Advancements in Conversational Finance AI: Insights from the ConvFinQA Dataset

David George Williams

Wednesday, March 6, 2024

Executive Summary
In the transformative era of conversational finance AI, our study delves into the ConvFinQA dataset's remarkable potential to revolutionize financial analytics and decision-making processes, presenting a compelling case for Tomoro AI's pioneering stance in the AI-native future of finance. Through a meticulous comparative analysis between GPT-4 and Claude-3, we uncover the nuanced capabilities of Large Language Models (LLMs) in navigating complex financial inquiries, highlighting Claude-3's superior performance and the importance of refined prompt engineering and error mitigation strategies.

This research not only benchmarks the current state of AI in finance but also charts a strategic path for Tomoro AI, emphasizing the critical role of continuous model evaluation and adaptation in staying at the forefront of technological innovation. Our findings, underpinned by an extensive literature review and a bespoke selection of the most relevant and insightful papers, serve as a beacon for Tomoro AI, showcasing how cutting-edge AI solutions can address existing challenges and unlock new opportunities in the financial sector, ensuring the company remains a leader in the transformative journey towards an AI-driven financial landscape.

Introduction
The financial sector stands on the brink of a transformative era with the advent of conversational finance AI, yet it navigates a landscape riddled with both challenges and unprecedented opportunities. The primary hurdle lies in the inherent complexity of financial data—its interpretation demands not only a granular understanding of numeric information but also the context in which this data operates. Moreover, the financial domain is regulated by stringent compliance and privacy requirements, further complicating the deployment of AI technologies.

However, these challenges are mirrored by significant opportunities. Conversational finance AI promises to revolutionize customer interactions, enabling personalized financial advice at scale and automating routine inquiries, thereby freeing human agents to tackle more complex customer needs. Furthermore, AI's capacity to digest and analyze vast datasets can unearth insights that were previously inaccessible, driving more informed decision-making and identifying novel revenue streams.

In this evolving landscape, the ConvFinQA dataset emerges as a pivotal resource for pushing the boundaries of what conversational finance AI can achieve. By providing a rich array of financial questions and answers, alongside supporting data and figures, ConvFinQA offers a foundation for training AI models to navigate the complexities of financial reasoning and numeric data interpretation. This dataset not only challenges AI systems with real-world scenarios but also serves as a benchmark for evaluating their ability to perform long-range, complex numerical reasoning—a critical capability for applications within the financial sector.

Albert Phelps's insights on "Prompt Engineering is dead* or is it just beginning?" from Tomoro AI, touch on the evolution of prompt engineering and its significance in enhancing AI's interaction with complex domains like finance. The shift towards more sophisticated human-AI workflows and meta-prompting, as discussed by Phelps, underscores the necessity of developing AI models that can understand and generate nuanced responses based on the intricate demands of financial data. This evolution from simple prompt-response interactions to complex, context-aware dialogues aligns perfectly with the challenges and opportunities presented by conversational finance AI. It exemplifies how advancements in AI, as represented by the study and application of datasets like ConvFinQA, are crucial for overcoming the sector's hurdles and unlocking its full potential.

Through the lens of Tomoro AI's exploration into the frontiers of AI application, we see a clear pathway for leveraging the ConvFinQA dataset. It not only furthers our understanding of conversational finance AI's capabilities but also aligns with the broader mission of transforming financial services through deep AI integration. This endeavor, powered by datasets like ConvFinQA and informed by cutting-edge research and practice in prompt engineering, marks a significant step toward realizing an AI-native future for the finance sector.

Relevance to Tomoro AI

The integration and evaluation of Large Language Models (LLMs) within financial domains present a unique intersection of opportunity and challenge, particularly for Tomoro AI, a pioneer in the transformative application of AI technologies. The creation of a GPT-4 based intelligent agent, as demonstrated through the processing of the ConvFinQA dataset, underscores an essential stride toward harnessing the nuanced potential of LLMs in finance. This approach not only validates but also seeks to productionize findings from the comprehensive financial benchmark, FinBen, thereby addressing a critical industry gap— the systematic evaluation of LLMs against complex financial tasks.

