# The State of the Art of Large Language Models on Chartered Financial Analyst Exams

**Mahmoud Mahfouz**[*1], **Ethan Callanan**[*2], **Mathieu Sibue**[*1], **Antony Papadimitriou**[*1], **Zhiqiang Ma**[1], **Xiaomo Liu**[1], and **Xiaodan Zhu**[2]

[1]J.P. Morgan AI Research

[2]Queen's University

{mahmoud.a.mahfouz, mathieu.sibue, antony.papadimitriou, zhiqiang.ma, xiaomo.liu}@jpmchase.com

{e.callanan, xiaodan.zhu}@queensu.ca

[*]Equal Contribution

## Abstract

The Chartered Financial Analyst (CFA) program is one of the most widely recognized financial certifications globally. In this work, we test a variety of state-of-the-art large language models (LLMs) on mock CFA exams to provide an overview of their financial analysis capabilities using the same evaluation standards applied for human professionals. We benchmark five leading proprietary models and nine open-source models on all three levels of the CFA through challenging multiple-choice and essay questions. We find that flagship proprietary models perform relatively well and can solidly pass levels I and II exams, but fail at level III due to essay questions. Open-source models generally fall short of estimated passing scores, but still show strong performance considering their size, cost, and availability advantages. We also find that using textbook data helps bridge the gap between open-source and proprietary models to a certain extent, despite reduced gains in CFA levels II and III. By understanding the current financial analysis abilities of LLMs, we aim to guide practitioners on which models are best suited for enhancing automation in the financial industry.

**If a firm's long-run average cost of production increases by 15 percent as a result of an 8 percent increase in production the firm is most likely experiencing:**
A. economies of scale.
B. diseconomies of scale.
C. constant returns to scale.

(a) Level I sample MCQ

*"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 [...]"*

**The NAVPS for Bissorte REIT is closest to:**
A. $129.34. B. $130.43. C. $133.51.

(b) Level II sample MCQ

*"Algonquin Enterprises is a US company that recently raised a substantial quantity of cash from the sale of a redundant factory site and would like to use this cash to retire a set of debt liabilities [...] Three different portfolios of investment-grade corporate bonds, ranging in maturity from 3 years to 10 years, have been proposed for the duration matching approach [...]"*

**Identify and justify with two reasons which of the three portfolios (P, Q, or R) should be chosen if the duration matching strategy is adopted.**

(c) Level III sample essay question

Figure 1: Public CFA example questions (CFA Institute, 2024a; Kaplan Schweser, 2023); the vignette/case description appears in blue.

## 1 Introduction

With over 190,000 charterholders in 160 markets, the Chartered Financial Analyst (CFA) program (CFA Institute, 2024a) is amongst the most sought-after credentials for investment professionals, requiring over a thousand hours of preparation on average. CFA charterholders achieve one of the highest distinctions in investment management, possessing in-depth training in the core skills of investment strategy and high-level money management (Curry and Adams, 2022). Studies have shown that CFA training enhances job performance and productivity for financial analysts in financial firms (Shukla and Singh, 1994; De Franco and Zhou, 2009).

Correspondence to mahmoud.a.mahfouz@jpmchase.com

Given the rapid advancement of large language models (LLMs) (Vaswani et al., 2017; OpenAI, 2020, 2023; Anthropic, 2024) and their potential for automation, it has become fundamental to ensure such models meet the necessary standards for professional application and decision-making in finance. In this regard, benchmarking the capabilities of LLMs on CFA exams constitutes a crucial foray.

This paper provides the most comprehensive study to date on the performance of state-of-the-art LLMs, both open-source and proprietary, on CFA exams — aiming to give an overview of the landscape of the financial analysis capabilities of LLMs. We share our observations on advantages and limitations of their application. Our contributions are summarized as follows:



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1068–1082
November 12-16, 2024 

- We benchmark the performance of leading LLMs, including five proprietary and nine open-source, on mock CFA exams. We show that proprietary models constitute the state of the art and outperform their open-source counterparts, passing CFA exam levels I and II. They also perform well on multiple-choice questions (MCQs) at level III, but still cannot reach the professional level of essay writing. None of the models were able to pass level III.
- We provide a comprehensive investigation on the strengths and weaknesses of LLMs on each CFA level and across key financial topic areas, focusing on general patterns and comparing top proprietary and open source models.
- We examine the benefits of providing external theoretical knowledge to open-source LLMs by implementing a retrieval-augmented generation (RAG) pipeline using CFA textbooks. We find that RAG helps bridge the gap between closed and open source on certain levels of the CFA, but not all.

