

Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks

Xianzhi Li¹, Samuel Chan¹, Xiaodan Zhu¹, Yulong Pei², Zhiqiang Ma², Xiaomo Liu² and Sameena Shah²

¹Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute
Queen's University

²J.P. Morgan AI Research

{li.xianzhi, 19syc2, xiaodan.zhu}@queensu.ca

{yulong.pei, zhiqiang.ma, xiaomo.liu, sameena.shah}@jpmchase.com

Abstract

The most recent large language models (LLMs) such as ChatGPT and GPT-4 have shown exceptional capabilities of generalist models, achieving state-of-the-art performance on a wide range of NLP tasks with little or no adaptation. How effective are such models in the financial domain? Understanding this basic question would have a significant impact on many downstream financial analytical tasks. In this paper, we conduct an empirical study and provide experimental evidences of their performance on a wide variety of financial text analytical problems, using eight benchmark datasets from five categories of tasks. We report both the strengths and limitations of the current models by comparing them to the state-of-the-art fine-tuned approaches and the recently released domain-specific pretrained models. We hope our study can help understand the capability of the existing models in the financial domain and facilitate further improvements.

In general, in the financial domain, LLMs is playing an increasingly crucial role in tasks such as investment sentiment analysis, financial named entity recognition, and question-answering systems for assisting financial analysts.

In this paper, we perform an empirical study and provide experimental evidence for the effectiveness of the most recent LLMs on a variety of financial text analytical problems, involving eight benchmark datasets from five typical tasks. These datasets are from a range of financial topics and sub-domains such as stock market analysis, financial news, and investment strategies. We report both the strengths and limitations of ChatGPT and GPT-4 by comparing them with the state-of-the-art domain-specific fine-tuned models in finance, e.g., FinBert (Araci, 2019) and FinQANet (Chen et al., 2022a), as well as the recently pretrained model such as BloombergGPT (Wu et al., 2023). Our main contributions are summarized as follows:

1 Introduction

The advancement of LLMs is bringing profound impacts on the financial industry. Through training with reinforcement learning from human feedback (RLHF) (Christiano et al., 2023) and masked language model objectives, the most recent models such as ChatGPT¹ and GPT-4² have demonstrated exceptional capabilities in a wide range of natural language processing (NLP) tasks (Bang et al., 2023a; Liu et al., 2023; Omar et al., 2023; Khoury et al., 2023).

These LLMs are trained on datasets that encompass a broad range of genres and topics. While their performance in generic NLP tasks is impressive, their applicability and effectiveness in specific domains like finance yet need a better understanding and can influence a wide range of appli-

- This study is among the first to explore the most recent advancement of *generically* trained large language models on financial text analytic tasks and it provides a comprehensive comparison.
- We demonstrate that ChatGPT and GPT-4 can outperform the most recently released domain-specifically pretrained model as well as fine-tuned models on many tasks. We provide detailed analysis and recommendations.
- We observe that the advancement made in generalist models continues to carry over to the financial domain; e.g., GPT-4 is significantly better than ChatGPT on nearly all the financial benchmarks used.
- Limitations of the existing LLMs are analyzed and discussed with these benchmark datasets.

¹<https://platform.openai.com/docs/models/gpt-3-5>

²<https://platform.openai.com/docs/models/gpt-4>

Category	Sentiment Analysis	Classification	NER	RE	QA
Complexity	Easy	Easy	Hard	Hard	Hard
Knowledge	Low	Low	High	High	High
Dataset	FPB/FiQA/TweetFinSent	Headline	NER	REFinD	FinQA/ConvFinQA
Eval. Metrics	Weighted F1	Weighted F1	Macro F1	Macro F1	Accuracy
#Test samples	970/223/996	2,114	98	4300	1,147/421

Table 1: Statistics of the five tasks and eight datasets used in this study.

2 Related Works

ChatGPT and Related Models. ChatGPT, GPT-3.5 (text-davinci-003), and GPT-4 are generically trained LLMs and have shown impressive performance on a wide range of tasks. Recent studies have shown that they outperform fine-tuned models on some tasks. But, they still fail in some other cases. Bang et al. (2023b) evaluated ChatGPT on multitasking, multilingual and multimodal tasks, highlighting addressing the failures to improve the overall performance. Qin et al. (2023) studied ChatGPT’s zero-shot capabilities on a diverse range of NLP tasks. While these models present unprecedented quality and retain accumulated knowledge with excellent generalization ability, by respecting the objective of being general problem solvers, how effective they are for financial text analytical tasks is an intriguing open question that needs a better understanding.

Domain-specific Models Currently, there have been only a handful of LLMs specifically trained within the finance domain. BloombergGPT (Wu et al., 2023), a language model with 50 billion parameters, is trained using a mixed approach to cater to the financial industry’s diverse tasks. The model is evaluated on standard LLM benchmarks, open financial benchmarks, and Bloomberg-internal benchmarks. The mixed training approach results in a model that significantly outperforms existing models in financial tasks and performs on par or even better in some general NLP benchmarks. Other researchers also attempted to adapt existing language models to tackle domain-specific tasks. For example, Lewkowycz et al. (2022) adapted T5 to the financial domain. Note that in addition to fine-tuning, A study has also been conducted to use parameter-efficient tuning for financial tasks such as intent detection (Li et al., 2022). The details of the related work can be found in Appendix A.

3 Experiment Setup

Tasks and Datasets. Our research utilizes a wide range of financial NLP tasks and challenges (Pei et al., 2022; Kaur et al., 2023; Shah et al., 2022), enabling us to establish a testbed with different types of NLP problems ranging from basic sentiment analysis and text classification to information extraction and question answering (see Table 1 and more details in Appendix B).

