Beyond Classification: Financial Reasoning in State-of-the-Art Language Models

Guijin Son[1,2] , **Hanearl Jung**[2] , **Moonjeong Hahm**[3] , **Keonju Na**[4] and **Sol Jin**[5]
[1]Yonsei University [2]OneLineAI [3]Chung-Ang University
[4]Seoul National University of Science and Technology [5]Seoul National University
spthsrbwls123@yonsei.ac.kr, earl@onelineai.com, daily6298@cau.ac.kr, keonju2@seoultech.ac.kr,
jinsol9770@snu.ac.kr

Abstract

Large Language Models (LLMs), consisting of 100 billion or more parameters, have demonstrated remarkable ability in complex multi-step reasoning tasks. However, the application of such generic advancements has been limited to a few fields, such as clinical or legal, with the field of financial reasoning remaining largely unexplored. To the best of our knowledge, the ability of LLMs to solve financial reasoning problems has never been dealt with, and whether it can be performed at any scale remains unknown. To address this knowledge gap, this research presents a comprehensive investigation into the potential application of LLMs in the financial domain. The investigation includes a detailed exploration of a range of subjects, including task formulation, synthetic data generation, prompting methods, and evaluation capability. Furthermore, the study benchmarks various GPT variants with parameter scales ranging from 2.8B to 13B, with and without instruction tuning, on diverse dataset sizes. By analyzing the results, we reveal that the ability to generate coherent financial reasoning first emerges at 6B parameters, and continues to improve with better instruction-tuning or larger datasets. Additionally, the study provides a publicly accessible dataset named sFIOG (Synthetic-Financial Investment Opinion Generation), consisting of 11,802 synthetic investment thesis samples, to support further research in the field of financial reasoning. Overall, this research seeks to contribute to the understanding of the efficacy of language models in the field of finance, with a particular emphasis on their ability to engage in sophisticated reasoning and analysis within the context of investment decision-making.

1 Introduction

Large Language Models(100+ billion parameters) have undergone remarkable advancements in recent years, enabling them with the ability to generate coherent and meaningful text [Wei *et al.*, 2022a]. These LLMs have demonstrated notable abilities in performing complex multi-step reasoning, either by thinking "step by step" [Kojima *et al.*,] or leveraging Chain-of-Thought(CoT) prompts [Wei *et al.*, 2022b]. Various fields have attempted to harness such reasoning ability, and among them, the field of clinical research has made notable progress by developing domain-specific LLMs like Med-Palm [Singhal *et al.*, 2022], retrained on massive amounts of domain-specific texts and tasks, which achieves performance comparable to that of human clinicians. In situations where data is insufficient to train dedicated language models, researchers have directed their efforts towards developing advanced prompt engineering techniques, such as Legal Prompt Engineering (LPE) [Trautmann *et al.*, 2022], or generation of synthetic data via LLMs and training of smaller language models on such samples [Yunxiang *et al.*, 2023]. However, there is a lack of comprehensive investigation for either of the methods in the financial domain, leaving the field of financial reasoning largely unexplored.

The research of natural language processing in the financial domain has predominantly been confined to token or sequence classification tasks [Araci, 2019; Shah *et al.*, 2022]. This is likely due to the lack of datasets or tasks suitable for training generative language models. Even dedicated financial language models like BloombergGPT, tend to prioritize tasks such as sentiment analysis, binary classification, and named entity recognition, with limited attention given to numerical reasoning tasks [Wu *et al.*, 2023].