Tomoro AI's mission to enable enterprise clients to realize competitive advantage through AI is directly complemented by this study's objectives and findings. By leveraging GPT-4's capabilities, we create an agent adept at navigating the intricacies of financial data, thereby offering a solution that is both innovative and necessary. This aligns seamlessly with Tomoro AI's efforts to enhance AI-driven decision-making tools and automate financial analyses, fortifying the company's position at the forefront of AI application in finance.

Moreover, the study's insights into the strengths and limitations of LLMs, including GPT-4, within the financial domain, provide Tomoro AI with invaluable data to tailor its AI solutions more precisely. The identification of areas requiring targeted enhancements, such as complex extraction and forecasting, informs Tomoro AI's strategic direction in developing or refining AI models that address these specific challenges. This not only augments the company's product offerings but also elevates its capability to solve existing challenges and unlock new opportunities within the financial sector.

Furthermore, the introduction of FinBen as a comprehensive, open-sourced evaluation benchmark for LLMs in finance represents a significant advancement in the field. For Tomoro AI, utilizing FinBen as a tool for continuous evaluation and improvement of its AI solutions can facilitate the development of more effective and efficient financial AI agents. This proactive approach to leveraging cutting-edge benchmarks ensures that Tomoro AI remains adaptive and responsive to the rapid advancements within AI technology, thereby maintaining its competitive edge.

In essence, this study's implementation of a GPT-4 based intelligent agent to navigate and elucidate the ConvFinQA dataset, coupled with the strategic application of FinBen, epitomizes Tomoro AI's commitment to pioneering the integration of AI within the financial domain. It underscores the company’s dedication to solving complex financial challenges through AI, reinforcing its mission to transform business operations and decision-making with state-of-the-art AI solutions. This alignment not only enhances Tomoro AI's product suite but also solidifies its role as a leader in the AI-driven transformation of finance, propelling the company towards realizing its vision of an AI-native future for its clients and the broader industry.

Related Work

The advent of AI in the finance sector has ushered in a transformative era, marked by the integration of complex numerical reasoning and conversational interfaces to interpret financial data. This literature review navigates through seminal works that have shaped our understanding and application of AI in finance, underpinning the research presented in this paper.

Early Contributions and Language Models Jun et al. (2022) underscored the predominance of text data in question answering systems, identifying a gap in datasets for table question answering in languages other than English. Their construction of Korean-specific datasets for table question answering, comprising 1.4M tables and 70k question-answer pairs, represents a foundational step towards diversifying AI's linguistic and format capabilities. This work not only highlights the importance of linguistic diversity but also the necessity of addressing the complexities inherent in tabular data.

Advancements in Conversational Finance AI Chen et al. (2022) introduced ConvFinQA, a pivotal dataset aimed at exploring the chain of numerical reasoning in conversational finance question answering. This dataset challenges existing AI models with its focus on long-range, complex numerical reasoning paths within real-world conversations. ConvFinQA's comprehensive experiments and analyses illuminate the reasoning mechanisms of neural symbolic and prompting-based methods, marking a significant stride towards emulating human-like reasoning in AI.

Hybrid Tabular-Textual Question Answering Zhang et al. (2023) proposed a novel non-autoregressive program generation framework, NAPG, to address the limitations of autoregressive decoding in hybrid tabular-textual question answering. By generating complete program tuples independently, NAPG significantly enhances program generation speed and reduces performance drop with increasing numerical reasoning steps, establishing new benchmarks in the field.

Optimizing Training Approaches Sun et al. (2023) introduced APOLLO, an optimized training approach that significantly advances the long-form numerical reasoning framework. By adopting a number-aware negative sampling strategy and designing consistency-based reinforcement learning, APOLLO achieves state-of-the-art performance, showcasing the importance of tailored training strategies in AI's reasoning performance.

