analyzing XBRL reports, we utilize two different metrics:

- FActScore [35]: For generation and analysis tasks (XBRL Term, Domain Query to XBRL Reports, Numeric Type Query to XBRL Reports), we adopt FActScore as a metric. It extracts atomic facts from both correct answers and LLM-generated responses for each test. Then, it calculates the percentage of matching atomic facts and computes the overall average alignment across all tests. It facilitates a detailed comparison of semantic content, allowing for a nuanced assessment of the model's ability to capture and articulate key information from XBRL reports.
- Hybrid Evaluation (Accuracy): For tasks involving computations (Financial Formula), we implement a two-stage evaluation process: a) We utilize ChatGPT to perform an automated evaluation of the LLM's calculations. b) then, the authors conduct a manual review that involves checking ChatGPT's judgments against the correct answers to ensure accuracy. The final accuracy is presented as the percentage of correct responses out of the total number of tests conducted.

The task details are given in Table 1.

[1] We have open-source this dataset in https://huggingface.co/datasets/KirkHan/XBRL_Terminology and https://huggingface.co/datasets/KirkHan/XBRL_Formula_Calculation.
[2] https://www.xbrl.org/guidance/xbrl-glossary/, https://xbrl.us/data-rule/dqc_0015-lepr/
[3] https://www.sec.gov/data-research/osd_xbrlglossary

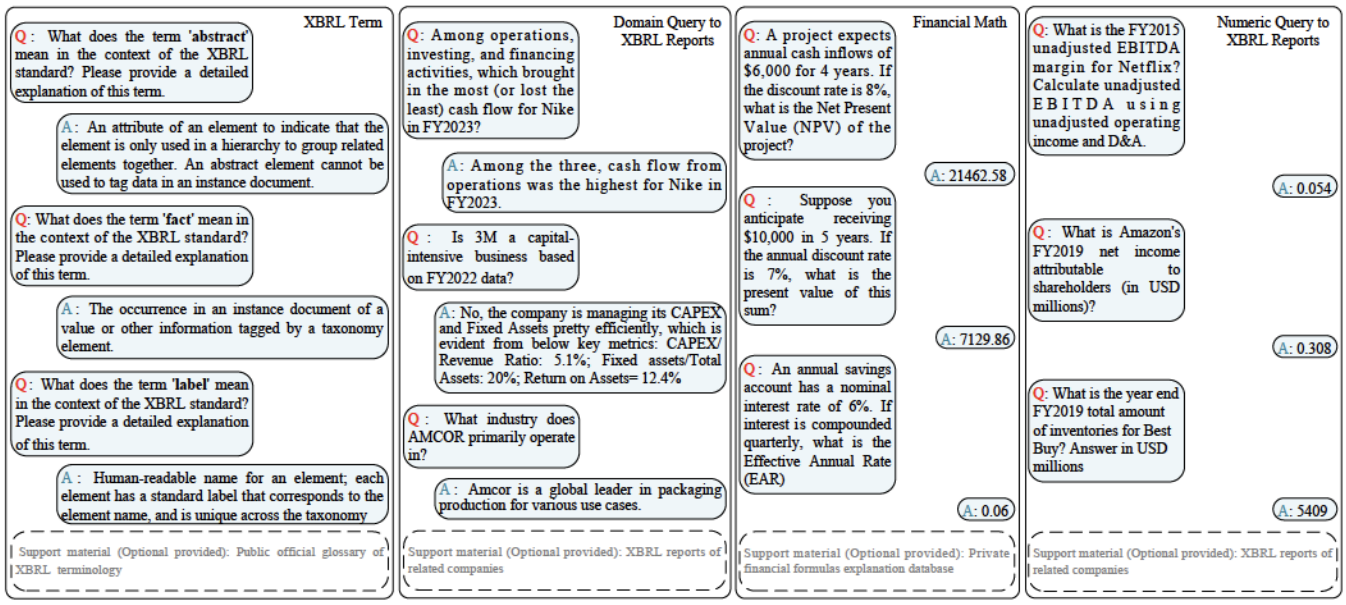

Figure 2: Sample questions and answers for four datasets with optional support materials.