The span of the tasks enables us to make observations along modeling complexity and different levels of financial knowledge required to perform the tasks. Regarding the modeling complexity of tasks, sentiment analysis and text classification are often regarded to be more straightforward, compared to information extraction (IE) tasks such as named entity recognition (NER) and relation extraction (RE). The latter often requires more understanding of syntax and semantics in the input contexts as well as the interactions of labels in the output space as the *structured prediction* problems. Compared to sentiment analysis and text classification, question answering (QA) is often thought of as being harder as it often requires a model to understand the embedded internal logic and numerical operation/reasoning. Regarding financial knowledge, the existing classification and sentiment analysis datasets are sourced from daily news and social media. On the other hand, IE and QA data are often from professional documents like financial filings and reports, which usually require more domain knowledge to comprehend.

Models. We test the representative state-of-the-art LLMs, ChatGPT and GPT-4 models. Specifically, we use gpt-3.5-turbo and GPT-4 (8k) for most of the experiments, except FinQA few-shot experiments, where the input tokens are extra long so we adopt gpt-3.5-turbo-16k.³ Both these LLMs are evaluated using zero-shot and few-shot learning as well as CoT learning for

³All the models are current versions as of July 7th, 2023.

QA reasoning tasks. Furthermore, we compare them with previous LLMs and the domain specific BloombergGPT (Wu et al., 2023). The state-of-the-art fine-tuned models on each dataset are employed to test the idea of training smaller models on individual tasks in comparison with prompting LLMs on all tasks without additional fine-tuning.

Evaluation Metrics. We use *accuracy*, *macro-F1 score*, and *weighted F1 score* (Wu et al., 2023) as the evaluation metrics. For the NER task, we calculate the *entity-level F1 score*. Table 1 shows the details of the experiment setup.

4 Results and Analysis

4.1 Sentiment Analysis

Sentiment analysis is one of the most commonly used NLP techniques in the financial sector and can be used to predict investment behaviors and trends in equity markets from news and social media data (Mishev et al., 2020). We use three financial sentiment datasets with different focuses.

Financial PhraseBank. PhraseBank is a typical three scale (positive, negative and neutral) sentiment classification task curated from financial news by 5-8 annotators (Malo et al., 2013). We use both the *50% annotation agreement* and the *100% agreement* datasets. Same as in (Wu et al., 2023), 20% sentences are used for testing. In Table 2, the first group of models (4 models) are OpenAI LLMs, followed by BloombergGPT, three previous LLMs (referred to as Prior LLMs), and the state-of-the-art fine-tuned models on this dataset (FinBert). Due to the space limit of Table 2, we put the name of these four groups in the next table (Table 3) for clarity. In Table 2, we can see that the performance of Prior LLMs greatly falls behind ChatGPT and GPT-4. With the enhancement of few-shot learning, GPT-4 is comparable to fine-tuned FinBert (Araci, 2019).

FiQA Sentiment Analysis. This dataset extends the task complexity to detect aspect-based sentiments from news and microblog in the financial domain (Maia et al., 2018). We follow BloombergGPT’s setting (Wu et al., 2023), where we cast this regression task into a classification task. 20% of labeled training data are held as test cases. The results in Table 3 present similar performance trends as in the previous dataset: ChatGPT and GPT-4 out-

Data	50% Agreement		100% Agreement	
Model	Accuracy	F1 score	Accuracy	F1 score
ChatGPT ₍₀₎	0.78	0.78	0.90	0.90
ChatGPT ₍₅₎	0.79	0.79	0.90	0.90
GPT-4 ₍₀₎	<u>0.83</u>	<u>0.83</u>	<u>0.96</u>	<u>0.96</u>
GPT-4 ₍₅₎	0.86	0.86	0.97	0.97
BloombergGPT ₍₅₎	/	0.51	/	/
GPT-NeoX ₍₅₎	/	0.45	/	/
OPT66B ₍₅₎	/	0.49	/	/
BLOOM176B ₍₅₎	/	0.50	/	/
FinBert	0.86	0.84	0.97	0.95

Table 2: Results on the Phrasebank dataset. The subscript (*n*) after an LLM name represents the number of shots. The best results are marked in bold and the second-best with underscored. The results of other LLMs like BloombergGPT are from the corresponding papers. ‘/’ indicates the metrics were not included in the original study. The notation convention used here applies to all the following experiments. Different few-shot settings are tested and discussed in Appendix C.

perform Prior LLMs. With a few-shot examples GPT-4 is better than all other models here. BloombergGPT has relatively close performance to zero-shot ChatGPT and is inferior to GPT-4. The fine-tuned RoBERTa-large model on this dataset is better than ChatGPT, but is slightly less effective than GPT-4. The latter achieves 88% on F1, which is less than that in Financial PhraseBank. We due this to the fact that FiQA requires modeling more details and needs more domain knowledge to understand the sentiment with the aspect finance tree in the data.

Model	Category	Weighted F1
ChatGPT ₍₀₎	OpenAI LLMs	75.90
ChatGPT ₍₅₎		78.33
GPT-4 ₍₀₎		<u>87.15</u>
GPT-4 ₍₅₎		88.11
BloombergGPT ₍₅₎	Domain LLM	75.07
GPT-NeoX ₍₅₎	Prior LLMs	50.59
OPT66B ₍₅₎		51.60
BLOOM176B ₍₅₎		53.12
RoBERTa-large	Fine-tune	87.09

Table 3: Results on the FiQA dataset.