Domain-Specific Large Language Models Wu et al. (2023) unveiled BloombergGPT, a specialized large language model trained on an extensive range of financial data. BloombergGPT's development, underscored by a mixed dataset training approach, demonstrates its superiority in financial tasks, emphasizing the need for domain-specific models to address the nuanced challenges of financial NLP.

General-Purpose Models and Financial Text Analytics Li et al. (2023) embarked on an empirical study to evaluate the effectiveness of general-purpose models like ChatGPT and GPT-4 in the financial domain. Their work highlights the strengths and limitations of these models against domain-specific pretrained models, urging further improvements to enhance their capability in financial text analytics.

Instructional Data and Evaluation Benchmarks Xie et al. (2023) introduced PIXIU, a comprehensive framework featuring the first financial LLM, FinMA, based on fine-tuning LLaMA with instruction data. PIXIU's contribution, including an evaluation benchmark covering critical financial tasks, marks a significant advancement in the open-source development of financial AI.

Evaluating Financial Reasoning Capabilities Callanan et al. (2023) assessed the financial reasoning capabilities of LLMs using mock CFA exams. Their in-depth analysis provides valuable insights into the models' performance and outlines potential strategies to enhance LLMs' applicability in finance.

Quantitative Reasoning Benchmarks Koncel-Kedziorski et al. (2023) introduced BizBench, a benchmark aimed at evaluating models' ability to reason about financial problems through program synthesis. This benchmark evaluates a model's financial background knowledge and its capacity to solve problems with code, contributing to a deeper understanding of AI's quantitative reasoning in finance.

Zero-Shot Question Answering and Financial Document Analysis Phogat et al. (2023) explored a zero-shot approach to answering complex questions requiring multi-hop numerical reasoning over financial reports. Their novel prompting technique significantly improves accuracy, demonstrating the potential of LLMs in extracting complex domain-specific numerical reasoning.

Comprehensive Surveys and Benchmarks Lee et al. (2024) provided a survey of FinLLMs, offering a chronological overview, performance evaluations, and discussing opportunities and challenges. Similarly, Yuan et al. (2024) introduced FinLLMs, a framework for generating financial question-answering data using Large Language Models, addressing the challenge of creating numerical reasoning datasets in the financial domain.

Integration of LLMs in Finance Zhao et al. (2024) provided an overview of applications and insights into the integration of LLMs into various financial tasks, showcasing their potential in automating financial report generation, forecasting market trends, and more. This comprehensive survey underlines the growing momentum of LLM deployment in finance.

Evaluation of Mathematical Reasoning Srivastava et al. (2024) delved into LLMs' mathematical reasoning capabilities within financial document question answering, introducing a novel prompting technique tailored to semi-structured documents. This work enhances our understanding of LLMs' abilities in complex mathematical scenarios.

Holistic Financial Benchmarking Xie et al. (2024) introduced FinBen, the first comprehensive evaluation benchmark for assessing LLMs in the financial domain. FinBen's evaluation of 15 representative LLMs reveals their strengths and limitations, indicating a need for targeted enhancements in complex extraction and forecasting.

This literature review underscores the dynamic evolution of AI in finance, from early dataset construction and hybrid question answering to the development of domain-specific models

and comprehensive evaluation benchmarks. The research presented in this paper builds upon these foundational works, aiming to further the exploration of numerical reasoning in conversational finance AI, contributing to the advancement of AI applications in the finance sector.

Methodology

I meticulously designed the methodology for evaluating and exploring Large Language Models (LLMs) for the ConvFinQA dataset to harness the capabilities of best-in-class models, ensuring their application is easy and efficient. A robust approach to prompt engineering, refined through my extensive experience across a diverse range of topics, underpins this process. At its core, I developed a custom prompt and encapsulated it within a function to ensure conversational accuracy and naturalism.

Prompt Engineering Insights

The function *create_message_body* plays a pivotal role in generating the prompt structure. It meticulously processes the given data to produce a cohesive document, emphasizing the correct use of pluralization to mirror natural conversational patterns. This process involves:

- Simplifying and unifying the *pre_text*, *table*, and *post_text* fields into a single, fluid document format, thereby avoiding the fragmented nature of JSON.

