Can GPT Models be Financial Analysts?
An Evaluation of ChatGPT and GPT-4 on Mock CFA Exams

Ethan Callanan[1,†], Amarachi Mbakwe[2,†,‡], Antony Papadimitriou[3,†], Yulong Pei[3,†], Mathieu Sibue[3,†],
Xiaodan Zhu[1], Zhiqiang Ma[3], Xiaomo Liu[3], and Sameena Shah[3]

[1]Queen's University
[2]Virginia Tech
[3]J.P. Morgan AI Research
[1]{e.callanan,xiaodan.zhu}@queensu.ca, [2]bmamarachi@vt.edu, [3]{first.last}@jpmchase.com

Abstract

Large language models (LLMs) have demonstrated remarkable performance on a wide range of natural language processing tasks, often matching or even outperforming state-of-the-art task-specific models. They have the potential to make a significant impact on financial professions and to have profound influence on the finance industry. In this study, we leverage mock exam questions of the Chartered Financial Analyst (CFA) program to conduct a comprehensive evaluation of ChatGPT and GPT-4 in financial analysis, considering zero-shot, chain-of-thought, and few-shot scenarios. We present an in-depth analysis of the models' performance and limitations, and estimate whether they would have a chance at passing the CFA exams. Finally, we outline insights into potential strategies and improvements to enhance the applicability of LLMs in finance. In this perspective, we hope this work paves a way for future studies to continue enhancing LLMs for financial analysis.[1]

1 Introduction

Tracking the progress of the most advanced large language models (LLMs) and their performance on major financial professional certifications has a profound impact on the financial industry. In general, language models and natural language processing (NLP) systems have played a pivotal role in enhancing various services, such as customer relations, financial question answering (Wang et al., 2022), document understanding (Kim et al., 2022), and report summarization (Abdaljalil and Bouamor, 2021). Despite these advancements, applying NLP in finance poses unique challenges, such as the distinct nature of financial tasks, linguistic structures, and specialized terminology. As a result, the performance of general NLP models often falls short when applied to finance-related tasks — the specific challenges of financial reasoning problems warrant further investigation.

LLMs have the potential to make a significant impact on financial professions, and by extension on professional qualifications such as the Chartered Financial Analyst (CFA) Program.[2] With more than 190,000 charterholders across 160 markets worldwide, the CFA Program is arguably the most recognized certification in finance. Its exams are known for their meticulous yet practical assessment of financial expertise, making their resolution an ideal use case to gauge the capabilities of LLMs in handling complex financial analyses and reasoning. A human being often spends years to learn the required knowledge for the CFA examination.

Which of the following is most likely an assumption of technical analysis?
A. Security markets are efficient
B. Market trends reflect irrational human behavior
C. Equity markets react quickly to inflection points in broad economy

(a) Level I sample question

Paris Rousseau, a wealth manager at a US-based investment management firm, is meeting with a new client. The client has asked Rousseau to make recommendations regarding his portfolio's exposure to liquid alternative investments [...]
[Table Evidence]
The AFFO per share for Autier REIT over the last 12-months is closest to:
A. $6.80; *B. $7.16;* *C. $8.43.*

(b) Level II sample question

Figure 1: CFA example questions (source: CFA Institute); the question appears in bold, the multiple choices in blue and italic, and the vignette/case description in orange and italic.

In this paper, we rigorously assess the out-of-the-box capabilities of LLMs in real-world financial reasoning problems by conducting an evaluation

[1] The code used in this paper is available upon request to any {first.last}@jpmchase.com among the authors.
[†] Equal contribution.
[‡] Work done while interning at J.P. Morgan AI Research.