TweetFinSent. Pei et al. (2022) created this dataset based on Twitter to capture retail investors’ mood to a specific stock ticker. Since tweets are informal texts which typically are not used to train LLMs, this could be a challenging task for LLMs to perform well. Furthermore, a tweet can sometimes contain several tickers (>5 is not unusual). The aspect modeling on this data is more complex. The evaluation results on 996 test instances

are shown in Table 4. GPT-4 with a few-shot examples achieves ~72% accuracy and F1, which is lower than the values in the previous two tasks. The fine-tuned RoBERTa-Twitter (Pei et al., 2022) has similar performance. We also conduct an ablation study by removing emojis. Both ChatGPT and GPT-4 show 2-3 points performance drop, indicating emojis in social media do convey meaningful sentiment signals. We do not have results of Prior LLMs as this dataset is not evaluated in the corresponding previous studies.

Model	Accuracy	Weighted F1
ChatGPT ₍₀₎	68.48	68.60
ChatGPT ₍₅₎	69.93	70.05
GPT-4 ₍₀₎	69.08	69.17
GPT-4 ₍₅₎	<u>71.95</u>	72.12
ChatGPT _{((0_no_emoji))}	64.40	64.43
ChatGPT _{((5_no_emoji))}	67.37	67.61
GPT-4 _{((0_no_emoji))}	67.26	67.45
GPT-4 _{((5_no_emoji))}	70.58	70.44
RoBERTa-Twitter	72.30	<u>71.96</u>

Table 4: Results on the TweetFinSent dataset.

4.2 Headline Classification

While sentiment analysis has been regarded as one of the most basic tasks and is mainly pertaining to some dimensions of *semantic orientation* (Osgood et al., 1957), the semantics involved in financial text classification tasks can be more complicated. Classification, particularly multi-class text classification, is often applied to a wide range of financial text such as news, SEC 10-Ks, and market research reports to accelerate business operations.

Same as in (Wu et al., 2023), we use the news headlines classification dataset (Sinha and Khandait, 2020) from the FLUE benchmark (Shah et al., 2022). This classification task targets to classify commodity news headlines to one of the six categories like “Price Up” and “Price Down”. We follow the setting in BloombergGPT, converting the multi-class classification to six individual binary classification problems (refer to Figure 7 as an example).

The model performance is listed in Table 5. Again GPT-4 outperforms ChatGPT and Prior LLMs as well as BloombergGPT. The fine-tuned BERT can achieve 95% on F1, 9% higher than 5-shot GPT-4. This task is considered to be challenging due to its multi-class and the need of domain knowledge of the commodity market.

Model	Weighted F1
ChatGPT ₍₀₎	71.78
ChatGPT ₍₅₎	74.84
GPT-4 ₍₀₎	84.17
GPT-4 ₍₅₎	86.00
BloombergGPT ₍₅₎	82.20
GPT-NeoX ₍₅₎	73.22
OPT66B ₍₅₎	79.41
BLOOM176B ₍₅₎	76.51
BERT	95.36

Table 5: Results on the headline classification task.

4.3 Named Entity Recognition

NER helps structure textual documents by extracting entities. It is a powerful technique to automate document processing and knowledge extraction from documents (Yang, 2021). In our evaluation, we use the NER FIN3 datasets, created by Salinas Alvarado et al. (2015) using financial agreements from SEC and containing four NE types: PER, LOC, ORG and MISC. Following the setting used in BloombergGPT, we remove all entities with the MISC label due to its ambiguity.

In Table 6, we can see that both GPT-4 and ChatGPT perform poorly under the zero-shot setup. Following BloombergGPT’s setting, the few-shot learning uses 20 shots on this dataset. We can see that GPT-4 is less effective than BloombergGPT, and is comparable or worse than Prior LLMs on this task. Since NER is a classic structured prediction problem, CRF model is also compared. When CRF is trained with FIN5, which is similar to the test data (FIN3), it performs better than all the other models (see the last row of the table). Note that CRF is very sensitive to domain shifting—when it is trained on the out-of-domain CoNLL data, it performs poorly on the FIN3 data (refer to the second to the last row of Table 6), inferior to the zero-shot LLMs. In general, in this structured prediction task, LLMs’ performance is not ideal and future improvement is imperative, particularly for the generalist models.

4.4 Relation Extraction

Relation extraction aims to detect linkage between extracted entities. It is a foundational component for knowledge graph construction, question answering and semantic search applications for the financial industry. In this study, we use a financial relation extraction dataset — REFinD, which was created from 10-K/Q filings with 22 relation types Kaur et al. (2023). In order for LLMs to predict the relationship between two entities, we provide

Model	Entity F1
ChatGPT (0)	29.21
ChatGPT (20)	51.52
GPT-4 (0)	36.08
GPT-4 (20)	56.71
BloombergGPT (20)	60.82
GPT-NeoX (20)	60.98
OPT66B (20)	57.49
BLOOM176B (20)	55.56
CRF (CoNLL)	17.20
CRF (FIN5)	82.70

Table 6: Results of few-shot performance on the NER dataset. CRF (CoNLL) refers to CRF model that is trained on general CoNLL data, CRF (FIN5) refers to CRF model that is trained on FIN5 data. Again, we choose the same shot as BloombergGPT for fair comparison. More detailed experiments using 5 to 20 shots can be found in Appendix C.

the original sentence, entity words, and their entity types in the prompts and ask the models to predict a relation type. Same as in Luke-base (Yamada et al., 2020), we use Macro F1. Table 7 shows that the fine-tuned Luke-base outperforms both ChatGPT and GPT-4 by a notable margin. On the other hand, GPT-4 demonstrates considerably better performance compared to ChatGPT. The outcomes from this IE task illustrated the strength of fine-tuning on complex tasks that need a better understanding of the structure of sentences.