- Employing regular expressions to clean and streamline the text and table data, ensuring readability and consistency.

- Dynamically adjusting the language based on the number of questions involved, thus maintaining linguistic accuracy.

By concatenating various components of the dataset (*pre_text*, *table*, and *post_text*), the function crafts a uniform document. The table data, joined by tab characters, is integrated seamlessly with narrative elements to provide a comprehensive view of the information. This refined input is crucial for the LLM's processing, as it relies on natural text rather than structured formats for analysis.

Adapting to New Developments

The introduction of Anthropic's Claude-3 on March 4th offered an unforeseen opportunity to elevate the research. Claude-3's promise of enhanced performance was corroborated by the accuracy improvements documented in the "Experiments and Results" section. This development shifted the analytical focus from assessing a single model's accuracy to comparing the relative accuracies of two models. By maintaining consistency in prompt engineering and output evaluation methodologies, the analysis facilitated a direct, equitable comparison between GPT-4 and Claude-3.

Implications for Tomoro AI

This methodological approach, centered around precise prompt engineering and adaptive analysis, enables Tomoro AI to make informed decisions regarding model utilization. By focusing on relative model comparisons, Tomoro AI can adopt a strategic, hill-climbing approach to model selection and evaluation. This strategy not only ensures continuous improvement in AI-driven solutions but also upholds fairness and ethical considerations in model deployment. Ultimately, the methodology detailed here lays the groundwork for Tomoro AI to navigate the evolving landscape of conversational finance AI, optimizing for both performance and ethical standards.

Experiments and Results

In the pursuit of understanding the capabilities of Large Language Models (LLMs) in the domain of conversational finance AI, a comparative analysis between GPT-4 and Claude-3 was conducted utilizing the ConvFinQA dataset. This analysis was designed to ensure fairness and clarity by holding constant the variables of the dataset used, the prompts provided to the models, and the functions employed for parsing and computing accuracy.

Key Findings

The core of my analysis lies in the direct comparison of model accuracies, offering a clear view of how GPT-4 and Claude-3 perform under uniform conditions. This comparative approach has yielded significant insights:

- **GPT-4** demonstrated an accuracy of 58.348%.

- **Claude-3** outpaced GPT-4 with an accuracy of 73.194%.

These results reveal that, when evaluated against the same ConvFinQA dataset and under equivalent conditions, Claude-3 significantly outperformed GPT-4, achieving roughly 73% accuracy in responding to queries, as opposed to GPT-4's 58%. This differential underscores the distinct capabilities of these models in processing and understanding complex financial data.

Sources of Error and Mitigation Strategies

The experiment identified several sources of potential error that could affect the outcome:

- **Dataset Errors**: Instances were observed where the ConvFinQA dataset provided non-numerical or ambiguously parsable answers, which could lead to discrepancies in evaluating model responses.

- **Variability in Expressing Percentages**: The dataset and LLM outputs varied in expressing percentages, either numerically or as strings with a percentage sign, along with semantic interpretations of percentage changes.

- **Rounding Differences**: Subtle differences in rounding between the LLMs' outputs and those in the ConvFinQA dataset were noted.

To address these challenges and ensure a level playing field, I implemented a function named *essentially_equals*. This function deemed the LLMs' output and the ConvFinQA dataset's answers as equivalent if they were within a 0.5 absolute difference or within a 1% margin of the target value. This innovative approach allowed for a fair comparison by accommodating minor discrepancies and emphasizing the conceptual correctness of the answers.

Practical Implications for Tomoro AI

The experimental results underscore several opportunities for Tomoro AI to leverage LLMs in enhancing its conversational finance AI offerings:

- **Model Selection**: The superior performance of Claude-3 suggests a potential preference for utilizing this model in applications requiring high accuracy in financial question answering.