Our research aims to comprehensively investigate the **financial reasoning** capabilities of language models, specifically their ability to generate logically coherent and persuasive investment opinions. The investigation involves both prompt engineering and specialized training of smaller language models [Fu *et al.*, 2023], seeking to advance our understanding on the ability of language models to engage in sophisticated reasoning and analysis within the context of investment decision-making. Accordingly, our research introduces an original financial reasoning task called "Financial Investment Opinion Generation (FIOG)", which involves the generation of investment opinions by language models with either parametric or injected knowledge. We then benchmark various GPT variants, ranging in size from 2.7B to 13B, with and without instruction-tuning [Ouyang *et al.*, 2022], on the dataset. Additionally, we propose a novel prompting method called In-Context Question Answering for controlled generation of context. Finally, we investigate the alignment between

LLM-based evaluators, such as G-Eval [Liu *et al.*, 2023], and human evaluators for financial texts, in order to gain insights into the efficacy of such evaluators in the financial domain.

To support further research on financial reasoning, we provide a publicly accessible dataset named sFIOG (Synthetic-Financial Investment Opinion Generation), which includes 11,802 synthetic investment opinion samples. This dataset is intended to enable benchmarking and experimentation in the field of financial language modeling and investment opinion generation.

2 Related Work

2.1 Reasoning with Language Models

Language Models (LMs) trained using conventional pre-training objectives have demonstrated the ability to acquire complex reasoning capabilities once they reach a certain scale [Wei *et al.*, 2022a]. However, recent research has shown that the parameter requirements for complex reasoning abilities of LMs can be significantly alleviated through a process called instruction tuning [Ouyang *et al.*, 2022]. Further research has suggested that narrowing down the model's focus to specialize in a specific field can result in additional alleviation of parameter requirements. This can be achieved by including task-specific Chain-of-Thought (CoT) data in the instruction-tuning process, allowing the model to acquire specialized reasoning capabilities [Fu *et al.*, 2023]. Some researchers have adopted this approach, leveraging domain-specific CoT data, which is often generated by the LLMs themselves, to enable domain-specific reasoning abilities [Yunxiang *et al.*, 2023]. However, the effectiveness of this approach across different domains and the potential variability in parameter and data requirements for specific domains remain relatively unexplored. Accordingly, it is plausible that domains characterized by complex nomenclature and reasoning steps, which significantly deviate from general, widely applicable patterns, may necessitate higher parameter and data requirements.

2.2 Financial Natural Language Processing

The financial domain has been quick to adopt advancements in generic natural language processing research. Notably, BloombergGPT, a language model with 50 billion parameters specifically dedicated for finance, stands out as a significant development in this field [Wu *et al.*, 2023].However, despite its significance, BloombergGPT and recent research of the field have limitations in terms of their investigation in reasoning abilities, which have been left out of the scope of research. The focus of predominant research in the financial domain has largely been limited to token or sequence classification tasks [Araci, 2019; Shah *et al.*, 2022], likely due to the scarcity of suitable datasets or tasks for training generative language models. For instance, corpora containing financial reasoning steps, which are essential for training language models for tasks such as investment opinion generation, are mostly confidential in nature and therefore excluded from the training data of publicly available language models [Scao *et al.*, 2022; Black *et al.*, 2022; Touvron *et al.*, 2023]. This limitation poses challenges for developing language models with specialized reasoning capabilities in the financial domain.

Though this study does not involve the development of a finance-native LM of its own, it distinguishes itself from previous research as it comprehensively investigates the circumstances under which specialized financial reasoning capabilities can be enabled.

3 Task Formulation

In this paper we introduce a novel task called Financial Investment Opinion Generation(FIOG), the term encompasses all tasks aiming to train or prompt language models to generate investment opinions in the context of finance, leveraging either parametric or injected knowledge. Our variant of the FIOG task involves providing language models with the necessary information as part of the input. The input information in our variant is provided in two types: full-text and question-and-answer (Q&A). In the full-text type, the input consists of complete text passages, while in the Q&A type, the input comprises pairs of questions and corresponding answers. The Q&A type is used to train and prompt our model via In-Context Question Answering, which will be explained later in the paper. Incorporating investment decision-relevant information as part of the input, enables us to investigate the ability of Language Models (LMs) as reasoning engines, rather than knowledge databases, and allows for a more targeted and effective training process.