Model	Macro F1
ChatGPT (0)	20.97
ChatGPT (10)	29.53
GPT-4 (0)	42.29
GPT-4 (10)	46.87
Luke-base (fine-tune)	56.30

Table 7: Results on the REFinD dataset.

4.5 Question Answering

The application of QA to finance presents a possible path to automate financial analysis, which at present is almost 100% conducted by trained financial professionals. It is conventionally thought of as being challenging since it often requires a model to understand not only domain knowledge but also the embedded internal logic and numerical operation/reasoning. We adopt two QA datasets: FinQA (Chen et al., 2022a) and ConvFinQA (Chen et al., 2022b). The former dataset focuses on a single question and answer pair. The latter decomposes the task into a multi-round structure: a chain of reasoning through conversation. Both of them concentrate on numerical rea-

soning in financial analysis, e.g. calculating profit growth ratio over years from a financial table. The experiment setting and prompt design details are in Appendix B and C. Since the labels of the ConvFinQA test set are not publicly available, we utilize its dev dataset (421 samples) instead to evaluate the models, while for FinQA use the testing dataset (1,147 samples).

Model	FinQA	ConvFinQA
ChatGPT (0)	48.56	59.86
ChatGPT (3)	51.22	/
ChatGPT (CoT)	63.87	/
GPT-4 (0)	68.79	76.48
GPT-4 (3)	69.68	/
GPT-4 (CoT)	78.03	/
BloombergGPT (0)	/	43.41
GPT-NeoX (0)	/	30.06
OPT66B (0)	/	27.88
BLOOM176B (0)	/	36.31
FinQANet (fine-tune)	68.90	61.24
Human Expert	91.16	89.44
General Crowd	50.68	46.90

Table 8: Model performance (accuracy) on the question answering tasks. FinQANet here refers to the best-performing FinQANet version based on RoBERTa-Large (Chen et al., 2022a). Few-shot and CoT learning cannot be executed on ConvFinQA due to the conservation nature of ConvFinQA.

From the performance in Table 8, we can see that GPT-4 substantially outperforms all the other LLMs in both datasets. For FinQA, GPT-4 has highest zero-shot accuracy of 68.79%, while ChatGPT has 48.56%. The performance gap between GPT-4 and ChatGPT persists on ConvFinQA. ChatGPT has a big edge over BloombergGPT (59.86% vs. 43.41%) and also Prior LLMs on ConvFinQA. This result demonstrates that the continuous improvement of reasoning developed through ChatGPT to GPT-4, which is also observed in other studies.

We further explore the impact of few-shot learning and Chain-of-Thought (CoT) prompting on GPT-4 and ChatGPT on the FinQA task. The results provide a compelling narrative of performance increase using these prompting strategies. Both ChatGPT and GPT-4 show a 1-3% accuracy increase using 3 shots. This is consistent with our observations from other tasks. The CoT strategy brings a massive lift, 10% and 15% percentage points, to ChatGPT and GPT-4 respectively. These results underscore the importance of detailed reasoning steps over shallow reasoning in boosting the performance of language mod-

els on complex financial QA tasks. The best GPT-4 result indeed exceeds the fine-tuned FinQANet model with a quite significant margin. It is surprising to us since we previously observe that fine-tuned models have advantages on more complex tasks. We reckon that the scale of parameters and pre-training approaches make ChatGPT and GPT-4 excel in reasoning than other models, particularly the numerical capability of GPT-4, which was demonstrated when the model was released by OpenAI. But their performance (70+% accuracy) still cannot match that of professionals (~90% accuracy). Furthermore, numerical reasoning is just one of many reasoning tasks. More studies are needed for symbolic reasoning and other logic reasoning (Qin et al., 2023) if more datasets in the financial sector are further available. Also, we think the pretraining strategy such as RLHF has not been designed to improve sequence-labeling and structured-prediction skills needed in IE, but can inherently benefit QA.

5 Discussions

Comparison over LLMs. We are able to benchmark the performance of ChatGPT and GPT-4 with four other LLMs on five tasks with eight datasets. ChatGPT and GPT-4 significantly outperforms others in almost all datasets except the NER task. It is interesting to observe that both models perform better on financial NLP tasks than BloombergGPT, which was specifically trained on financial corpora. This might be due to the larger model size of the two models. Finally, GPT-4 constantly shows 10+% boost over ChatGPT in straightforward tasks such as Headlines and FiQA SA. For challenging tasks like RE and QA, GPT-4 can introduce 20-100% performance growth. This indicates that GPT-4 could be the first choice for financial NLP tasks before a more powerful LLM emerges.

Prompt Engineering Strategies. We adopted two commonly used prompting strategies: few-shot and chain-of-thoughts. We constantly observe 1% to 4% performance boost on ChatGPT and GPT-4 from few-shot over zero-shot learning across various datasets. Chain-of-thoughts prompting is very effective in our test and demonstrates 20-30% accuracy improvement over zero-shot and few-shot as well. According our findings, we argue that these two strategies should always be considered first when applying LLMs to finan-

cial NLP tasks.

LLMs vs. Fine-tuning. One attractive benefit of using LLMs in business domains is that they can be applied to a broad range of NLP tasks without conducting much overhead work. It is more economical compared to fine-tuning separate models for every task. Whereas, our experiments show fine-tuned models still demonstrate strong performance in most of the tasks except the QA task. Notably, for tasks like NER and RE, LLMs are less effective than fine-tuned models. In the QA tasks, LLMs illustrated the advantage over fine-tuned model. But the reasoning complexity of the tested QA tasks is still deemed as basic in financial analysis. Although ChatGPT and GPT-4 have proven to be able to perform multi-step reasoning, including numerical reasoning, to some extent, simple mistakes have still been made.