[2] https://www.cfainstitute.org/en/programs/cfa/exam

on mock exam questions of the CFA Program. Our work focuses on two closed-source, non-domain specific LLMs, ChatGPT and GPT-4, using various popular prompting techniques. Although there are other LLMs available, the top models in the GPT series, e.g., GPT-4, do represent the state of the art on most benchmarked and in-house tasks, and are adequate to support the main conclusions of this study. In summary, our contributions are as follows:

- We conduct the first comprehensive evaluation of state-of-the-art LLMs on CFA mock exams, considering zero-shot, few-shot, and chain-of-thought prompting scenarios. We demonstrate that some of the models have a decent chance to pass the tests.
- We present an in-depth analysis of the models' performance and limitations in solving these financial analysis and reasoning problems, including investigations at different topics and levels of the exams.
- We outline insights into potential strategies and improvements to enhance the applicability of LLMs in finance, suggesting new avenues for research and development.

2 Related Work

LLMs and Finance. As highlighted in (Brown et al., 2020; Wei et al., 2022), LLMs exhibit remarkable generalization across diverse topics. However, their application to finance, a domain demanding intricate reasoning with specific concepts, mathematical formulas, and visual aids, poses significant challenges. Approaches like continued pre-training (Araci, 2019; Wu et al., 2023), supervised fine-tuning (Mosbach et al., 2023; Yang et al., 2023), and retrieval augmented generation using external knowledge (Lewis et al., 2020) have been proposed to address these challenges. Notably, (Li et al., 2023) has extensively benchmarked the out-of-the-box capabilities of newer instruction-tuned LLMs in finance.

Evaluation of LLMs on Exams. Various studies have scrutinized LLMs in exams like the United States medical licensing exam (Kung et al., 2023), free-response clinical reasoning exams (Strong et al., 2023), college-level scientific exams (Wang et al., 2023), and the Bar exam (Katz et al., 2023). Notably, (Wang et al., 2023) found LLMs lacking in complex scientific reasoning, while (Bang et al., 2023) demonstrated ChatGPT's outperformance in NLP tasks. Our paper contributes by evaluating the financial reasoning abilities of ChatGPT and GPT-4 (Li et al., 2023) on the CFA exams. Refer to Appendix C for more detailed related work.

	Level I			Level II		
Topic	**Calc.**	**#Tab**	**Len**	**Calc.**	**#Tab**	**Len**
Ethics	0.7%	0.01	125	0.0%	0.00	1013
Quant. Meth.	70.5%	0.26	131	27.8%	0.00	1256
Economics	50.6%	0.25	121	66.7%	2.00	1115
Fin. Reporting	57.7%	0.35	151	53.6%	2.79	1383
Corp. Issuers	59.3%	0.28	120	44.4%	1.67	930
Equity Invest.	52.5%	0.19	112	45.8%	1.00	1053
Fixed Income	43.0%	0.06	87	50.0%	1.45	779
Derivatives	20.7%	0.00	65	75.0%	2.00	816
Alter. Invest.	36.4%	0.06	85	66.7%	2.00	840
Port. Manage.	38.3%	0.18	110	56.3%	2.13	1077
Overall	42.4%	0.17	116	45.5%	1.47	1058

Table 1: Question characteristics by topic; percentage of questions requiring calculation, average number of table evidence per question, and average prompt length (estimated using the tiktoken Python package).

3 The Chartered Financial Analyst Exam and Data

The CFA Program, a three-part exam, assesses investment tools, asset valuation, portfolio management, and wealth planning fundamentals. It is pursued by individuals in finance, accounting, economics, or business for roles in investment, risk, and asset management upon successful completion.

The CFA Institute does not release official past exams, so we use mock CFA exams written by CFA Charterholders and based on past CFA Institute assessments to conform to current testing formula and level of difficulty.

Other than availability, evaluating a model on the CFA Program poses another challenge in that the level III questions are open-ended written response questions, necessitating expensive human expert grading. As such, in this work, we focus on levels I and II and leaving Level III for future work.

Each exam level adheres to a specific format. Level I has 180 multiple choice questions (MCQs) on ten finance topics (Table 1). Level II includes 22 vignette-based item sets with 88 MCQs. Level III combines vignette-supported essay questions and MCQs. Example MCQs from the CFA Institute are illustrated in Figure 1.