4 Dataset Creation

To support further research on financial reasoning, we provide a publicly accessible dataset named sFIOG (Synthetic-Financial Investment Opinion Generation). The sFIOG dataset is generated through the following steps.

1. Collection of expert-written analyst reports: We gathered 1,087 analyst reports from various sources, including J.P Morgan, Truist Financial Corp, and Oppenheimer & Co. These reports cover 752 companies in the U.S stock market.
2. Expert-Written investment thesis set construction: We extracted the "Investment Thesis" and "Related Risk" sections from each analyst report, resulting in a set of expert-written investment theses.
3. Full-Text type input construction: We constructed the Full-Text type input by collecting the marginal sections from the analyst reports.
4. Q&A type input question generation: Using the GPT3.5-Turbo API, we fed the Full-Text type input and required it to generate questions addressing important information.
5. Dummy answer generation: We used the GPT3.5-Turbo API to generate dummy answers for the questions generated in step 4. Human annotators were hired to eliminate answers that deviated greatly from reality.
6. Investment opinion generation: The GPT3.5-Turbo API was employed to generate investment opinions for both types of inputs.

Expert-Written		**Full-Text Type**		**Q&A Type**		
Coverage	Investment Thesis	Full-Text	Investment Thesis	Question	Q&A Pair	Investment Thesis
752	1,087	1,087	4,386	10,437	26,138	11,802

Table 1: Dataset Overview

In step 4, we extract questions from a given text rather than relying solely on a LLM to few-shot generate questions on a given topic. This approach is expected to generate questions that inquire about information deemed important by human experts rather than generating random questions. For comparison, we also construct a set of few-shot generated questions. To assess the lexical and syntactic diversity of each method, we use three metrics: Mass and HD-D for lexical diversity, and Syntactic Sim. for syntactic diversity. Mass and HD-D are established metrics for measuring lexical richness and have been shown to be reliable across texts of different lengths [Torruella and Capsada, 2013; McCarthy and Jarvis, 2010]. A higher HD-D score indicates greater lexical richness, while a higher Mass score indicates the opposite. For syntactic diversity, we use Syntactic Sim., which measures the average pairwise similarity of the dependency tree across generated samples [Oya, 2020]. A higher Syntactic Sim. value indicates greater similarity in syntactic structures across generated samples. As presented in Table 2, our approach resembling question extraction yields synthetic data with a higher degree of both lexical and syntactic diversity.

Generation	**Few-Shot**	**Step 4.**
HD-D	0.811	0.873
Mass	0.034	0.025
Syntactic Sim.	0.578	0.42

Table 2: Quantitative assessment of questions generated via few-shot generation against ours (step 4).

Step 5, adds multiple dummy answers for the questions generated in the prior step. These dummy answers were carefully screened by a human annotator to eliminate those that deviate greatly from reality. We expect this process to add to the diversity of the dataset aiding the fine-tuning of complex reasoning, similar to diverse reasoning [Ho *et al.*, 2022].

Table 1 includes the statistics for the constructed sFIOG dataset. Our dataset encompasses three types of the investment thesis. First, we have 1,087 expert-written investment theses. Second, we have 4,386 investment theses generated with full-text type input. It is noteworthy that the investment thesis generated with the full-text type input exhibits a balanced distribution of buy, hold, and sell opinions, with 1,462 samples for each. Finally, we have 11,802 samples generated with Q&A type input. Each sample was generated with 13 or more Q&A pairs, ensuring that a sufficient amount and diversity of information was provided for the language models to formulate comprehensive investment opinions. More than one sample was generated for each set of Q&A pairs to add to the diversity of the dataset.

The publicly accessible sFIOG dataset is limited to the Q&A type input subset of the dataset due to the restriction of third-party sharing of the expert-written analyst reports collected from the web. To the best of our knowledge, the publicly accessible version of the sFIOG dataset is comprised only of synthetically generated questions, answers, and investment opinions.