## 2 Background

Earning the CFA certification requires a bachelor's degree, three years of qualified work experience, and passing all CFA exam levels (CFA Institute, 2024a). The examination process is structured into three levels (I, II, III; see Table 1). It is designed to test: (1) the mastery of a range of financial concepts such as economics, financial reporting, and quantitative methods; (2) the ability to reason over situations with context; (3) the ability to conduct case analyses. CFA exams include both MCQs and essay questions, with levels I to III progressively increasing in difficulty and incorporating more real-world financial scenarios (CFA Institute, 2024a).

Level I of the CFA examination tests candidates' understanding of basic financial analysis across 10 topic areas (Table 1) using MCQs, as illustrated in Figure 1a. Therefore, it is generally considered the easiest level to pass. Level II transitions to vignette-based MCQs, requiring the application of investment tools and concepts in diverse contexts and the evaluation of asset classes, as depicted in Figure 1b. Level III differs by introducing essay questions that simulate professional scenarios, such as portfolio management decision-making and problem-solving (Figure 1c). Level III is assessed by tallying the total marks from MCQs (worth 3 points each) and the total marks from essay questions (points can vary) (CFA Institute, 2024b). The same grading process is followed in our research.

In summary, from level I to III, LLMs must progress from answering questions based on concept memorization and simple calculations to understanding context and reasoning, and finally to organizing thoughts in essay writing. Each level presents increasingly challenging tasks for AI.

## 3 Experimental Setup

**Dataset.** As official CFA exams are not public, we use CFA mock exams purchased from AnalystPrep (AnalystPrep, 2024) in this study, covering all three levels of the CFA program. The dataset includes both MCQs and essay questions, each accompanied with corresponding answers, explanations, grading details, as well as metadata such as the CFA topic each question belongs to. We use the set of mock exams of the year 2023, which corresponds to the 2023 CFA curriculum. Given that the mock exam data is secured behind a paywall, the risk of data contamination is reduced for LLMs. The distribution of question topics is shown in Table 1 (more details in Appendix A).

| Topic area | Level I | Level II | Level III |
|---|---|---|---|
| **Ethical Standards** | 16% | 11% | 9% |
| **Investment Tools** | 39% | 43% | 0% |
| Corporate Finance | 5% | 10% | - |
| Economics | 10% | 7% | - |
| Financial Reporting | 14% | 16% | - |
| Quantitative Methods | 10% | 10% | - |
| **Asset Classes** | 38% | 37% | 32% |
| Alternative Investments | 9% | 3% | - |
| Derivatives | 3% | 7% | - |
| Equity Investments | 16% | 14% | - |
| Fixed Income | 10% | 13% | - |
| **Portfolio Management** | 7% | 9% | 59% |
| **#Mock exams** | 5 | 2 | 2 |
| **#Questions per exam** | 180 | 88 | 44 |

Table 1: CFA mock exam topic areas and weights; Level III uses a different subtopic breakdown.

**LLM Models.** To perform a comprehensive study, we investigate a wide variety of LLMs as listed in Table 2. Specifically, the models highlighted in grey represent the state-of-the-art proprietary models (OpenAI, 2020, 2023; Open AI, 2024; Anthropic, 2024; Mistral AI, 2024). In contrast, open-source models (Jiang et al., 2024; Team et al., 2024; Meta, 2024; Cohere, 2024; Abdin et al., 2024; Groeneveld et al., 2024) provide more access to model details, are flexible for customization, and are often more cost-effective.

**Evaluation.** We implement an experimental setup designed to ensure consistency, fairness, and reproducibility across all tested models. Following recommendations from Callanan et al. (2023), each LLM is assessed using a one-shot learning setting, zero temperature, and prompted for chain-of-thought (CoT) reasoning (1S-CoT). Further details can be found in Appendix B.

To evaluate level I and II MCQs, we use the Accuracy metric. More precisely, to determine whether a model returns the correct answer to a question, we clean its CoT prediction by removing any reasoning from the output text using LLaMA 3 70B and only retain the final choice A, B, or C. To evaluate level III essay questions, we employ a model-assisted human evaluation strategy. We first prompt GPT-4o to perform marking by providing it with the ground-truth answers as well as the answer explanation and grading details from the mock exam data, which specify where and how to allocate marks. Then, a human CFA charter-holder verifies the generated scoring as demonstrated in Appendix G. The overall score for level III is the combination of the total marks from MCQs and essay questions according to the provided weighting.

To account for variation in the models' responses and a limited amount of data, each question is repeated five times with different seeds for selecting the one-shot example. We then calculate the mean score for each exam for each seed, and report the median of means. The costs for running our experiments are reported in Tables 9 and 10. We also perform ablation experiments (Appendix C) to study the effect of varying the number of examples and temperature, and a retrieval augmented generation (RAG) study in Section 4.3 to investigate the effect of incorporating external theoretical information.