We collected a total of five Level I exams and two Level II exams. In our experiments, we ensure each topic is represented in a similar proportion to the original CFA section (Figures 2 and 3 in

Appendix A). Table 1 summarizes the statistics of the exam questions we collected.

4 Experiment Setup

Prompting Paradigm. Our study examines the following typical prompting methods:

- **Zero-shot (ZS) prompting.** Tracking the off-the-shelf performance of generically trained AI models such as `ChatGPT` and `GPT-4` is critical for programs like the CFA. We accordingly test ZS prompting performance.
- **Few-shot (FS) prompting.** We test 2-shot (2S), 4-shot (4S), 6-shot (6S), and 10-shot (10S) settings. When selecting examples, we apply two different strategies: (i) randomly sampling from the entire set of questions within the exam level (2S, 4S and 6S), and (ii) sampling one question from each topic in the exam level (10S). The latter aims at enabling the models to discern distinct attributes of the topic.
- **Chain-of-thought (CoT) prompting.** We follow ZS CoT (Wei et al., 2022), which has the added benefit of allowing us to analyze the "problem-solving process" of the models and determine where and why an answer goes wrong.

Implementation Details. We conduct the experiments using the OpenAI ChatCompletion API (`gpt-3.5-turbo-0613` and `gpt-4-0613` versions, 32K context window for FS prompting), with the temperature parameter set to zero. The prompt templates and settings are in Appendix B. To confirm the models had not memorized the mock exams as part of their training data, we employ memorization tests as in (Kıcıman et al., 2023).

Metric. We compare the predictions against the exams' solution set. *Accuracy* served as our sole evaluation metric throughout this study.

5 Overview of the Experiment Results

LLMs struggle more on Level II than on Level I. No matter the prompting paradigm employed, both models encounter more difficulties correctly answering the Level II item-sets than the independent questions from Level I (Table 2). We suggest that three factors might have negatively affected the performance of LLMs in Level II.

Firstly, the case description of a Level II item-set increases the length of the input prompt and could dilute the useful information it contains. Indeed, prompts for Level II are on average 10× longer than the Level I ones; confronting Tables 1 and 2 shows that topics associated with poor performance usually present longer contexts in both Level I and II. In addition, the detailed case descriptions from Level II depict realistic day-to-day situations that contrast with the more general questions from Level I — LLMs thus need to abstract from case-specific details so as to identify the underlying finance concepts. Secondly, each item from the grouped item-set in Level II tends to go more in-depth about a specific finance topic than those in Level I, thus leading to more specialized and intricate problems. Lastly, Level II features a slightly higher proportion of questions requiring calculations and a much higher proportion containing table evidence (Table 1). Given the known limitations of LLMs on numerical and table reasoning (Frieder et al., 2023; Chen et al., 2022), this could also result in the low accuracy on Level II.

`GPT-4` outperforms `ChatGPT` in almost all experiments, but certain finance topics remain challenging for both. As shown in Table 2, `GPT-4` consistently beats `ChatGPT` in all topics in Level I and most topics in Level II, irrespective of the prompting paradigm.

In Level I, both LLMs perform best in Derivatives, Alternative Investments, Corporate Issuers, Equity Investments, and Ethics. The explicit mention of common finance notions in the questions (e.g., options, arbitrage, etc.) could be a factor, notions which `ChatGPT` and `GPT-4` might have encountered during pretraining or instruction-tuning and that may help facilitate resolution. For Derivatives and Ethics, the question complexity is reduced due to the low amount of calculations and table understanding required to answer correctly (Table 1). However, both models perform relatively poorly in Financial Reporting and Portfolio Management (especially in ZS and CoT), with `ChatGPT` also struggling a lot more on highly computational topics such as Quantitative Methods. Indeed, the problems within these topics are more case-based, applied, computational, and CFA-specific than the ones from the aforementioned topics. They also tend to include more table evidence and complex details (Table 1).