5 In-Context Question Answering

Both LLM or their smaller variants have been pointed out to hallucinate, or generate context unfaithful from real world information [Ji *et al.*, 2023]. Even if these LMs manage to accurately retrieve real-world information that they have memorized during the pre-training stage, there are still risks of the information being outdated or non-stationary [Son *et al.*, 2023]. To address this issue, we propose In-Context Question Answering, where a list of question-and-answer pairs are provided instead of full-text contexts. Through experiments, we demonstrate that our approach has several advantages compared to previous full-text in-context learning approaches when zero-shot prompting LLMs.

First, our findings indicate that generations grounded on Q&A pairs exhibit a higher degree of controlled behavior, or a lower likelihood to generate unintended context, compared to conventional in-context learning generations. For instance, approximately 11.12% of the samples generated with conventional in-context learning included analysis on the pandemic, even though the investment opinion was intended for the post-pandemic era. In contrast, when using in-context question answering, the chances of generated samples to discuss pandemic-related issues, despite their absence in the provided Q&A sets, was merely 1.63%. This suggests that the proposed in-context question answering may be a more effective approach to zero-shot prompt LLMs to generate controlled outputs, making it more suitable for specific contexts and scenarios, such as post-pandemic era financial analysis. We speculate that such behavior is because in-context question answering delivers a refined version of information with most of the irrelevant text removed, resulting in a more concise and focused input. Language models are susceptible to distraction from irrelevant text [Shi *et al.*, 2023], and the provision of context in a Q&A format allows them to concentrate on the core information without being influenced by unnecessary or irrelevant sentences. This conciseness and absence of irrelevant text in the Q&A format may enable language models to better align with the intended task, leading to improved performance and controlled behavior in generating contextually relevant and accurate content.

Second, we conducted a survey with hired human annotators using a subset of 1,000 samples from each type. In order to assess the performance of our LLM-based evaluators in comparison to human annotators, we also conducted the identical survey using GPT-4 as a respondent, following previous

research on G-Eval [Liu *et al.*, 2023]. The survey presented respondents with three samples at a time, one from each of the expert-written, full-text type, and Q&A type. They were then required to answer two questions:

1. *Which investment thesis contains the most investment helpful information?*
2. *Which investment thesis presents a more logically structured and reasonable argumentation?*

Figure 1, indicates that human evaluators perceived Q&A type generation to contain the most investment-helpful information in 61.2% of cases and demonstrated the most coherent argumentation in 48% of cases. In contrast, Full-Text type generation was found to have relatively fewer investment-helpful information, which may be attributed to the presence of irrelevant text that could disrupt the language model's output. Notably, the generated samples in either full-text or Q&A type were preferred by human annotators over the expert-written samples for both questions. We speculate that this preference for generated samples over expert-written thesis may be due to the fact that expert-written thesis are tailored for professionals with domain-specific expertise, and may omit explanations or assumed background knowledge, potentially affecting their comprehensibility to human evaluators. An investigation of the inter-annotator agreement was conducted on a subset of 350 samples for each question, revealing a decent Krippendorff's alpha of 0.63 for question 1 and 0.68 for question 2.

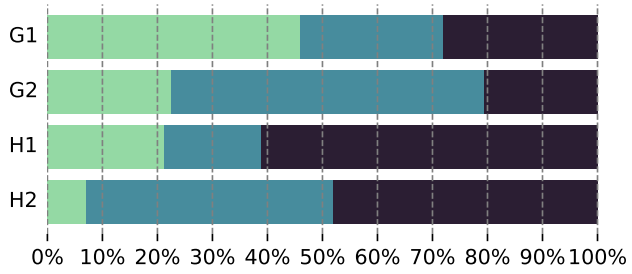

Figure 1: Qualitative Evaluation of Collected Investment Theses: Green denotes expert-written, blue represents full-text type, and dark blue indicates Q&A type. G1 and G2 refer to GPT-4 answers for Question1 and Question2, respectively. H1 and H2 denote human answers for Question1 and Question2, respectively.