Using LLMs in Financial Services. This study suggests that one can consider adopting the state-of-the-art generalist LLMs to address the relatively simple NLP tasks in financial applications. For more complicated tasks such as structured prediction, the pretraining plus fine-tuning paradigm is still a leading option. Although ChatGPT and GPT-4 excel on QA compared to other models and are better than the general crowd, they are still far from satisfactory from the industry requirement standpoint. Significant research and improvement on LLMs are required before they can act as a trustworthy financial analyst agent.

6 Conclusion

This study is among the first to explore the most recent advancement of *generically* trained LLMs, including ChatGPT and GPT-4, on a wide range of financial text analytics tasks. These models have been shown to outperform models fine-tuned with domain-specific data on some tasks, but still fall short on others, particularly when deeper semantics and structural analysis are needed. While we provide comprehensive studies on eight datasets from five categories of tasks, we view our effort as an initial study, and further investigation of LLMs on financial applications is highly desirable, including the design of more tasks to gain further insights on the limitations of existing models, the integration of LLMs in the loop of human decision making, and the robustness of the models in high-stakes financial tasks.

Acknowledgement

This research was funded in part by the Faculty Research Awards of J.P. Morgan AI Research. The authors are solely responsible for the contents of the paper and the opinions expressed in this publication do not reflect those of the funding agencies.

Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

- Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models](#).
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#).
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023b. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#).
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022a. [Finqa: A dataset of numerical reasoning over financial data](#).
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. [Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering](#).
- Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. [Deep reinforcement learning from human preferences](#).
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. [Training compute-optimal large language models](#).
- Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. [Refind: Relation extraction financial dataset](#). *arXiv preprint arXiv:2305.18322*.
- Raphaël Khoury, Anderson R. Avila, Jacob Brunelle, and Baba Mamadou Camara. 2023. [How secure is code generated by chatgpt?](#)
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#).
- Xianzhi Li, Will Aitken, Xiaodan Zhu, and Stephen W. Thomas. 2022. [Learning better intent representations for financial open intent classification](#).
- Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. [Evaluating the logical reasoning ability of chatgpt and gpt-4](#).
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. [BioGPT: generative pre-trained transformer for biomedical text generation and mining](#). *Briefings in Bioinformatics*, 23(6).
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www'18 open challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. [Good debt or bad debt: Detecting semantic orientations in economic texts](#).
- Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. 2020. [Evaluation of sentiment analysis in finance: from](#)

- lexicons to transformers. *IEEE access*, 8:131662–131682.
- Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. [Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots](#).
- Charles Osgood, George Suci, and Percy Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.
- Yulong Pei, Amarachi Mbakwe, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu, and Sameena Shah. 2022. [TweetFinSent: A dataset of stock sentiments on Twitter](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 37–47, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is chatgpt a general-purpose natural language processing task solver?](#)
- 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.
- Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. [What language model to train if you have one million gpu hours?](#)
- Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. [Large language models encode clinical knowledge](#).
- Ankur Sinha and Tanmay Khandait. 2020. [Impact of news on the commodity market: Dataset and results](#).
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. [Galactica: A large language model for science](#).
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#).
- Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. [LUKE: Deep contextualized entity representations with entity-aware self-attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.
- Sharon Yang. 2021. [Financial use cases for named entity recognition \(ner\)](#).
- Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2020. [Global table extractor \(gte\): A framework for joint table identification and cell structure recognition using visual context](#).

A Details of the Related Work

ChatGPT and Related Models. ChatGPT, GPT-3.5 (text-davinci-003), and GPT-4 are all part of a series of large language models created by OpenAI. GPT-4, as the latest and most advanced version, builds on the achievements of its forerunners. ChatGPT is an earlier version, tailored to offer users engaging and responsive conversational experiences. GPT-3.5 acted as a transitional stage between GPT-3 and GPT-4, improving upon the former and paving the way for the latter.

ChatGPT presents unprecedented quality when interacting with humans conversationally while retaining accumulated knowledge and generalization ability, achieved through large-scale conversational-style dataset pre-training and reward model fine-tuning. This allows ChatGPT to answer follow-up questions, admit mistakes, challenge incorrect premises, and reject inappropriate requests. Secondly, it is trained with a human-aligned objective function using Reinforcement Learning from Human Feedback (RLHF), which results in its output being more closely aligned with human preferences.

Recent studies have shown that ChatGPT outperforms multiple state-of-the-art zero-shot LLMs on various tasks and even surpasses fine-tuned models on some tasks. However, like many LLMs, ChatGPT still fails in many cases, such as generating overly long summaries or producing incorrect translations. A recent study (Bang et al., 2023b) evaluated ChatGPT’s performance on multitasking, multilingual and multimodal tasks, highlighting the importance of addressing these failure cases for improving the overall performance of the model.

Qin et al. (2023) studied ChatGPT’s zero-shot capabilities on a diverse range of NLP tasks, providing a preliminary profile of the model. Their findings suggest that while ChatGPT shows certain generalization capabilities, it often underperforms compared to fine-tuned models on specific tasks. Compared to GPT-3.5, ChatGPT outperforms it on natural language inference, question answering, and dialogue tasks, while its summarization ability is inferior. Both ChatGPT and GPT-3.5 face challenges on sequence tagging tasks.