The results are more nuanced in Level II. `ChatGPT` struggles on Alternative Investments and Fixed Income compared to `GPT-4`, while `ChatGPT` outperforms `GPT-4` in Portfolio Management and Economics. Interestingly enough, both models now

Exam	Level I						Level II					
	ChatGPT			GPT-4			ChatGPT			GPT-4		
Category	**ZS**	**CoT**	**2S**	**ZS**	**CoT**	**10S**	**ZS**	**CoT**	**10S**	**ZS**	**CoT**	**4S**
Ethics	59.2	59.2	**64.6**	80.3	78.9	**82.4**	31.3	**37.5**	21.9	43.8	56.3	**62.5**
Quantitative Methods	53.9	50.0	**59.7**	**78.0**	76.0	76.0	44.4	**55.6**	54.2	66.7	66.7	**72.2**
Economics	**68.0**	63.7	**68.0**	74.1	73.6	**76.2**	**66.7**	58.3	62.5	41.7	**58.3**	**58.3**
Financial Reporting	54.0	53.4	**60.1**	68.2	**70.8**	70.0	39.6	31.3	**44.8**	54.2	**66.7**	56.3
Corporate Issuers	71.4	69.8	**74.2**	74.4	74.6	**75.3**	**55.6**	50.0	50.0	77.8	77.8	**83.3**
Equity Investments	59.4	60.9	**62.5**	**80.3**	70.5	68.8	60.4	60.4	**60.9**	**65.0**	58.8	62.5
Fixed Income	55.6	60.2	**63.6**	**74.9**	60.2	73.6	**38.9**	27.8	34.4	60.0	**62.2**	55.6
Derivatives	61.1	68.5	**73.0**	90.5	93.8	**96.0**	50.0	**58.3**	47.9	**66.7**	58.3	58.3
Alternate Investments	60.7	60.7	**62.9**	75.9	**77.1**	72.1	33.3	33.3	**58.3**	66.7	50.0	**83.3**
Portfolio Management	58.3	48.3	**61.7**	63.7	71.7	**79.6**	47.2	**66.7**	59.7	36.1	55.6	**61.1**
Overall	58.8	58.0	**63.0**	73.2	74.0	**74.6**	46.6	47.2	**47.6**	57.4	61.4	**61.9**

Table 2: ChatGPT and GPT-4 accuracy across topics on Level I and II exams for ZS, CoT, and FS prompting. For FS, the table only retains the results from the k-shot that achieves highest overall performance.

Model	Setting	Level I	Level II
ChatGPT	ZS	58.8	46.6
	CoT	58.0	47.2
	2S	**63.0**	46.6
	4S	62.3	45.7
	6S	62.2	47.0
	10S	62.4	**47.6**
GPT-4	ZS	73.2	57.4
	CoT	74.0	61.4
	2S	73.9	60.2
	4S	73.8	**61.9**
	6S	74.5	60.2
	10S	**74.6**	60.2

Table 3: Overall performance (accuracy) of ChatGPT and GPT-4 on Level I and II in ZS, CoT, and FS settings.

demonstrate low answer accuracy in the Ethics item-sets of Level II. This could originate from the more in-depth, situational, and detailed character of the problems from Level II in comparison to Level I.

CoT prompting yields limited improvements over ZS. CoT does not help LLMs in the evaluation as much as we expected, although CoT performs better than ZS in almost all cases (Table 3). In Level I, CoT prompting hardly benefits GPT-4 and deteriorates the performance of ChatGPT instead. Particularly, both models are affected in Quantitative Methods due to the hallucinations in mathematical formula and calculations. In Level II, CoT prompting yields a decent 7% improvement over ZS for GPT-4, but a disappointing 1% for ChatGPT. CoT benefits both LLMs in Ethics and Portfolio Management, where its explicit step-by-step reasoning over long and intricate evidence is usually helpful. Section 6 further investigates the reasons explaining such observations. We note that it is not easy to identify topics where CoT consistently improves or worsens the models' performance across levels, e.g., GPT-4 sees an accuracy improvement of 23% in Level II Financial Reporting, while ChatGPT has a 21% decrease.