Furthermore, we conduct the identical survey using GPT-4, following G-Eval, we use the following prompt:

> *You are a professional financial researcher. You will be given an investment thesis. Your task is to rate the thesis on the following metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*
>
> *Evaluation Criteria:*
>
> *Investment-Helpfulness (1-5) - the quality and diversity of financial facts provided in the passage. The investment thesis should provide a diverse set of quantitative information. Quantitative information must include numerical values. Concentrate on the diversity and amount of facts provided. Ignore the argumentation for the moment.*
>
> *Financial Argumentation (1-5) - the quality of the financial reasoning and supporting evidence in the passage. This includes the logical coherence of the financial argument, the strength of the financial evidence provided, and the overall persuasiveness of the financial argument. Specifically, this criterion evaluates the effectiveness of the financial analysis and the quality of the financial data used to support the investment thesis.*

The responses from LLMs were compared with the decision of human annotators to investigate the efficacy of LLM applications for the evaluation of financial reasoning. Unlike previous research [Gilardi *et al.*, 2023], our study found a notable disparity between GPT-4 and human judgments, with low correlation observed regardless of the presence of CoT explanations. Figure 2 displays the confusion matrix comparing the decisions of human and LLM evaluators. The results indicate that the agreement rate between the two evaluators was only 29.26%, and 34.6% for each question correspondingly. Moreover, the Spearman correlation coefficients between human and LLM decisions were -0.07 for question one and -0.073 for question two, significantly lower than that of previous research that reported 0.514 [Liu *et al.*, 2023]. This disparity may be attributed to two key factors. First, unlike prior research that focused on LLMs' evaluation of summarization quality or zero-shot classification of tweets, our study required the LLMs to evaluate financial reasoning, which is a more intricate and complex task. Additionally, LLMs were never trained for such tasks, which may have impacted their performance in evaluating the quality of financial reasoning. Secondly, the financial domain poses unique complexities, including diverse nomenclature and domain-specific knowledge, which may present a challenge for generic LLMs to fully comprehend and accurately evaluate the coherence of financial reasoning. Following our findings, LLMs are no longer used as evaluators in this paper.

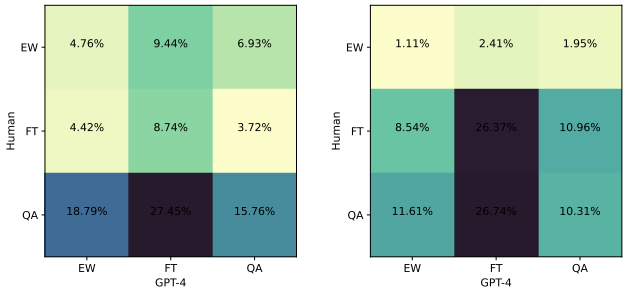

Figure 2: Left for Q1, Right for Q2.

Overall, the aforementioned experiments yield two important findings. Firstly, the results discover that LLMs are inadequate as evaluators for financial reasoning tasks, given the limited alignment observed between LLMs and human eval-

Base Model	Instruction-Tuning	ROUGE-L				BERTScore			
		type#1	type#2	type#3	average	type#1	type#2	type#3	average
LLama	✓	**0.283**	**0.178**	**0.359**	**0.273**	0.830	**0.849**	**0.855**	**0.845**
Galactica	✓	0.108	0.028	0.114	0.083	0.794	0.807	0.799	0.800
GPT-J	✓	0.159	0.023	0.183	0.122	**0.836**	0.692	0.836	0.788
Pythia(2.8B)	✓	0.022	0.000	0.023	0.015	0.731	0.769	0.735	0.745
LLama	✗	0.080	0.123	0.180	0.128	0.592	0.778	0.723	0.698
Galactica	✗	0.086	0.027	0.097	0.070	0.777	0.804	0.773	0.785
GPT-J	✗	0.054	0.023	0.139	0.072	0.773	0.692	0.818	0.761
Pythia(2.8B)	✗	0.017	0.012	0.018	0.016	0.729	0.795	0.728	0.751