Domain-specific Models. Currently, there has been only a handful of financial-domain-specific LLMs available, which are often trained exclusively on domain-specific data. These LLMs have shown promising results in their respective domain tasks. For instance, Luo et al. (2022) developed an LLM for the legal domain, which was trained exclusively on legal texts, and (Taylor et al., 2022) trained a healthcare LLM.

Most recently, BloombergGPT (Wu et al., 2023), a language model with 50 billion parameters, is trained using a mixed approach to cater to the financial industry’s diverse tasks while maintaining competitive performance on general-purpose LLM benchmarks. A training corpus with over 700 billion tokens is created by leveraging Bloomberg’s proprietary financial data archives and combining them with public datasets. The model, designed based on the guidelines from (Hoffmann et al., 2022) and (Scao et al., 2022), is validated on standard LLM benchmarks, open financial benchmarks, and Bloomberg-internal benchmarks. The mixed training approach results in a model that significantly outperforms existing models in financial tasks and performs on par or even better in some general NLP benchmarks.

It is worth mentioning that other researchers opt to adapt large general-purpose language models to tackle domain-specific tasks. For example, Singhal et al. (2022) applied GPT-3 in the legal domain, and Lewkowycz et al. (2022) adapted T5 to the financial domain. Despite being trained on a general purpose corpus, these models have also demonstrated excellent performance when applied to domain-specific tasks. Note that in addition to fine-tuning, research has also been conducted to use parameter efficient tuning for financial tasks such as intent detection on Banking77 dataset (Li et al., 2022).

B Dataset Details

Financial PhraseBank. This is a dataset introduced by Malo et al. (2013), which is a sentiment classification dataset derived from financial news sentences. It is designed to assess the impact of news on investors, with positive, negative, or neutral sentiment labels being assigned to each news sentence from an investor’s perspective. Containing 4,845 English sentences, the dataset is sourced from financial news articles found in the Lexis-Nexis database. These sentences were annotated

by individuals with expertise in finance and business, who were tasked with assigning labels based on their perception of the sentence’s potential influence on the mentioned company’s stock price.

FiQA Sentiment Analysis. This second sentiment analysis task is part of the FiQA challenge (Maia et al., 2018) focusing on the prediction of sentiment specifically related to aspects within English financial news and microblog headlines. This was initially released as part of the 2018 competition that centered on financial question answering and opinion mining. The primary dataset was marked on a continuous scale, but we follow BloombergGPT’s setting and transform it into a classification system with three categories: negative, neutral, and positive. We’ve created our own test split incorporating both microblogs and news. We use a 0-shot learning and our results are calculated through the weighted F1 score. We fine-tuned a RoBERTa-large model on this task for comparison with OpenAI and other LLMs.

TweetFinSent. This third sentiment analysis task is introduced by (Pei et al., 2022). The unique attribute of the TweetFinSent dataset is that it annotates tweets not merely on emotional sentiment, but also on the anticipated or realized gains or losses from a specific stock. Previous studies have revealed the TweetFinSent dataset as a challenging problem with significant room for improvement in the realm of stock sentiment analysis.

Headlines. This binary classification task, created by Sinha and Khandait (2020), involves determining whether a news headline contains gold price related information. This dataset contains 11,412 English news headlines which span from 2000 to 2019. The headlines were collected from various sources, including Reuters, The Hindu, The Economic Times, Bloomberg, as well as aggregator sites. We note that the dataset we have access to consists of six tags: “price up”, “price down”, “price stable”, “past price”, “future price”, and “asset comparison”, while the test reported in BloombergGPT used a version of nine categories. We contacted the original dataset authors, they claimed that they had performed some additional filtering and provided this six-label dataset.

We also conducted an experiment where we prompted ChatGPT and GPT-4 to generate answers simultaneously in response to six distinct questions. Our preliminary findings suggest that

these models handle single-question prompts more effectively than those involving multi-tag binary classification. We noticed a significant drop in performance related to three tags: ‘past information’, ‘future information’, and ‘asset comparison’. This suggests that the models struggle to provide separate and accurate responses to a series of questions presented at once.

NER. This named entity recognition task focuses on financial data collected for credit risk assessment from financial agreements filed with the U.S. Securities and Exchange Commission (SEC). The dataset, created by Salinas Alvarado et al. (2015), consists of eight manually annotated documents with approximately 55,000 words. These documents are divided into two subsets: “FIN5” for training and “FIN3” for testing. The annotated entity types follow the standard CoNLL format (Tjong Kim Sang and De Meulder, 2003) and include PERSON (PER), LOCATION (LOC), ORGANIZATION (ORG), and MISCELLANEOUS (MISC).

REFinD. This relation extraction dataset is created by Kaur et al. (2023). REFinD is currently the most extensive of its kind, consisting of approximately 29K instances and 22 relations amongst 8 types of entity pairs. This specialized financial relation extraction dataset is constructed from raw text sourced from various 10-X reports (including but not limited to 10-K and 10-Q) of publicly traded companies. These reports were obtained from the website of the U.S. Securities and Exchange Commission (SEC).

ConvFinQA. This is an extension of the FinQA dataset, named as ConvFinQA (Chen et al., 2022b), which is designed to address numerical reasoning chains in a format of conversational question-answering tasks. ConvFinQA expands the original FinQA dataset to include 3,892 conversations with 14,115 questions derived from earnings reports of S&P 500 companies. This task not only demands numerical reasoning and understanding of structured data and financial concepts, but also emphasizes the ability to relate follow-up questions to previous conversation context.

