Empowering Financial Intelligence: Unraveling the Impact of Large Language Models in Downstream Finance Tasks

Peter Findley, Tanzila Binti Alam, Asif Mahmud, Abul Kalam Faruk, Yeahia Sarker
Anchorblock Technology LLC
{peter, tanzila, asif, faruk, yeahia}@anchorblock.vc

Abstract

This study provides a thorough assessment of General Purpose Large Language Models (LLMs) on a range of financial datasets, followed by a comparison of their performance with models from the BloombergGPT paper. Examining the performance of Open-llama-7B-instruct, Llama-2-7B-chat, and OpenAI's GPT-3.5 in various financial NLP tasks is the main goal of the study. The datasets include named entity recognition, news headline classification, aspect-specific sentiment analysis, sentiment classification, and question-answering over financial data. Results show the general-purpose LLMs' promising performance across the datasets, highlighting their potential for financial applications and providing insightful information for upcoming open finance research.

1 Introduction

Large Language Models (LLMs), with their remarkable language understanding capabilities, have been a transformative force in the financial industry, revolutionizing data-driven decision-making and automation. Thanks to LLMs like GPT-3.5, financial companies, and professionals now have access to powerful tools that can automate economic research, enhance customer service, leverage data to drive strategies for investment, manage risk, uncover fraud, and optimize portfolios. Using their capacity to analyze vast amounts of textual data, LLMs provide essential insights from financial reports, news articles, and market sentiment data, providing real-time, accurate information to investors, traders, and analysts. However, there are limitations to this emerging technology, such as the interpretability of models, ethical concerns, and potential data biases. To effectively use LLMs in finance while addressing these issues, it is crucial to do continuing research, collaborate, and employ LLMs in the right ways. A more data-driven and informed financial landscape will be made feasible as a result.

‘

Large language Models. The introduction of large language models (LLMs) is a groundbreaking advancement in natural language processing (NLP). Researchers have scaled pre-trained language models (PLMs) by increasing model size or data size, improving model performance on downstream tasks. Significant research and applications in many areas, including NLP, information retrieval, and computer vision, have been sparked by LLMs. Large parameter size LLMs, such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), demonstrate startling emergent skills, addressing complicated problems that smaller PLMs are unable to. Notably, ChatGPT, based on GPT-4 (OpenAI, 2023), exhibits an amazing capacity for human dialogue.

A recent publication proposes GPT-4 as a prototype for an AGI system (Bubeck et al., 2023). Research in numerous AI fields is revolutionized by LLMs' quick development. LLMs play a significant role in general-purpose language problem-solving in NLP, which has caused a paradigm shift in research approaches. LLMs like ChatGPT are posing a threat to established search engines, and initiatives like New Bing have aimed to improve search results by utilizing LLMs. To enable multimodal dialogues, researchers in the Computer Vision field are working to create vision-language models similar to ChatGPT. GPT-4 has facilitated the integration of multimodal input by including visual data. This wave of technological development could support a vibrant ecosystem of LLM-based real-world applications. Notably, OpenAI makes use of ChatGPT plugins to enable specialized functionalities, whereas Microsoft 365 integrates LLMs like Copilot to automate office tasks.

However, there are still certain issues to be resolved, such as comprehending the underlying theories driving LLMs' enhanced abilities, the high computational cost of training, and addressing po-

Table 1: Dataset Information. This table provides details about various datasets used in the study, including their names, types, and sizes for training, validation, and testing.

Dataset Name	Type	Dataset Size		
		Train	Valid	Test
FPB	Sentiment Classification	3488	388	969
FiQA SA	Sentiment Analysis	822	117	234
Headline	News Headlines Classification	7989	1141	2282
NER	Named Entity Recognition	932	232	302
ConvFinQA	Question Answering	900	180	320

tential dangers associated with the creation of damaging content. Despite these difficulties, LLMs have the potential to bring revolution.