Table 3: Results for LLama, Galactica, GPT-J, and Pythia (2.8B), both with and without instruction-tuning, obtained on the sFIOG test dataset. The evaluation was carried out across three distinct subsets. Type#1 consisted of companies and questions from the training set with new corresponding answers. Type#2 featured companies from the training set paired with new, previously unencountered question-and-answer combinations. Lastly, Type#3 introduced companies not present in the training set, accompanied by new question-and-answer pairs.

uators. Secondly, the proposed In-Context Question Answering method represents a promising alternative to traditional prompting methods, exhibiting improved controlledness and generating better-quality reports. Notably, this method could be applicable to a broader range of fields beyond finance, wherever controlled generation of information-rich texts is required.

6 Experiments

6.1 Experimental Setup

In this research, we assessed four GPT variants (2.8B to 13B parameters) with and without instruction tuning, as detailed in Table 4. This comparison aimed to identify the point at which the ability to generate financial reasoning emerges.

Base Model	Instruction-Tuned	Param.
Pythia	dolly-v2-3b	2.8B
GPT-J	dolly-v1-6b	6B
Galactica	galpaca-6.7b	6.7B
LLama	vicuna-13b-delta-v1.1	13B

Table 4: Summary of GPT variants employed in the experiments, detailing their parameter sizes and whether they underwent instruction tuning. Checkpoints for instruction-tuned models were imported from HuggingFace.

The test dataset for this study is comprised of three distinct subsets to evaluate the performance of the GPT variants in different settings. The first subset included companies and questions that appeared in the training set but with new corresponding answers. The second subset featured companies from the training set but paired with new, previously unencountered question-and-answer combinations. Lastly, the third subset introduced companies that did not appear in the training set, accompanied by new question-and-answer pairs. Through this dataset split we assess the models' capabilities in generating financial reasoning across varying degrees of familiarity and novelty.

To evaluate the generated context, we used both automated metrics and human evaluations. Automated metrics included rouge-2 and rougeL [Lin, 2004], measuring text overlap, and BERTScore [Zhang *et al.*, 2019], assessing semantic similarity. As mentioned previously, we excluded LLM-based evaluators due to their misalignment with human judgments.

Models in this study were trained using Lora [Hu *et al.*, 2021] and quantization for enhanced hardware efficiency, with a maximum token length of 2048 and an AdamW optimizer. Each model was trained in three epochs on the full sFIOG dataset, which is consisted of 11,802 samples. During the test phase, decoding settings were configured to enhance the quality and diversity of generated outputs, while ensuring a fair comparison across models. The parameters were set as follows: top_k=50, top_p=0.95, no_repeat_ngram_size=3, and max_new_tokens=512. By setting a fixed maximum number of tokens, we prevented models that generate longer sequences from appearing to outperform others in the evaluation.

6.2 Model Scale and Financial Reasoning

In Table 3, we present the results for LLama [Touvron *et al.*, 2023], Galactica [Taylor *et al.*, 2022], GPT-J, and Pythia (2.8B) [Biderman *et al.*, 2023], with and without instruction-tuning, on the sFIOG test dataset. Our findings indicate that the ability to generate coherent investment opinions emerges in models with sizes between 2.8B $\tilde{6}$B and continues to improve as the model scales. For instance, LLama demonstrates superior performance, achieving the highest average scores in ROUGE-L (0.217) and BERTScore (0.821). There are two possible explanations for the scaling behavior of financial reasoning abilities in these models: (1) larger models are typically trained on more tokens, thereby accumulating a greater amount of knowledge essential for generating well-informed investment theses, and (2) the architecture of larger models inherently allows for improved reasoning capabilities, enabling them to better analyze and synthesize the information they have learned. Consequently, as model size expands, it leads to a stronger ability to effectively generate financial reasoning, as demonstrated by the superior performance of the LLama model in our experiments. An exception in the scaling behavior is observed between GPT-J and Galactica, with GPT-J surpassing Galactica in performance, despite its smaller size. We posit that this discrepancy may arise from two factors: (1) GPT-J is trained on a substan-