For the ConvFinQA dataset, we employ a turn-based approach, where we collect the answer generated by the models after each turn, append it to the previous question, and use them along with the next question as the prompt input for the next

round. As shown in Figure 11, we collect Answer 1 (A1) after Question 1 (Q1) and then prefix A1 together with Question 2 (Q2) to proceed to the next round, and so on, until we reach the end of the conversation chain.

We notice some fundamental issues of ChatGPT from the tests on the ConvFinQA dataset. Firstly, it makes some basic mistakes, such as miscalculating “\$753 million + \$785 million + \$1,134 million” to be \$3,672 million instead of \$2,672 million. Even though all the intermediate results are correct, the final summation step produces an incorrect final answer. Note that such mistakes can be critical in the financial domain, particularly in high-stakes setups. We also found that ChatGPT struggles with understanding contextual information and coreference in conversations. For example, in ConvFinQA, questions often use the word “that” to refer to an entity mentioned in the previous question, but ChatGPT sometimes responds with a request for clarification, indicating its limitations in handling coreference. In contrast, GPT-4 shows significant improvement and faces this issue much less.

FinQA. Chen et al. (2022a) propose an expert-annotated dataset consisting of 8,281 financial question-answer pairs, along with their corresponding numerical reasoning processes. Created by eleven finance professionals, FinQA is based on earnings reports from S&P 500 companies (Zheng et al., 2020). The questions necessitate extracting information from both tables and unstructured texts to provide accurate answers. The reasoning processes involved in answering these questions comprise common financial analysis operations, including mathematical operations, comparison, and table aggregation operations. FinQA is the first dataset of its kind designed to address complex question-answering tasks based on real-world financial documents.

When composing prompts, we use text and tables as context input, following the pattern ‘pre_text’ + ‘table’ + ‘post_text’, where the ‘pre’ and ‘post’ texts provide the necessary context for the table, and the table itself contains the structured data that the model is expected to reason on and generate responses from. We also convert tables into a markdown format. For the FinQA dataset, we simply ask the question right after the context. Figure 12 demonstrates the complete prompt format.

We use the function call feature to assist CoT prompting. This Question_Answering function required both models to generate two arguments: a) “thinking process” which contains each step of the reasoning process and evidence of how they locate information in the original documents and perform calculations, and b) “answer”, which is the final numerical response.

We also conduct experiments with each of these models being subjected to different steps complexity, classified as 1-step programs, 2-step programs, and programs that involve more than 2 steps of calculation. For problems involving less than 2 steps, The models’ performance follows the same trend as overall results, where GPT-4 maintains the lead, outperforming FinQANet and ChatGPT. However, the conclusion changes with the increase in problem complexity. When faced with problems requiring more than 2 steps, ChatGPT outperformed FinQANet by a significant margin, scoring accuracy of 32.14% as opposed to FinQANet’s 22.78%. It is intriguing to note that despite struggling with less complex tasks, ChatGPT managed to outpace FinQANet when problem complexity escalated.

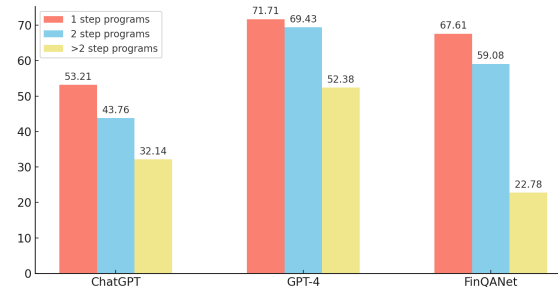


Figure 1: FinQA program steps analysis

C Few Shots Experiments

We conducted few-shot experiments on 6 widely used datasets out of 8. We argue that ConvFinQA task itself is designed with a multi-step QA setup so we didn’t conduct few-shot experiments on this dataset. For each shot number, we ran the experiment 10 times and generated box plots, which can be found in Figure 2 to 6 below. The general trend shows that as we increase the number of shots, the performance of ChatGPT improves by approximately 1% to 4% across various datasets, in comparison to zero-shot. For simpler tasks, such as Sentiment Analysis (illustrated in Figure 3), ChatGPT only requires 6 shots to perform ef-

fectively. However, as we continue to increase the number of shots, the rate of improvement tapers off. For NER tasks, 5 shots do not impart sufficient domain information to ChatGPT, thus necessitating more than 15 shots to adequately guide the model. Additionally, we observed that performance can still fluctuate even with the same number of shots. The dispersion illustrated in the box plots indicates a certain level of volatility, suggesting that ChatGPT is quite sensitive to the shots used. This underlines the importance of careful selection and design of shots and prompts.

We also listed the zero-shot prompt we used for each dataset, please find them in Figure 7 to 12. We use slightly different prompts for few-shot and CoT experiments since the shots and function call already provide guidance on how to structure the output.

Category	Tag Question
price up	Does the news headline talk about price going up?
price stable	Does the news headline talk about price staying constant?
price down	Does the news headline talk about price going down?
past price	Does the news headline talk about price in the past?
future price	Does the news headline talk about price in the future?
asset comparison	Does the news headline compare gold with any other asset?