Large Language Models in Finance: BloombergGPT (Wu et al., 2023) is a large language model with 50 billion parameters that was trained using a massive dataset of 363B tokens derived from financial data sources and an additional 345B tokens from general-purpose datasets. Without losing its performance on general LLM benchmarks, BloombergGPT surpasses previous models on financial benchmarks when tailored to financial activities. This model is an invaluable tool for a variety of financial technology applications, including sentiment analysis, named entity recognition and question answering due to its accuracy and domain-specific expertise.

Given that BloombergGPT was created and optimized specifically for financial activities, it is intriguing to consider how general-purpose LLMs would perform in similar downstream financial applications.

In order to do that, we plan to compare general-purpose LLMs' ability for handling financial tasks, and then compare their performance to BloombergGPT. We test these models on a variety of downstream finance tasks to investigate their advantages, disadvantages, and future applications in the financial sector.

2 Datasets for benchmarking

In this section, we provide details on the datasets used in our research to evaluate the performance of various Large Language Models (LLMs) on finance-related tasks. These datasets have been carefully curated to represent different aspects of financial sentiment analysis, opinion-based question answering, named entity recognition, and numerical reasoning over structured financial data.

A. **FPB (Financial Polar Sentiment Dataset)**: This dataset (Malo et al., 2014) comprises 4840 sentences extracted from English language financial news articles, annotated with polar sentiment labels. The sentences are categorized based on the agreement rate of 5-8 annotators, ensuring diverse and reliable sentiment annotations.

B. **FiQA SA (Financial Question Answering Sentiment Analysis Challenge)**: This dataset (Macedo Maia, 2018) involves predicting aspect-specific sentiment in English financial news and microblog headlines, released as part of the 2018 challenge on financial question answering and opinion mining. Although the original dataset had continuous sentiment annotations, we convert it into a classification setup, with negative, neutral, and positive classes. Similar to our approach with the FPB dataset, we create our own data splits, incorporating both microblogs and news and evaluate the model's performance using a 5-shot setup, reporting the weighted F1 score.

C. **Headline Dataset**: This dataset (Sinha and Khandait, 2020) facilitates a binary classification task to determine whether a news headline in the gold commodity domain includes specific information. Human-annotated, the dataset consists of English news headlines about "gold" with tags such as "price up," "price down," "price stable," "past price," "future price," "past general," "future general," and "asset comparison". Each tag is transformed into a question, and performance is evaluated using a 5-shot setup with an average weighted F1 score.

D. **NER (Named Entity Recognition) Dataset**: This dataset (Salinas Alvarado et al., 2015) involves named entity recognition on financial data collected for credit risk assessment

Table 2: Evaluation results (accuracy) of Llama 7B and Open Llama 7B on Common Sense QA datasets

Model	hellaswag	truthfulqa mc mc1	arc challenge	openbookqa
LLaMA 7B	0.56	0.21	0.68	0.29
OpenLLaMA 7B	0.53	0.23	0.72	0.30

Table 3: Performance comparison between Llama-2, Falcon, and MPT 7B and GPT 3.5 175B models on Common Sense QA dataset evaluation

Model	MMLU	TriviaQA	Hellaswag	OpenbookQA	Winogrande
MPT 7B	26.8	59.6	76.4	51.4	68.3
Falcon 7B	26.2	56.8	74.1	51.6	66.3
Llama-2 7B	45.3	68.9	77.2	58.6	69.2
GPT 3.5 175B	70	–	85.5	–	81.6

from financial agreements filed with the SEC. Annotated with entity types following the CoNLL format, including PER, LOC, ORG, and MISC, the dataset focuses on predicting named entities in few-shot setups. Sentences without any entities and MISC tags are excluded, and entity-level F1 scores are reported with a 9-shot setup.

E. **ConvFinQA (Conversational Financial Question Answering) Dataset**: This dataset (Chen et al., 2022) challenges models to answer conversational questions based on SP 500 earnings reports containing textual and structured financial data. The task requires numerical reasoning, an understanding of financial concepts, and relating follow-up questions to dialog turns. The model receives the entire gold conversation as input and context, and exact match accuracy on the public development set is reported.