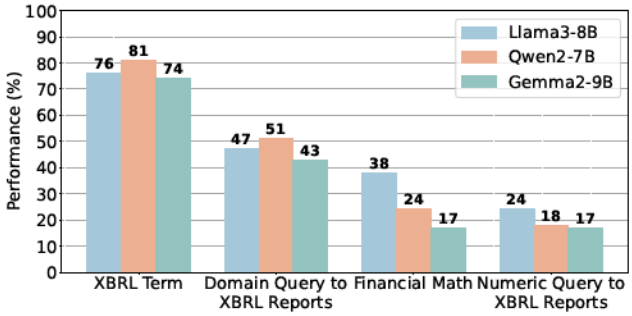

Figure 3: Results of motivating experiment.

3.2 Results

We configure the generation temperature to be 0.5. Generation temperature, ranging from 0 to 1, controls the creativity and diversity of language model outputs. Lower values produce more predictable text, while higher values increase randomness. A temperature of 0.5 is often used as it balances predictability and creativity [17, 42]. We present questions directly to the evaluation models to examine the LLMs' capacity to analyze XBRL reports without any supplementary context. Figure 3 illustrates the performance results of three LLMs in XBRL report analysis.

XBRL Domain Query: LLMs demonstrate moderate proficiency in financial terminology but encounter difficulties with specific XBRL report interpretations. Performance in this category is relatively better, yet the accuracy rates still necessitate improvement. Even the best-performing model, Qwen2-7B, only achieves an 81% score in XBRL Term and a mere 51% in Domain Query to XBRL Reports. This indicates substantial scope for improvement in understanding financial terminology and interpreting domain-specific information. Llama3-8B and Gemma2-9B perform even lower performance, with FActScore declining to 47% and 43% for the domain query task. This suggests a widespread challenge among LLMs in comprehending complex financial concepts and applying them within the context of specific XBRL reports.

Numeric Type Query: LLMs demonstrate significant limitations in handling mathematical data and financial calculations. The performances in this category are particularly concerning. Even the best-performing model, Llama3-8B, only achieves 38% accuracy in Financial Formula Calculation and 24% in Numeric Query to XBRL Reports task. This demonstrates a severe limitation in LLMs' ability to handle mathematical data and perform financial calculations. Qwen2-7B and Gemma2-9B exhibit even poorer performance, with FActScore as low as 18% and 17% in some numeric tasks. This emphasizes a critical weakness in processing and analyzing numerical information in XBRL reports.

3.3 Findings

Overall, the results in Section 3.2 underscore shortcomings in LLMs' capabilities for XBRL report analysis:

Limited financial domain knowledge. The models demonstrate insufficient mastery of specialized financial knowledge and terminology, hindering their ability to provide accurate and granular interpretations of XBRL reports.

Deficient mathematical capabilities The LLMs exhibit a notable weakness in processing and interpreting numeric information, encounter difficulties in performing complex financial calculations and derive meaningful insights from numerical data in XBRL reports.

The above findings indicate a gap between existing LLMs' capabilities and the professional requirements of XBRL report analysis, while they also serve as guidance for further improvements.

4 XBRL Agent

Our analysis reveals two inherent limitations of LLMs for XBRL report analysis, which are difficult to address through internal