tially larger corpus of tokens (402 billion) from a general domain, while Galactica has been trained on a smaller, science-specific corpus (106 billion); (2) The size difference between the two models is relatively minimal, at just 0.7B. This observation is consistent with recent research, suggesting that training smaller models with an increased number of tokens beyond the chinchilla optimal point can yield improved performance [Touvron *et al.*, 2023]. Furthermore, this finding emphasizes the potential trade-offs of domain-specific training, which could compromise a model's robustness across broader contexts.

6.3 Instruction-Tuning and Financial Reasoning

We observe that instruction-tuning plays a significant role in enhancing the performance of all models across both evaluation metrics. However, the degree of improvement varies among models, which may be due to the difference of instruction-tuning datasets used to fine-tune each model. It is noteworthy that Pythia (2.8B), the smallest model employed in our experiments, failed to demonstrate the ability to generate coherent financial reasoning, even when instruction-tuning was applied. This finding implies that the ability to generate financial reasoning could be an emergent property that becomes evident as the model size exceeds a specific threshold.

6.4 Dataset and Financial Reasoning

In examining the performance of the models across each subset of the dataset, we find that the models exhibit their weakest performance in type#2 questions, which involve companies included in the training set but are accompanied by new question-and-answer pairs. This observation departs from the authors' initial assumption that type#3 questions, featuring companies not present in the training set, would pose the greatest challenge. The results demonstrate that generating financial opinions for novel question-answer pairs concerning familiar companies is a more demanding task for the models. This finding aligns with past research, suggesting that the non-stationary knowledge acquired during the training process may hinder the models' capacity to generalize their knowledge effectively and apply it to novel situations involving known entities [Son *et al.*, 2023].

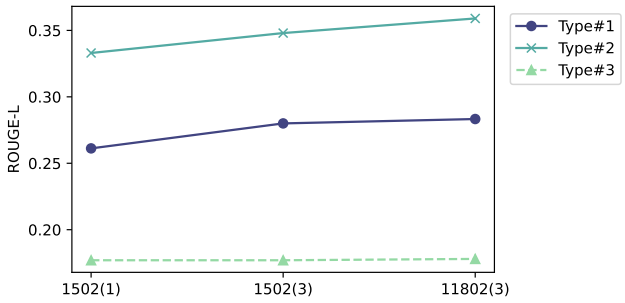

Figure 3: Performance of Vicuna across varying training steps. The x-axis denotes the training step, presented in the format sample_size(epoch). The y-axis displays the corresponding ROUGE-L scores.

Furthermore, we evaluate the financial reasoning abilities of the best-performing model, instruction-tuned LLama 13B, across different dataset sizes and training steps. Specifically, we conducted experiments by training the model for (1) 3 epochs on an 11,802-sample dataset, (2) 3 epochs on a smaller 1,502-sample dataset, and (3) 1 epoch on the same 1,502-sample dataset, where each company in the full dataset was represented by 2 samples. Our results reveal that LLama's performance improved with an increasing number of training steps. However, even the model trained on the smallest configuration exhibited superior performance compared to the instruction-tuned GPT-J, which was the second-best model trained on the full dataset. These findings suggest that model size may be a critical factor in generating coherent financial reasoning, while dataset size may not be as significant.