A few in-context exemplars help more than CoT. Compared with ZS and CoT prompting, FS offers significant performance improvement for ChatGPT on the Level I exams (Table 3). 2S prompting yields the best performance across all topics and overall in Level I for ChatGPT. The dominance is not as significant in Level II, but FS prompting still manages to achieve the best overall score for both models. Interestingly, for Level II, ChatGPT gains the most from 10S prompting, which suggests a more holistic FS approach across multiple topics helps the model crack complex questions. The overall trend in the results is that FS prompting seems to offer better assistance to less complex models (ChatGPT) when tested on seemingly simpler exams (Level I)

We argue that the better performance from FS credits to the answers associated with the examples in FS, which also might help the model understand how to best use the table evidence or other information contained in a question. The advantage from FS vanishes as the question complexity increases in Level II, where a combination of FS and CoT might be a potential approach worth further exploration.

6 Detailed Analysis on CoT

Surprisingly, CoT only marginally improves the models' performance on most tests and is even slightly detrimental to the performance of ChatGPT on Level I exams (Table 3). We dive deeper into this phenomenon, as ZS CoT is often reported to outperform ZS prompting (Kojima et al., 2023).

To better understand CoT errors, we examine all instances where non-CoT is correct while CoT

Type of Error	ChatGPT	GPT-4
Knowledge	55.2%	50.0%
Reasoning	8.6%	10.7%
Calculation	17.2%	28.6%
Inconsistency	19.0%	10.7%

Table 4: Error modes introduced when using CoT on Level I questions. These errors do not occur with non-CoT prompting.

Type of Error	ChatGPT	GPT-4
Knowledge	70%	80%
Reasoning	20%	20%
Out of Tokens	10%	0%

Table 5: Error modes introduced when using CoT on Level II questions. These errors do not occur with non-CoT prompting.

is incorrect, and categorize the errors as one of: Knowledge, Reasoning, Calculation, or Inconsistency. Knowledge errors are those where the model lacks critical knowledge required to answer the question. This includes an incorrect understanding of some concept, not knowing the relationship between concepts, or using an incorrect formula to answer a question requiring calculations. Reasoning errors occur when the model has all the correct knowledge, but either over-reasons in its response, or hallucinates some additional requirements or information not present in the question. Calculation errors are errors pertaining to some incorrect calculation (using a correct formula), or failing to accurately compare or convert results. Errors of inconsistency are when the model's thinking is entirely correct, yet it chooses the wrong answer.

6.1 Underperformance of CoT on Level I

ChatGPT. Table 4 underlines that knowledge-based errors are the most common error mode for ChatGPT, constituting over half of all errors VS. non-CoT. This implies that, with CoT reasoning, the gaps in the internal knowledge of LLMs are magnified. As the model begins to think through its answer, it states its incorrect assumptions, which it proceeds to rationalize in the context of the question, thereby skewing the rest of the answer towards a wrong choice. Without using CoT reasoning, the model is able to make an "educated guess" where any incorrect knowledge has less of an opportunity of skewing the guess towards an incorrect answer. With a 1/3 chance of guessing correctly (plus any contextual hints that may lie in the question), guessing is a more accurate strategy when ChatGPT lacks the knowledge to reason correctly.

This same principal similarity explains calculation and reasoning errors, where one or a few off-track token generations can throw off the rest of the answer, resulting in an incorrect conclusion.

The instances where the model is entirely correct but makes an incorrect conclusion or selects the wrong answer are more enigmatic. In about half of these cases, it seemingly fails to generate a stop token upon coming to the conclusion, leading it to restate the concluding sentence with another option selected. In other cases, there appears to be a disconnect between the thought process and the answer selection. As we are using OpenAI's API to retrieve structured output, our leading suspicion is that in these cases the ordering outlined in the system prompt is missed or ignored, and the answer is generated first.