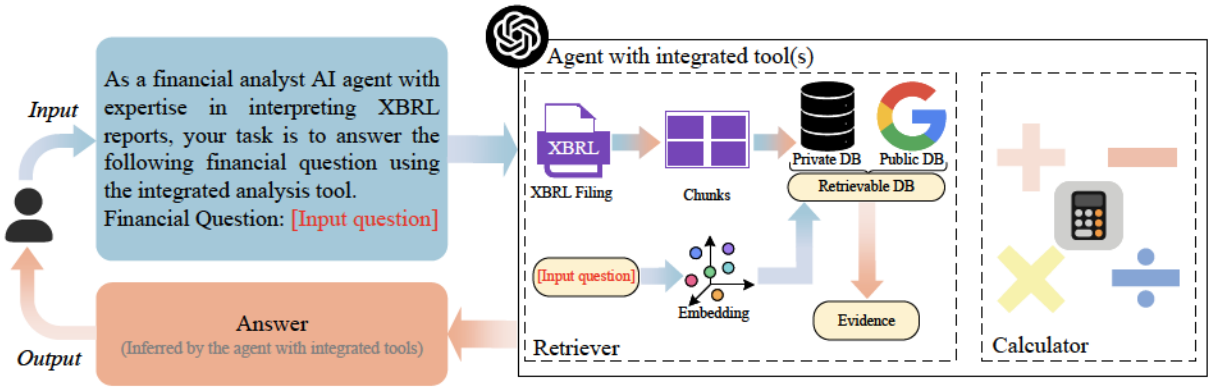

Figure 4: This diagram illustrates an LLM-powered XBRL-agent for financial analysis. The process begins with a financial question input, which is then processed by the XBRL-agent with integrated tools. The retrieval process involves segmenting XBRL filings into chunks, embedding them in a private database, and then combining it with the public database to create an information retrieval database. When a question is posed, the agent retrieves relevant evidence from this database to enhance the response. The calculator is called to perform accurate mathematical calculations when needed.

mechanisms such as prompt engineering alone [7, 51]. The inherent nature of these limitations stems from the general-purpose training of LLMs, which often is insufficient to encompass the depth of specialized financial knowledge required for XBRL analysis [23, 45]. Furthermore, the static nature of LLMs' knowledge emerges as a significant drawback in rapidly evolving financial landscapes, where up-to-date information is crucial [25, 34]. Additionally, LLMs face challenges in representing and manipulating precise numerical data within their neural architectures, a crucial requirement for accurate financial analysis [43, 48].

Inspired by prior LLM agent frameworks, we establish XBRL Agent, a LLM agent integrated with specialized tools for XBRL analysis. As shown in Figure 4, we aim to mitigate the limitations of LLMs with external tools to generate more accurate text. Specifically, we introduce a retriever, implemented through a Retrieval-Augmented Generation (RAG) system, to enhance the LLM's ability to access and utilize up-to-date information from the external database, thereby compensating for the deficiency of specialized financial knowledge. In addition, we incorporate a financial calculator to bolster the LLMs' mathematical calculation capabilities when faced with numerical analyses in XBRL reports. These solutions aim to bridge the gap between current LLMs' performances and the sophisticated requirements of XBRL report analysis.

4.1 Single Tool Use

To address the two specific limitations of LLMs in analyzing XBRL identified in Section 3.3, we propose to implement the following two tools under an agent framework for targeted mitigation, as illustrated in Fig. 4.

Retriever To address the limited financial domain knowledge of LLMs in domain query task, we propose implementing a retriever tool through the RAG process. This tool is designed to enhance the LLMs' capability to handle domain-specific financial tasks.

The motivating experiments have revealed that the current LLMs face challenges in in-depth financial analysis, particularly when dealing with XBRL reports. This limitation severely impacts their ability to provide accurate and insightful analyses of domain-related problems. RAG technology has demonstrated effectiveness in general domains for augmenting LLMs' knowledge bases with specialized information, resulting to more accurate and relevant outputs. Our proposed retriever operates as follows:

(1) Pre-processing: A retrievable database is constructed using professional financial domain knowledge and public databases.
(2) Retrieval: Upon receiving a query, the tool retrieves relevant background information from this knowledge base.
(3) Augmentation: The retrieved evidence is then combined with the original query.
(4) Generation: This augmented input is input into the LLM, enabling it to generate more accurate and detailed responses to complex XBRL-related queries.

By implementing this retriever, we bridge the gap between the LLM's general language understanding capabilities and the specialized knowledge required for financial reporting interpretation. This approach promises to improve the LLM's understanding of financial terminology, concepts, and XBRL-specific information, thereby enhancing its overall performance in analyzing XBRL reports.

Calculator. To mitigate the deficient mathematical capabilities of LLMs in numeric type query, we introduce a calculator tool. This tool is designed to overcome the model's limitations in executing complex mathematical operations in XBRL reports analysis. Implementation of the calculator involves the following steps:

(1) API Integration: We allow an LLM to access a calculator's APIs.
(2) Task Recognition: When encountering mathematical calculations within XBRL reports, the LLM invokes math calculators.
(3) Outsourcing: The LLM outsources these calculations to a calculator by invoking its APIs.
(4) Result Interpretation: After receiving the calculated results, the LLM incorporates them and generates responses.