6.5 Human Preference

To comprehensively evaluate the performance of each instruction-tuned model, a human preference test was conducted on their generated outputs. A panel of human evaluators was presented with four texts, each from one of the models, namely LLama, Galactica, GPT-J, and Pythia(2.8B), and asked to indicate their preference based on several factors, including coherence, relevance, and fluency. The results of the human preference test, depicted in Figure 4, reveal that the LLama model was the most preferred choice, followed by the GPT-J model. This outcome is consistent with the findings of our previous investigation, which utilized automated metrics.

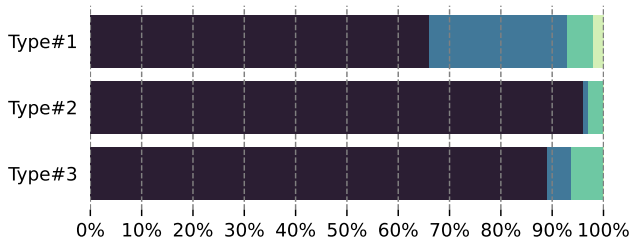

Figure 4: Human preference on generated samples. Dark Blue for LLama, Green for Galactica, Blue for GPT-J, and Yellow for Pytha(2.8B)

7 Limitations and Future Work

It is worth noting that due to hardware constraints, we were unable to investigate the emergent characteristic of financial reasoning ability on models beyond 13B parameters. Additionally, we do not open-source expert-written samples due to copyright issues. Nevertheless, this work still represents the most comprehensive investigation to date on the behavior of language models for financial reasoning generation and the first to make a dataset for financial reasoning publicly available. Going forward, we encourage the financial natural language processing community for collaborative efforts to create larger datasets for financial reasoning tasks and to experiment with larger language models. We believe that such ef-

forts will enable more comprehensive evaluations of language models and their potential for financial reasoning generation, ultimately advancing the state of the art in this field.

8 Conclusion

To the best of our knowledge, this work represents the first public effort to investigate the financial reasoning ability of language models. Our research seeks to contribute to the understanding of the efficacy of language models in the field of finance, with a particular emphasis on their ability to engage in sophisticated reasoning and analysis within the context of investment decision-making. We confirm that the ability to generate coherent investment opinions first emerges in models with 6B parameters and scales as the model gets larger until 13B parameters. Additionally, this study introduced a novel prompting method, In-Context Question-Answering, truth-faithful generation of LLMs. The research also identified the limitations of LLMs in aligning with human evaluators for evaluating financial texts. Finally, we make a valuable contribution to the field by open-sourcing sFIOG, a dataset consisting of 11,802 synthetic investment thesis samples.

References

[Araci, 2019] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[Biderman *et al.*, 2023] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023.

[Black *et al.*, 2022] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

[Fu *et al.*, 2023] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*, 2023.

[Gilardi *et al.*, 2023] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[Ho *et al.*, 2022] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[Kojima *et al.*,] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Liu *et al.*, 2023] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[McCarthy and Jarvis, 2010] Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Oya, 2020] Masanori Oya. Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the euclidean distance of their dependency trees. In *Proceedings of the 34th pacific Asia conference on language, information and computation*, pages 225–233, 2020.

[Scao *et al.*, 2022] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[Shah *et al.*, 2022] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*, 2022.

[Shi *et al.*, 2023] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*, 2023.

[Singhal *et al.*, 2022] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[Son *et al.*, 2023] Guijin Son, Hanwool Lee, Nahyeon Kang, and Moonjeong Hahm. Removing non-stationary knowledge from pre-trained language models for entity-level

sentiment classification in finance. *arXiv preprint arXiv:2301.03136*, 2023.

[Taylor *et al.*, 2022] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[Torruella and Capsada, 2013] Joan Torruella and Ramón Capsada. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95:447–454, 2013.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Trautmann *et al.*, 2022] Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

[Wei *et al.*, 2022a] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[Wei *et al.*, 2022b] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[Wu *et al.*, 2023] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[Yunxiang *et al.*, 2023] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.