Table 9: Each tag and its corresponding converted question

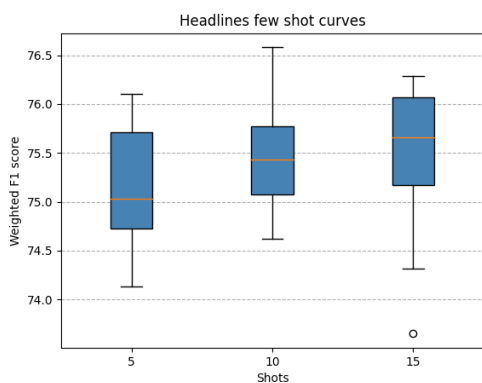


Figure 2: Headlines few shot results curve

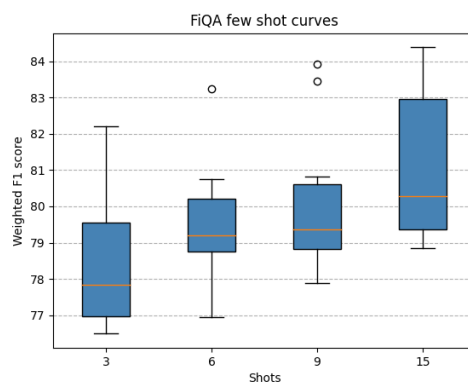


Figure 3: FiQA few shot results curve

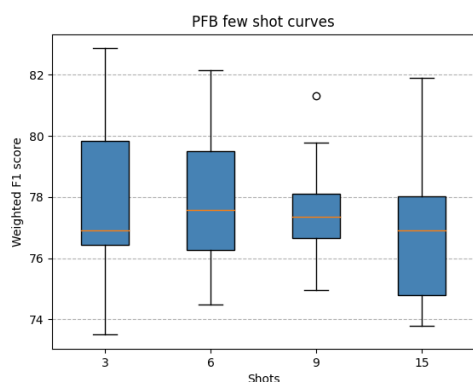


Figure 4: PFB few shot results curve

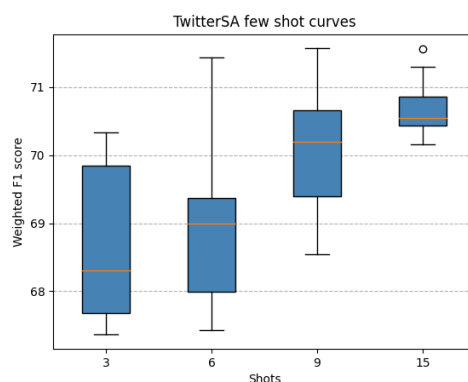


Figure 5: TweetFinSent few shot results curve

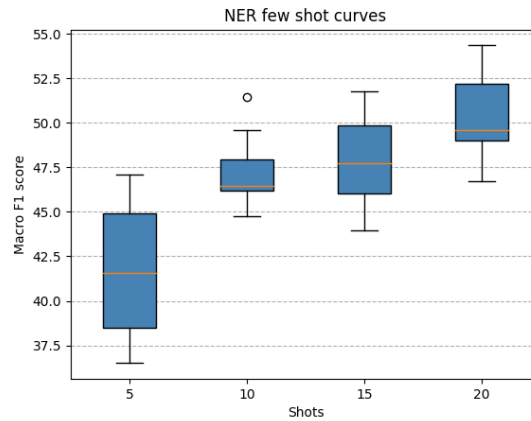


Figure 6: NER few shot results curve

Headlines

Given a news headline about gold, using Yes or No to answer the following question:

Text: gold prices narrowly higher; metals shares rise
 Question: Does the news headline talk about price going up?
 Answer: Yes

Figure 7: prompt for Headlines dataset

Financial PhraseBank

Label the sentiment of the given text. The answer should be exact 'positive', 'neutral' or 'negative'. The answer is

Text: A Helsinki : ELLiV today reported EPS of EUR1 .13 for 2009 , an increase over EPS of EUR1 .12 in 2008
 Answer: Positive

Figure 8: prompt for FPB dataset, same for other sentiment analysis tasks

Fin_NER

Please identify Person, Organization, Location Entity from the given text.

Text: Subordinated Loan Agreement - Silicium de Provence SAS and Evergreen Solar Inc . 7 - December 2007 [HERBERT SMITH LOGO]
 2007 SILICIUM DE PROVENCE SAS and EVERGREEN SOLAR , INC
 Answer:

Figure 9: prompt for NER dataset

Relation Extraction

Please identify the relationship between two entities in a sentence, you need to choose one from <22 relationships types>

Input: "Sentence" + "Entity1"(E1 type) + "Entity2"(E2 type)

Answer:

Figure 10: prompt for Relation Extraction dataset

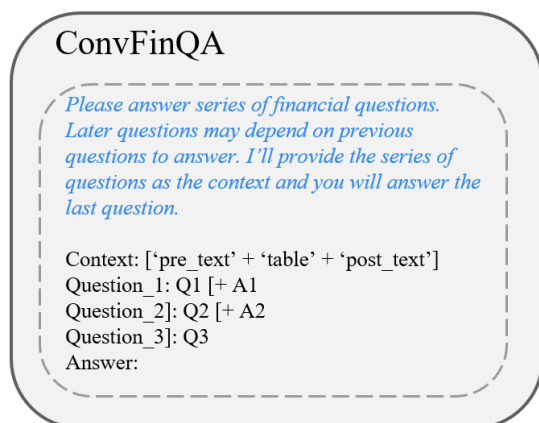


Figure 11: prompt for ConvFinQA dataset

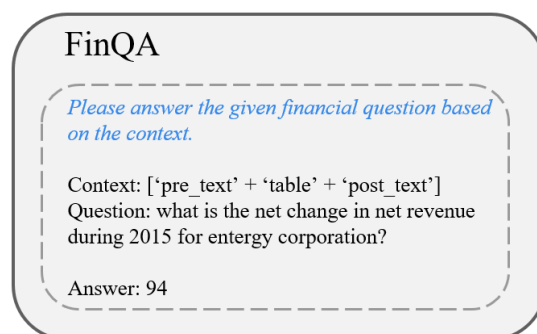


Figure 12: prompt for FinQA dataset