By utilizing this dedicated tool for numerical operations, we advance the LLM's mathematical capabilities, effectively mitigating

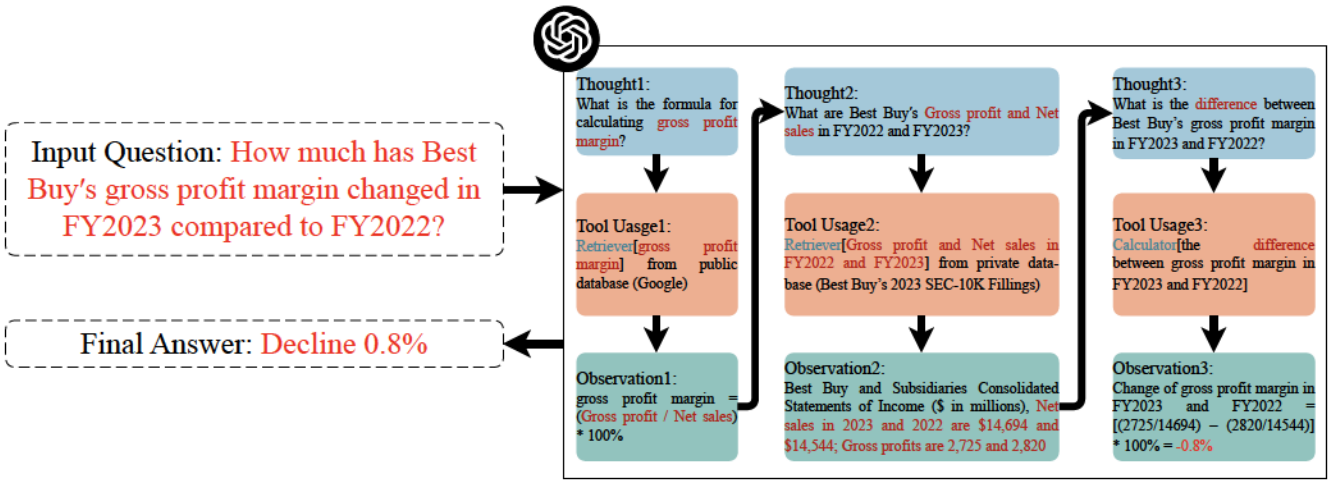

Figure 5: ReAct framework workflow as an example. Following the diagram in Fig. 4, we use the ReAct framework with a retriever and calculator to analyze financial problems until get the final answer.

their existing limitations in the numerical analysis required for XBRL report interpretation.

4.2 Enhancement Using Multiple External Tools

While the retriever and calculator individually mitigate specific limitations in XBRL report analysis, we recognize that numeric type query often demands both enhanced mathematical capabilities and deep financial knowledge. To further elevate the LLM's performance in analyzing numeric type queries in XBRL reports, we suggest integrating multiple external tools within an agent framework.

This multi-tool approach involves an agent that orchestrates the seamless interaction between domain knowledge retrieval and numerical computation. Based on the characteristics of the query, the agent intelligently determines whether to leverage the RAG tool, the Calculator, or a combination thereof.

Integrating multiple tools offers significant advantages in enhancing the LLMs' capabilities. By leveraging both financial domain knowledge and precise calculations, the LLMs are able to deliver accurate answers to complex financial queries. This approach enables the LLMs to tackle multi-faceted problems that require both qualitative understanding and quantitative analysis, a common requirement in XBRL report interpretation.

By implementing these enhancements, we improve LLMs' ability to understand and interpret XBRL reports, elevating their performance towards that of a professional financial analyst.

5 Performance Evaluation

This section presents a comprehensive evaluation of our proposed enhancement methods: the integration of financial knowledge through a retriever and the enhancement of mathematical capabilities through a calculator.

5.1 Experiment Setup

Our experimental setup maintains consistency with the motivating experiment in terms of LLMs, evaluation metrics, and overall structure while introducing targeted enhancement tools to address specific limitations. The use of tools by LLMs is inevitably accompanied by agent orchestration. Currently, the predominant agent framework is the Reasoning and Action (ReAct) agent framework [58]. The ReAct agent uses a standardized prompt template to navigate the LLM through a Chain of Thought (CoT) process. This process involves iterating through three stages: thought, tool utilization, and observation, to generate the correct answer step by step. A detailed example is shown in Figure 5. The ReAct agent enables LLMs to decompose complex tasks into manageable steps, leveraging external tools when required. As a result, the ReAct agent can tackle a wider range of tasks more effectively, producing more accurate and pertinent outputs. We leverage the LangChain[1] library to implement the ReAct agent framework, providing a flexible environment for integrating our enhanced tools. Prompt details are presented in Table 2.