- **Error Handling and Data Parsing**: The strategies developed to mitigate errors and parse responses can be integrated into Tomoro AI's systems to improve reliability and user trust in automated financial advice.

- **Custom Evaluation Metrics**: Adopting customized evaluation metrics like *essentially_equals* can enhance the precision of Tomoro AI's AI-based solutions, ensuring they accommodate the nuances of financial data.

Furthermore, acknowledging and adjusting for dataset imperfections and semantic variations in financial terminology can refine Tomoro AI's model training processes. This, in turn, paves the way for developing more robust, accurate, and user-friendly conversational finance AI applications.

This comparative analysis not only highlights the current state of LLM performance in financial question answering but also outlines a roadmap for Tomoro AI to harness these insights. By carefully selecting models and tailoring evaluation strategies, Tomoro AI can advance its mission to deliver cutting-edge AI-driven financial solutions, furthering its leadership in the transformative era of conversational finance AI.

Conclusion

This study embarked on a journey through the complexities and potentials of conversational finance AI, illuminated by the ConvFinQA dataset. The meticulous comparative analysis between GPT-4 and Claude-3 models underlined the significant advancement that conversational finance AI has achieved, and poised to contribute towards the financial sector's transformative era. Key takeaways underscore Claude-3's superior performance over GPT-4 within the specific context of financial question answering, reflecting a notable accuracy rate that highlights the rapid evolution of Large Language Models (LLMs) and their application in understanding and processing complex financial inquiries.

The exploration revealed not only the capabilities and limitations of current AI technologies but also the importance of nuanced prompt engineering and error mitigation strategies to harness these tools effectively. By adopting a structured and analytical approach to model evaluation, this research presents a method for Tomoro AI to continually refine and leverage AI solutions, ensuring they remain at the cutting edge of technology while adhering to ethical standards and practical effectiveness.

I am immensely excited about the prospect of contributing to Tomoro AI, armed with the insights and methodologies honed during this analysis. The possibility of applying my skills to further enhance Tomoro AI's offerings and help realize its ambitious vision for the future of conversational finance AI is a motivating prospect. The advancement of AI in finance is not just a technical challenge but a transformative opportunity to redefine the interaction between finance professionals and their tools, and ultimately, the customers they serve.

In conclusion, this study not only highlights the potential of conversational finance AI but also charts a course for Tomoro AI's continued leadership and innovation in this space. I look forward to the opportunity to contribute further to this exciting journey, leveraging my skills and insights to support Tomoro AI's mission to pioneer the AI-native future of finance.

References

At the outset of my research into the profound intersections between Large Language Models (LLMs) and financial analytics, I delved into the seminal work "ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering." An initial

examination led me to the Connected Papers graph for this study, accessible at https://tinyurl.com/yssvwhdw.

This graph provided a visually intuitive exploration of the paper's academic context and its interconnections with related research. However, to ensure a comprehensive understanding of ConvFinQA's contributions and its broader implications for the financial sector, I embarked on a bespoke search through arXiv.org, focusing specifically on the development and application of LLMs within the financial domain.

This tailored exploration, driven by a quest to encompass the past, present, and future impacts of ConvFinQA and its relevance to business, unearthed a selection of references markedly superior in relevance and insight than those initially indicated by the Connected Papers graph. Through a meticulous curation process, I have compiled a list of key scholarly articles that significantly enrich our comprehension of how LLMs are revolutionizing financial question answering and analytical processes.

Herein lies the final list of papers that underpinned the foundational research for this work, serving as a testament to the dynamic evolution of LLMs in finance and their potential to reshape our understanding of numerical reasoning within this vital sector.