These diverse and carefully curated datasets (table 1) provide comprehensive evaluation scenarios for LLMs in finance-related tasks, enabling us to assess their capabilities, strengths, and limitations in the financial domain.

3 Model

Open llama 7B open instruct model: A model utilized in this study is open-llama-7B-open-instruct by VMware, which is an instruction-tuned version of the pretrained Open Llama 7B (Geng and Liu, 2023) language model. The Open Llama models represent a permissively licensed collection of open-source replicas derived from Meta AI's esteemed LLaMA large language model. In the pursuit of disseminating advanced language capabilities to the research community, they introduced a set of meticulously crafted models: the 3B, 7B, and 13B variants, all of which have been rigorously trained on an extensive dataset comprising 1 trillion tokens. The Open Llama models were evaluated on various evaluation datasets using lm-evaluation-harness (Gao et al., 2021) powered by Eluther AI. The evaluation results (table 2) show that Open Llama models perform almost as good as the Llama models (Touvron et al., 2023) trained by Meta.

Using the Alpaca prompt template, the Open llama 7B model was improved with a focus on instruction-based learning. This instruct model's training data comes from the Open-instruct-v1 dataset (oas, 2023), which also contains the oasst, dolly, and hhrlhf datasets.

We aim to investigate the performance of this instruction-tuned model on financial datasets and assess its efficiency in comparison to existing financial language models in the field.

Llama-2 7B chat model: Llama 2 is a family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models with parameter scales ranging from 7 billion to 70 billion. It is the succession of Meta's Llama model series. It was trained on 40% more data than the first Llama model series. The Llama-1 models were not open-source, but the Llama-2 models are. That's why we have chosen the Llama-2 models for our research. Also, Llama-2 models have demonstrated superior performance (table-3) than most models in various benchmarks and hold their ground in human evaluations for helpfulness and safety, comparable to popular closed-source models like ChatGPT and PaLM. In this study, we specifically focus

Table 4: Comparison of accuracy between general purpose LLMs and Financial LLMs

Model	FPB	FiQA SA	Headline	NER (9-shots)	ConvFinQA (1-shot)
Open llama 7B instruct	56.34	70.91	83.36	49.08	–
Llama-2 7B chat	72.97	47.93	45.88	23.02	–
GPT-3.5 175B	47.62	77.53	87.21	58.14	–
BloombergGPT 50B	51.07	75.07	82.20	60.82	43.41
BLOOM 176B	50.25	53.12	76.51	55.56	36.31

on the 7B fine-tuned model known as Llama-2-Chat, optimized for dialogue use cases.

Regarding training data, Llama 2 was pretrained on an extensive dataset of 2 trillion tokens derived from publicly available sources. For fine-tuning, both publicly available instruction datasets and over one million new human-annotated examples were used.

GPT-3.5 Model: The GPT-3.5 language model is a very powerful and cutting-edge generative text model created by OpenAI. This model, which has an incredible size of 175 billion parameters, is made to perform well in a variety of natural language processing (NLP) applications. The GPT-3.5 model demonstrates extraordinary ability in comprehending and producing language that is similar to that of humans by utilizing substantial pretraining on a large corpus of heterogeneous text data and fine-tuning on human-annotated instruction data. It has demonstrated remarkable performance across a range of benchmarks and has been widely used in a variety of applications, including chatbots, language translation, and question-and-answer systems. The GPT-3.5 model, which pushes the limits of language creation and understanding and opens up new possibilities for AI-driven language applications, is seen as a turning point in the field of NLP. OpenAI has performed evaluation (OpenAI, 2023) on different common sense QA datasets, and the result (table 3) is really promising.

4 Results & Analysis

In this study, we assessed the performance (table-4) of three cutting-edge language models on a variety of financial datasets: Open-llama-7B-instruct, Llama-2-7B-chat, and GPT-3.5. We compared their findings to those published in the BloombergGPT paper, which serves as a reliable benchmark in the financial industry. To ensure accurate results, the experimental design used cross-validation methods and hyperparameter adjustment. When using financial data, privacy, and security measures were scrupulously adhered to, and ethical issues were taken into account. The outcomes are thoroughly examined in the section that follows, which clarifies how well the language models perform in ensuing financial tasks.