We utilize the same four datasets as in Section 3.1 Motivating Experiment. The XBRL-agent calls external tools:

- For XBRL Domain Query Tasks: We deploy a retriever to mitigate the deficiency in domain-specific expertise.
- For Numeric Type Query Tasks: We initially use an integrated financial calculator to address the deficiency in mathematical abilities. Then, we combine both the retriever and calculator to address potential knowledge gaps alongside computation needs.

5.2 Results

Retriever for Domain Query Task. Implementing a retriever for domain-related queries improves the performance of all three tested LLMs, as shown in the left two columns of Figure 6. For XBRL Term, Qwen2-7B achieves 89% accuracy, followed by Llama3-8B (84%) and Gemma2-9B (83%). These results represent the retriever's effectiveness in enhancing comprehension of XBRL terminology. The more complex Domain Query to XBRL Report task exhibits substantial improvements: Qwen2-7B (65%), Llama3-8B (64%), and

[1]https://www.langchain.com/

Table 2: Prompts used in our experiments.

Experiment	Prompts
Motivating	As a financial analyst with expertise in interpreting XBRL reports, your task is to answer the following financial question: Financial Question: {Input question}
ReAct Agent with Tool(s)	As a financial analyst AI agent with expertise in interpreting XBRL reports, your task is to answer the following financial question using the integrated analysis tool(s): {tool_name(s)}. Use the following format: **Thought:** you should always think about what to do **Tool Usage:** the action to take, should be one of {tool_name(s)} **Observation:** the result of the action ... (this Thought/Tool Usage/Observation can repeat N times until have the final answer) Financial Question: {Input question}

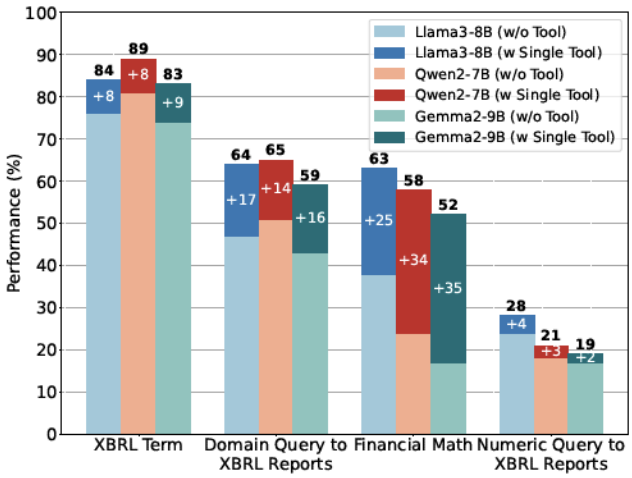

Figure 6: A single tool for different analysis of XBRL reports (in %).

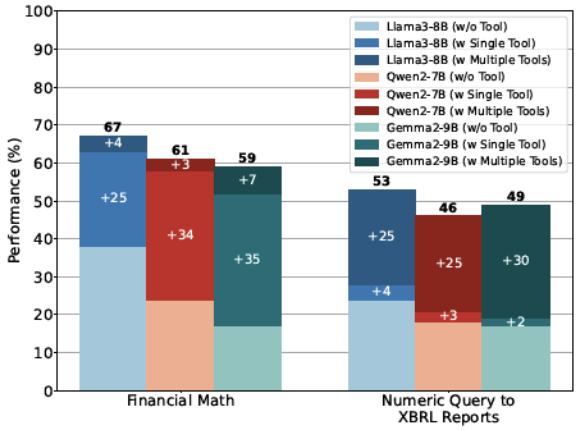

Figure 7: Retriever and calculator for numeric type query (in %).

Gemma2-9B (59%), representing increases of 14 to 17 percentage points. This highlights the retriever's specific efficacy in handling complex XBRL report queries. The consistent improvement across all models and tasks suggests that the retriever effectively enriches LLMs' domain knowledge by providing crucial contextual evidence during queries.

Calculator for Numeric Type Query Task. Integrating a calculator into LLMs improves their performance on numeric type queries (Fig. 6, right columns). For Financial Math, Llama3-8B achieves an accuracy of 63%, followed by Qwen2-7B (58%) and Gemma2-9B (52%), showing 25-35 percentage point improvements. This demonstrates the calculator's effectiveness in enhancing complex financial calculations. The Numeric Query to XBRL Reports task exhibits more modest: Llama3-8B (28%), Qwen2-7B (21%), and Gemma2-9B (19%), representing improvement of 2 to 4 percentage points. The substantial improvements in Financial Math tasks highlight the calculator's efficacy for computational tasks. However, the modest gains in numeric Query to XBRL Report task suggest these complex queries require more than just calculation assistance, likely necessitating both enhanced computational capabilities and domain knowledge.