GPT-4. The instances where CoT introduces errors for GPT-4 is half the number of instances where CoT introduces errors for ChatGPT. In these instances, GPT-4 also displays knowledge errors as the most common error mode. However, unlike ChatGPT, almost none of these knowledge errors stem from using incorrect formulas. This, along with the fact that there are less knowledge errors in total, shows that GPT-4 has more complete internal knowledge of both financial information and especially financial formulas and calculation methods. Even when GPT-4 finds the correct formula for a question involving calculations, it still struggles to perform the required calculation correctly. ChatGPT also frequently makes these sorts of errors in conjunction with wrong formula usage, which underlines the well-known and more foundational shortcoming of LLMs' mathematical abilities (Frieder et al., 2023).

GPT-4 also displays far fewer inconsistency errors than ChatGPT. It appears to have a much stronger ability to connect its reasoning to the answers and to make comparisons. The one error type that GPT-4 makes more frequently than ChatGPT when it fails is reasoning errors. It seems that, along with GPT-4's greater ability to reason, it has a greater chance of "talking itself" into incorrect lines of reasoning.

6.2 CoT Benefits on Level II

Level II exam questions require more interpretation of the information provided than Level I questions, as test-takers must determine what parts of the case are relevant to the question, and some information may be missing altogether. Using CoT helps the model reason over the information and filter what is relevant to the question from the case, as evidenced by the results in Table 3. However, knowledge errors still persist in Level II, and outnumber reasoning errors for both ChatGPT and GPT-4 (Table 5).

7 Can LLMs pass the CFA exams?

7.1 CFA Pass Scores

The most intriguing question in this study probably is "Can LLMs pass the CFA exams?". Conclusively determining whether a given score would suffice to pass the CFA exams is difficult because the CFA Institute refrains from disclosing the minimum passing score (MPS) for its examinations. The MPS is uniquely established for each individual exam, guided by the standards established by the CFA Institute in 2011. The CFA Institute employs the "Angoff Standard Setting Method" to ascertain the pass rates for CFA exams, and involves a group of CFA Charter holders convening to assess the difficulty level of the questions.

	ChatGPT			GPT-4		
Exam	**ZS**	**CoT**	**FS**	**ZS**	**CoT**	**FS**
Level I	F	F	F	P	P	P
Level II	F	F	F	U	P	P

Table 6: ChatGPT and GPT-4 ability to pass Level I and Level II Exams. P stands for pass, F stands for fail, and U stands for undetermined.

Although the CFA Institute maintains an air of secrecy surrounding its pass/fail thresholds, drawing from feedback provided by CFA exam takers on Reddit suggests that, for Level I, in general scoring approximately 70% in a majority of sections appears to more often than not lead to a pass. Attaining scores above 70% in *all* topics is not a requirement for pass, but maintaining an average score of 70% across topics considerably enhances the likelihood of a positive outcome[3].

The estimates from the Reddit community regarding the MPS for Level II and III indicate that the two advanced exams have consistently featured lower passing thresholds. In June 2019, their approximation on the MPS for Level III was at a mere 57.4%, and 62.8% for Level II. The section passing scores are ambiguous for Level II, but we can attempt to apply the same logic as aforementioned Level I exam but make an assumption that the cutoff for each is 60% instead of 70%[4].

7.2 Pass Criteria and Outcomes

Given the information above, our proposed pass criteria are as follows:

- Level I - achieving a score of at least 60% in each topic and an overall score of at least 70%
- Level II - achieving a score of at least 50% in each topic and an overall score of at least 60%

Table 6 shows which model implementations were able to pass the exams. The FS implementations in both levels leverage the number of shots indicated in Table 2. Most of the settings showed a clear outcome, except for ZS on GPT-4 in Level II, which was a borderline case. ZS on GPT-4 attains >60% in six topics and a score between 50% and 60% in one topic. The topic performance seems high but the overall score, 57.39%, falls slightly short of the passing score proposed earlier, which therefore turns to be an unclear case.