Open llama 7B open instruct model results: In this study, we conducted a comprehensive evaluation of the Open-llama-7B-instruct language model using our curated datasets. Our assessment focused on assessing the model's F1 score in performing various financial tasks, except for ConvFinQA, we used exact match accuracy evaluation for this dataset. Notably, we observed an impressive score of 56.34% in the FPB dataset, 70.91% in the FiQA dataset, 83.36% on the Headline dataset, 49.08% on the NER dataset, and finally, (pending) exact match accuracy in the ConvFinQA dataset.

Llama-2 7B chat model results: In the context of our research, we rigorously evaluated the Llama-2 7B chat model on our curated financial datasets. Through extensive experimentation, we assessed its proficiency in handling various financial language tasks. Remarkably, the model demonstrated notable performance, achieving a remarkable F1 score of 72.97% in the FPB dataset, 47.93% in the FiQA dataset, 45.88% on the Headline dataset, 23.02% on the NER dataset, and finally, (pending) exact match accuracy in the ConvFinQA dataset.

GPT 3.5 model results: We also conducted a comprehensive evaluation of the GPT-3.5 OpenAI model using our curated datasets. Our assessment focused on assessing the model's F1 score in performing various financial tasks. We observed a score of 47.62% in the FPB dataset, 77.53% in the FiQA dataset, 87.21% on the Headline dataset, 58.14% in the NER dataset, and finally, (pending) exact match accuracy in the ConvFinQA dataset.

We have observed the performance of various language models on different financial tasks. Notably, GPT-3.5 exhibited outstanding results, outperforming other models on the FiQA sentiment analysis (SA) and Headline datasets.

BloombergGPT and Open llama 7B instruct achieved competitive scores. Conversely, for the FPB task, the Llama-2 7B chat model demonstrated a remarkably high score, surpassing other models, while the Open Llama 7B instruct model closely followed with the second-best accuracy. Open Llama model achieving really good scores in the classification tasks symbolizes its expertise in those tasks. Also, these findings suggest that Llama models excel in sentiment and other classification tasks.

Regarding Named Entity Recognition (NER), a token classification task, Bloomberg GPT, GPT-3.5, and Bloom 176B showcased competitive results, with Bloomberg GPT achieving the highest accuracy of 60.82%. However, it is noteworthy that the Llama models struggled to deliver satisfactory scores in the NER task. This disparity in performance can partially be attributed to the fact that the Open Llama 7B instruct and Llama-2 7B chat models have a smaller parameter size of only 7B, which may affect their performance compared to larger models.

The most significant observation of our research lies in the remarkable performance of general-purpose LLMs on financial datasets. Surprisingly, in some cases, these LLMs even outperformed specialized financial LLMs on financial tasks. This finding highlights the versatility and potential of general-purpose LLMs to excel in various domains, including finance.

5 Conclusion

This research emphasizes how Large Language Models (LLMs) have revolutionized process automation and data-driven decision-making in the finance industry. Our analysis of state-of-the-art general-purpose language models, such as Open-llama-7B-instruct, llama-2-7B-chat, and GPT-3.5-OpenAI, on carefully curated financial datasets, demonstrated their outstanding performance in diverse financial tasks. Although, due to limited computational power, we focused our research on smaller models to provide valuable insights within our constraints. However, the results have given us some valuable insight into how general-purpose llms can perform almost as good as financial llms and in some cases can outperform them.

References

2023. Open instruct v1 oasst-dolly-hhrlhf dataset.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Alexandra Balahur Siegfried Handschuh Manel Zarrouk Ross McDermott Brian Davis Macedo Maia, André Freitas. 2018. Financial question answering. https://sites.google.com/view/fiqa/home [Accessed: 1 Aug, 2023].

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

OpenAI. 2023. Gpt-4 technical report.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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.