Retriever & Calculator for Numeric Type Query Task. Combining the retriever and calculator for numeric type query task yields significant improvements (Figure 7). For Financial Math, Llama3-8B led by 67% accuracy, followed by Qwen2-7B (61%) and Gemma2-9B (59%). This incorporating that adding financial knowledge enhances formula application. Numeric Query to XBRL Reports task exhibits profound improvements: Llama3-8B (53%), Gemma2-9B (49%), and Qwen2-7B (46%), representing increases of 25 to 30 percentage points compared to the single tool approach. This combined approach effectively mitigates LLMs' mathematical limitations and domain knowledge gaps in XBRL report analysis, resulting in more comprehensive and accurate numeric data interpretation within financial reports.

5.3 Ablation Study

To further understand the impact of individual tools on numeric type query task, we conducted an ablation experiment using only the retriever. The results shown in Fig. 8 demonstrate notable improvements over the baseline (without tool) but are inferior to the combined retriever-and-calculator approach.

In the Financial Math task, Llama3-8B achieves the highest accuracy at 66%, followed by Qwen2-7B at 58% and Gemma2-9B at 55%. These improvements suggest that domain knowledge provided

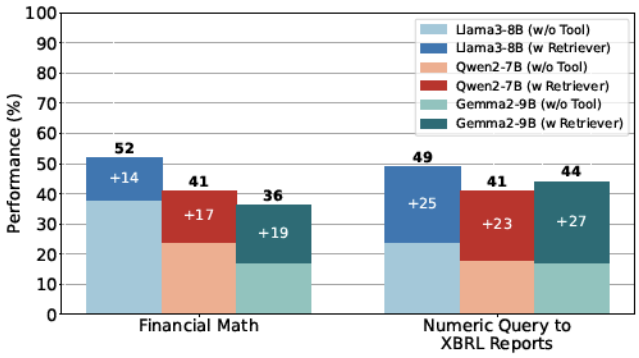

Figure 8: Ablation study of retriever's performance for numeric type query (in %).

by the retriever contributes significantly to the models' ability to understand and apply financial formulas, even without explicit calculation assistance. For the more complex Numeric Query to XBRL Reports task, we have observed similar trends. Llama3-8B achieves an accuracy of 77%, with Qwen2-7B at 64% and Gemma2-9B at 63% (a 27 percentage point increase). While these improvements are significant, they do not match the performance achieved when combining the retriever with the calculator.

5.4 Findings

Our experiments reveal that integrating specialized tools significantly enhances LLM performance in XBRL report analysis. The retriever technology improves domain-related queries, while the financial calculator boosts accuracy in numerical calculations. Notably, combining both tools yields synergistic effects, addressing both the need for domain knowledge and the deficiency in the computational accuracy of financial analysis. The ablation study underscores the importance of domain knowledge, with the retriever-only approach showing significant improvements over the baseline. However, the combined retriever-calculator approach consistently outperforms single-tool implementations across all tasks and models. These findings highlight the potential of tailored tool integration to significantly enhance LLMs' capabilities in specialized domains such as XBRL report analysis.

6 Conclusion and Future Work

In this paper, we conduct motivating experiments to reveal deficiencies in LLM's domain knowledge and mathematical abilities when analyzing XBRL reports. To overcome these challenges, we integrate a retriever to improve domain knowledge retrieval and a financial calculator to bolster numerical processing. Our experimental results demonstrate substantial improvements across various XBRL analysis tasks. We have shown that the RAG technology significantly boosts performance in domain query tasks, while the calculator markedly enhances accuracy in financial calculations. Notably, the combination of both enhancements yielded the most comprehensive improvements, particularly in the complex Numeric Query task. This research not only advances the application of LLMs in financial analysis but also paves the way for more efficient and accessible XBRL report interpretation, potentially transforming how financial data is processed and understood in the industry.