8 Conclusion and Discussion

We conduct a thorough evaluation of ChatGPT and GPT-4 on the CFA exams and find that ChatGPT is unable to pass while GPT-4 is able under some FS and CoT settings. We note that CoT prompting provides marginal improvement for the models, but also exposes them to reasoning errors. Meanwhile, FS yields the best performance in most cases.

With these observations in mind, we propose future systems that could display greater performance by utilizing various tools. The most prevalent error mode of CoT, knowledge errors, could be addressed through retrieval-augmented generation using an external knowledge base containing CFA-specific information, or through fine-tuning on textbook data. Calculation errors could be avoided by offloading calculations to a function or API such as Wolfram Alpha. The remaining error modes, reasoning and inconsistency, could be reduced by employing a critic model to review and second guess the thinking before submitting the answer,

[3] https://www.efinancialcareers.com.au/news/finance/whats-the-minimum-score-you-can-get-on-cfa-level-i-and-still-pass

[4] https://www.efinancialcareers.com.au/news/finance/whats-the-minimum-score-you-can-get-on-cfa-level-i-and-still-pass/

or combining FS and CoT together to give richer examples of expected behavior. We hope this work paves the way for future studies to continue enhancing LLMs for financial reasoning problems through rigorous evaluation.

Acknowledgments

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 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

Samir Abdaljalil and Houda Bouamor. 2021. An exploration of automatic text summarization of financial reports. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 1–7.

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

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. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

TH Kung, M Cheatham, A Medenilla, C Sillos, L De Leon, C Elepaño, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. plos digit health 2 (2): e0000198.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks. *arXiv preprint arXiv:2305.05862*.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Eric Strong, Alicia DiGiammarino, Yingjie Weng, Preetha Basaviah, Poonam Hosamani, Andre Kumar, Andrew Nevins, John Kugler, Jason Hom, and Jonathan Chen. 2023. Performance of chatgpt on free-response, clinical reasoning exams. *medRxiv*, pages 2023–03.

Bin Wang, Jiangzhou Ju, Yunlin Mao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2022. A numerical reasoning question answering system with fine-grained retriever and the ensemble of multiple generators for finqa. *arXiv preprint arXiv:2206.08506*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Appendix

A Topic Distribution in Each Level

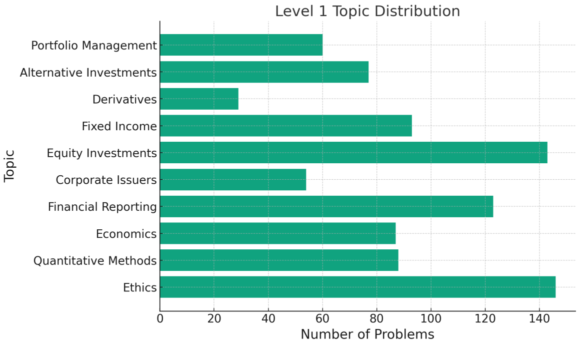

Figure 2: Level I exam topic distribution

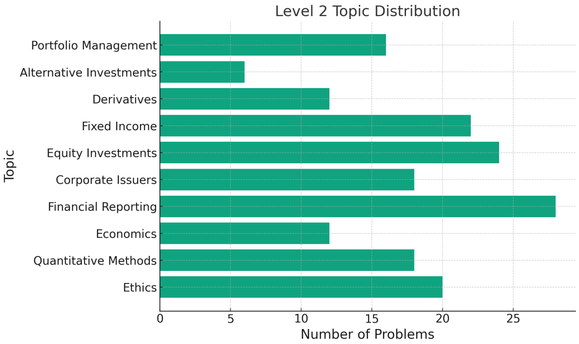

Figure 3: Level II exam topic distribution

B Prompt Templates Used in Our Work

B.1 Level I

Listing 1: ZS

```
SYSTEM: You are a CFA (chartered
    financial analyst) taking a test to
    evaluate your knowledge of finance.
    You will be given a question along
    with three possible answers (A, B,
    and C).

Indicate the correct answer (A, B, or
    C).

USER: Question:
{question}
A. {choice_a}
B. {choice_b}
C. {choice_c}
```