Callanan, E., Mbakwe, A., Papadimitriou, A., Pei, Y., Sibue, M., Zhu, X., Ma, Z., Liu, X., & Shah, S. (2023). Can GPT models be Financial Analysts? An Evaluation of ChatGPT and GPT-4 on mock CFA Exams. *arXiv:2310.08678*. https://doi.org/10.48550/arXiv.2310.08678

Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., & Wang, W. Y. (2022). ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. *arXiv:2210.03849*. https://doi.org/10.48550/arXiv.2210.03849

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering. *arXiv:2311.11944*. https://doi.org/10.48550/arXiv.2311.11944

Jun, C., Choi, J., Sim, M., Kim, H., Jang, H., & Min, K. (2022). Korean-Specific Dataset for Table Question Answering. *arXiv:2201.06223v2*. https://doi.org/10.48550/arXiv.2201.06223

Koncel-Kedziorski, R., Krumdick, M., Lai, V., Reddy, V., Lovering, C., & Tanner, C. (2023). BizBench: A Quantitative Reasoning Benchmark for Business and Finance. *arXiv:2311.06602*. https://doi.org/10.48550/arXiv.2311.06602

Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A Survey of Large Language Models in Finance (FinLLMs). *arXiv:2402.02315*. https://doi.org/10.48550/arXiv.2402.02315

Li, J., Bian, Y., Wang, G., Lei, Y., Cheng, D., Ding, Z., & Jiang, C. (2023). CFGPT: Chinese Financial Assistant with Large Language Model. *arXiv:2309.10654v2*. https://doi.org/10.48550/arXiv.2309.10654

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. *arXiv:2305.05862v2*. https://doi.org/10.48550/arXiv.2305.05862

Phogat, K. S., Harsha, C., Dasaratha, S., Ramakrishna, S., & Puranam, S. A. (2023). Zero-Shot Question Answering over Financial Documents using Large Language Models. *arXiv:2311.14722*. https://doi.org/10.48550/arXiv.2311.14722

Srivastava, P., Malik, M., Gupta, V., Ganu, T., & Roth, D. (2024). Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering. *arXiv:2402.11194v2*. https://doi.org/10.48550/arXiv.2402.11194

Sun, J., Zhang, H., Lin, C., Gong, Y., Guo, J., & Duan, N. (2023). APOLLO: An Optimized Training Approach for Long-form Numerical Reasoning. *arXiv:2212.07249v2*. https://doi.org/10.48550/arXiv.2212.07249

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. *arXiv:2303.17564v3*. https://doi.org/10.48550/arXiv.2303.17564

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., Deng, Z., Jiang, Y., Yao, Z., Li, H., Yu, Y., Hu, G., Huang, J., Liu, X.-Y., Lopez-Lira, A., Wang, B., Lai, Y., Wang, H., Peng, M., Ananiadou, S., & Huang, J. (2024). The FinBen: An Holistic Financial Benchmark for Large Language Models. *arXiv:2402.12659*. https://doi.org/10.48550/arXiv.2402.12659

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2023). PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. *arXiv:2306.05443*. https://doi.org/10.48550/arXiv.2306.05443

Yuan, Z., Wang, K., Zhu, S., Yuan, Y., Zhou, J., Zhu, Y., & Wei, W. (2024). FinLLMs: A Framework for Financial Reasoning Dataset Generation with Large Language Models. *arXiv:2401.10744*. https://doi.org/10.48550/arXiv.2401.10744

Zhang, L., Cai, W., Liu, Z., Yang, Z., Dai, W., Liao, Y., Qin, Q., Li, Y., Liu, X., Liu, Z., Zhu, Z., Wu, A., Guo, X., & Chen, Y. (2023). FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. *arXiv:2308.09975*. https://doi.org/10.48550/arXiv.2308.09975

Zhang, T., Xu, H., van Genabith, J., Xiong, D., & Zan, H. (2023). NAPG: Non-Autoregressive Program Generation for Hybrid Tabular-Textual Question Answering. *arXiv:2211.03462v2*. https://doi.org/10.48550/arXiv.2211.03462

Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., Xu, S., Dai, H., Zhao, L., Mai, G., Liu, N., & Liu, T. (2024). Revolutionizing Finance with LLMs: An Overview of Applications and Insights. *arXiv:2401.11641*. https://doi.org/10.48550/arXiv.2401.11641