However, we also recognize the intricacies of financial accounting rules across different jurisdictions. The extensive knowledge and subtle nuances embedded in XBRL reports often require a more sophisticated approach. While our enhanced method incorporating additional tools has shown promise, mathematical analysis remains a significant challenge for LLMs. Future research needs to focus on further enhancing LLMs' mathematical capabilities, potentially through the development of more advanced numerical reasoning modules or the integration of specialized financial calculation engines. Additionally, incorporating comprehensive financial domain knowledge graphs across different countries can help address the varying accounting standards and reporting practices globally.

Acknowledgments

Bo Jin, Xiao-Yang Liu Yanglet, and Steve Yang acknowledge the support from NSF IUCRC CRAFT center research grant (CRAFT Grant 22017) for this research. The opinions expressed in this publication do not necessarily represent the views of NSF IUCRC CRAFT. Shijie Han, Haoqiang Kang, and Xiao-Yang Liu Yanglet acknowledge the support from Columbia's SIRS and STAR Program, The Tang Family Fund for Research Innovations in FinTech, Engineering, and Business Operations.

All authors thanks Mohammed J. Zaki and anonymous reviewers for providing detailed revision comments.

References

[1] 2024. Qwen2 Technical Report. (2024).
[2] Toyin Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, Charese Smiley, and Sameena Shah. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. *Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (2024).
[3] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
[4] Anonymous. 2024. Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering. arXiv:2402.11194 [cs.CL]
[5] Elizabeth Blankespoor. 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research* 57, 4 (2019), 919–967.
[6] Enrique Bonsón, Virginia Cortijo, and Tomás Escobar. 2009. Towards the global adoption of XBRL using International Financial Reporting Standards (IFRS). *International Journal of Accounting Information Systems* 10, 1 (2009), 46–60.
[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
[8] Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, KP Subbalakshmi, and Papa Momar Ndiaye. 2024. RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data. *arXiv preprint arXiv:2404.07452* (2024).
[9] Yufei Chen and Jianfeng Liu. 2023. Challenges in applying large language models to financial data analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1267–1279.
[10] I de Zarzà, J de Curtò, Gemma Roig, and Carlos T Calafate. 2023. Optimized financial planning: Integrating individual and cooperative budgeting models with llm recommendations. *AI* 5, 1 (2023), 91–114.
[11] Roger Debreceny, Stephanie Farewell, Maciej Piechocki, Carsten Felden, and André Gräning. 2010. Does it add up? Early evidence on the data quality of XBRL filings to the SEC. *Journal of Accounting and Public Policy* 29, 3 (2010), 296–306.
[12] Roger Debreceny, Carsten Felden, Bartosz Ochocki, Maciej Piechocki, and Michał Piechocki. 2009. *XBRL for Interactive Data: Engineering the Information Value Chain*. Springer Science & Business Media, Berlin.
[13] Roger S Debreceny, Stephanie M Farewell, Maciej Piechocki, Carsten Felden, Andre Gräning, and Alessandro d'Eri. 2011. Flex or break? Extensions in XBRL