Listing 2: CoT

```
SYSTEM: You are a CFA (chartered
    financial analyst) taking a test to
    evaluate your knowledge of finance.
    You will be given a question along
    with three possible answers (A, B,
    and C).

Before answering, you should think
    through the question step-by-step.
    Explain your reasoning at each step
    towards answering the question. If
    calculation is required, do each
    step of the calculation as a step
    in your reasoning.

Indicate the correct answer (A, B, or
    C).

USER: Question:
{question}
A. {choice_a}
B. {choice_b}
C. {choice_c}
```

Listing 3: FS (2S example)

```
SYSTEM: You are a CFA (chartered
    financial analyst) taking a test to
    evaluate your knowledge of finance.
    You will be given a question along
    with three possible answers (A, B,
    and C).

Indicate the correct answer (A, B, or
    C).

USER: Question:
{question}
A. {choice_a}
B. {choice_b}
C. {choice_c}

ASSISTANT: {answer}

USER: Question:
{question}
A. {choice_a}
B. {choice_b}
C. {choice_c}

ASSISTANT: {answer}

USER: Question:
{question}
A. {choice_a}
B. {choice_b}
C. {choice_c}
```

B.2 Level II

For Level II, the case description of each item-set was inserted after the system prompt, before each question from the user.

C Related Work

LLMs and Finance. LLMs are trained on massive datasets that cover a broad range of topics and domains. Previous work has demonstrated the ability of LLMs to generalize surprisingly well to unseen downstream tasks, with little to no additional training data (Brown et al., 2020; Wei et al., 2022).

This raises an interesting question on the competitiveness of LLMs on specialized domains, such as finance. Indeed, the characteristics of most financial reasoning tasks — which rely on specific concepts and mathematical formula, frequently leverage diagrams and tables, often need multistep reasoning with calculations — make finance a challenging domain of application for LLMs. Several paths have been proposed to incorporate or emphasize domain-specific knowledge in LLMs: continued pre-training (Araci, 2019; Wu et al., 2023) and supervised fine-tuning on new data (Mosbach et al., 2023; Yang et al., 2023), retrieval augmented generation using a vector database of external knowledge (Lewis et al., 2020), etc. However, before considering such enhancements, only few papers have proceeded to extensively benchmark the out-of-the-box capabilities of newer instruction-tuned LLMs in finance (Li et al., 2023).

Evaluation of LLMs on Human Exams and other Benchmarks. Several previous studies have evaluated LLMs on various standard exams, such as United States medical licensing exam (Kung et al., 2023), free-response clinical reasoning exams (Strong et al., 2023), college-level scientific exams (Wang et al., 2023), and the Bar exam (Katz et al., 2023). The crucial contribution of these works is their analysis of the strengths and weaknesses of LLMs in realistic domain-specific settings, which guide subsequent research and practical use case resolutions.

For example, (Wang et al., 2023) evaluated `ChatGPT` and `GPT-4` on a collection of Physics, Chemistry, and Math problems, and then concluded that current LLMs do not deliver satisfactory performance in complex scientific reasoning yet to be reliably leveraged in practice. In contrast, (Bang et al., 2023) found that `ChatGPT` outperformed fine-tuned task-specific models on four different NLP tasks, thus suggesting `ChatGPT` could be directly applied to solve industry use cases. Our paper aims at delving into the assessment of the inner financial reasoning abilities of `ChatGPT` and `GPT-4`.