disclosures to the SEC. *Accounting Horizons* 25, 4 (2011), 631–657.
[14] Hui Du, Miklos A Vasarhelyi, and Xiaochuan Zheng. 2013. XBRL mandate: Thousands of filing errors and so what? *Journal of Information Systems* 27, 1 (2013), 61–78.
[15] Charles Hoffman and Liv Apneseth Watson. 2009. *XBRL for Dummies*. John Wiley & Sons, Hoboken, NJ.
[16] Rani Hoitash and Udi Hoitash. 2018. Measuring accounting reporting complexity with XBRL. *The Accounting Review* 93, 1 (2018), 259–287.
[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. [n. d.]. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.
[18] Tim Hopkinson and Carlo Magni. 2016. Net Present Value as a Decision-Making Tool in Project Management and Capital Budgeting. *Advances in Economics, Business, and Management Research* 219 (2016), 110–121.
[19] Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research* 40, 2 (2023), 806–841.
[20] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944* (2023).
[21] Diane J Janvrin and Won Gyun No. 2012. XBRL implementation: A field investigation to identify research opportunities. *Journal of Information Systems* 26, 1 (2012), 169–197.
[22] Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination. *Workshop on Failure Modes in the Age of Foundation Models, NeurIPS* (2023).
[23] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. 2022. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445* (2022).
[24] Devrimi Kaya. 2014. The influence of firm-specific characteristics on the extent of voluntary disclosure in XBRL: Empirical analysis of SEC filings. *International Journal of Accounting and Information Management* 22, 1 (2014), 2–17.
[25] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[27] Wei Li and Jane Zhang. 2023. Large language models for XBRL: Opportunities and challenges. *Journal of Information Systems* 37, 1 (2023), 53–72.
[28] Hongwei Liu et al. 2024. MathBench: A Hierarchical Mathematics Benchmark for LLMs. In *Findings of the Association for Computational Linguistics*.
[29] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. Data-centric FinGPT: Democratizing internet-scale data for financial large language models. *Workshop on Instruction Tuning and Instruction Following, NeurIPS* (2023).
[30] Xiao-Yang Liu, Jie Zhang, Guoxuan Wang, Weiqing Tong, and Anwar Walid. 2024. FinGPT-HPC: Efficient pretraining and finetuning large language models for financial applications with high-performance computing. *arXiv preprint arXiv:2402.13533* (2024).
[31] Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *the Journal of Finance* 69, 4 (2014), 1643–1671.
[32] Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54, 4 (2016), 1187–1230.
[33] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining* 16, 1 (2023), 20.
[34] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842* (2023).
[35] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. [n. d.]. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
[36] Jennifer Neglia. 2023. Achieving trust and efficient financial reporting through XBRL. https://www.pwc.com/us/en/services/trust-solutions/xbrl.html
[37] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903* (2024).
[38] Victor A Onuchak and Maria A Khalturina. 2022. Transition of Business Information to Electronic Exchange: Foreign and Russian Experience. In *Industry 4.0: Fighting Climate Change in the Economy of the Future*. Springer, 327–336.
[39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[40] Robert Pinsker and Shaomin Li. 2008. Costs and benefits of XBRL adoption: Early evidence. *Commun. ACM* 51, 3 (2008), 47–50.
[41] Karlo Puh and Marina Bagić Babac. 2023. Predicting stock market using natural language processing. *American Journal of Business* 38, 2 (2023), 41–61.
[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
[43] Jesse Roberts, Kyle Moore, and Doug Fisher. 2024. Do Large Language Models Learn Human-Like Strategic Preferences? *arXiv preprint arXiv:2404.08710* (2024).
[44] Ali Saeedi, Jim Richards, and Barry Smith. 2007. An Introduction to XBRL. In *British Accounting Association's Annual Conference*.
[45] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2024).
[46] Gemma Team. 2024. Gemma. (2024). https://doi.org/10.34740/KAGGLE/M/3301
[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[48] Michele Tufano, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. 2024. AutoDev: Automated AI-Driven Development. *arXiv preprint arXiv:2403.08299* (2024).
[49] Aliaksei Vertsel and Mikhail Rumiantsau. 2024. Hybrid LLM/Rule-based Approaches to Business Insights Generation from Structured Data. *arXiv preprint arXiv:2404.15604* (2024).
[50] Hui Wang and Rachel Johnson. 2023. Ethical considerations in AI-driven financial analysis: A focus on XBRL and large language models. *AI and Ethics* 3, 2 (2023), 215–228.
[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
[52] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[53] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Xiao-Yang Liu, et al. 2024. The FinBen: An holistic financial benchmark for large language models. *NeurIPS, Special Track on Datasets and Benchmarks* (2024).
[54] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *International Conference on Neural Information Processing Systems* (2023).
[55] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *FinLLM at IJCAI* (2023).
[56] Steve Yang, Fang-Chun Liu, and Xiaodi Zhu. 2015. An exploratory study of financial reporting structures: A graph similarity approach using XBRL. *Stevens Institute of Technology School of Business Research Paper* 2015-58 (2015).
[57] Steve Yang, Fang-Chun Liu, and Xiaodi Zhu. 2018. The impact of XBRL on financial statement structural comparability. In *Network, Smart and Open: Three Keywords for Information Systems Innovation*. Springer, 193–206.
[58] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.
[59] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *FinLLM at IJCAI* (2023).
[60] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *ACM International Conference on AI in Finance*. 349–356.
[61] Chao Zhang, Yuren Mao, Yijiang Fan, Yu Mi, Yunjun Gao, Lu Chen, Dongfang Lou, and Jinshu Lin. 2024. FinSQL: Model-Agnostic LLMs-based Text-to-SQL Framework for Financial Analysis. In *International Conference on Management of Data*. 93–105.