# 4 Experiment Results & Analysis

## 4.1 Overall Performance

**Proprietary models constitute the state-of-the-art on CFA exam performance.** The results, shown in table 2, indicate a wide performance range across different LLMs on the CFA exams. Our results show that the leading proprietary models have the best overall performance, with GPT-4o showing the highest overall score on levels I and III, and Claude 3 Opus narrowly doing the best on level II.

**Mixtral and LLaMA 3 offer competitive alternatives while being smaller and often cheaper.** Of the open-source models, Mixtral-8x22B and LLaMA 3 70B perform the best. Both LLaMA 3 models do surprisingly well on all of the exams. Despite the far smaller size, the gap between LLaMA 3 70B and the leading proprietary models is only $\sim 20\%$ on each level, and while LLaMA 3 70B slightly underperforms Mixtral-8x22B, it is still within a few percentage points at roughly half the size. Furthermore, LLaMA 3 8B is able to outperform GPT-3.5 Turbo on MCQs from levels II and III. In comparison, OLMo 7B, an open-data and open-weights model, shows decent performance for its size on level I (despite a limited proportion of finance content in its training data), but falls short in levels II and III due to a reduced context length. Relative to the other open-source models, the LLaMA 3 models thus offer impressive financial reasoning capabilities for their parameter size class.

**All models struggle on level III essay questions.** These results yield surprising upsets compared to the level III MCQ results. While GPT-4o and GPT-4 Turbo still remain best-in-class, Claude 3 Opus struggles a lot more, performing on par with Mistral Large. In fact, the leading open source model Mixtral-8x22B outperforms its proprietary counterpart and Claude 3 Opus. Many models, such as OLMo 7B, simply do not have a large enough context length to answer the questions, or otherwise fail to provide an answer to the question. When models are able to answer, the ones that perform best are generally better at filtering the large context for only the most pertinent information. Worse performing models tend to recite too much and may come to the right answer but insufficiently explain their reasoning, or fail to interpret all the context and come to an outright incorrect conclusion.

**A major limitation for open-source models is their ability to catch nuance.** Although all models are given the exact same instructions for each question, we observe that the proprietary models are categorically better at following instructions exactly as presented compared to the open-source models. When prompted to "Think step by step and respond with your thinking and the correct

| Provider | Model | Parameters | Architecture | Level I | Level II | Level III | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | MCQ | Essay | Overall |
| OpenAI | GPT-3.5 Turbo | – | – | $63.8 \pm 1.1$ | $52.3 \pm 1.7$ | $44.2 \pm 6.0$ | $17.4 \pm 2.1$ | $31.4 \pm 2.2$ |
| | GPT-4 Turbo | – | – | $\underline{84.6} \pm 0.5$ | $\underline{76.7} \pm 0.7$ | $52.5 \pm 3.3$ | $\underline{42.4} \pm 4.4$ | $\underline{49.2} \pm 3.1$ |
| | GPT-4o | – | – | $\mathbf{88.1} \pm 0.3$ | $\underline{76.7} \pm 0.7$ | $\underline{63.4} \pm 4.2$ | $\mathbf{46.2} \pm 3.3$ | $\mathbf{55.0} \pm 2.8$ |
| Anthropic | Claude 3 Opus | – | – | $82.7 \pm 0.2$ | $\mathbf{77.8} \pm 2.9$ | $\mathbf{65.8} \pm 3.3$ | $6.8 \pm 1.4$ | $36.0 \pm 2.2$ |
| Mistral | Mixtral-8x7B | 46.7B | Mixture of Experts | $63.6 \pm 1.0$ | $49.4 \pm 0.8$ | $43.3 \pm 5.3$ | $18.9 \pm 1.3$ | $31.8 \pm 2.2$ |
| | Mixtral-8x22B | 141B | Mixture of Experts | $69.1 \pm 1.7$ | $61.4 \pm 1.4$ | $52.5 \pm 3.3$ | $28.8 \pm 2.9$ | $39.8 \pm 1.4$ |
| | Mistral Large | – | – | $69.0 \pm 1.4$ | $63.1 \pm 2.3$ | $47.5 \pm 5.5$ | $6.8 \pm 0.8$ | $28.0 \pm 2.8$ |
| Google | Gemma 2B | 2.5B | Decoder-only | $38.9 \pm 1.4$ | $35.2 \pm 2.4$ | $43.0 \pm 3.7$ | $6.1 \pm 1.0$ | $24.6 \pm 2.3$ |
| | Gemma 7B | 8.5B | Decoder-only | $46.0 \pm 1.7$ | $39.8 \pm 3.3$ | $43.3 \pm 6.2$ | $7.6 \pm 1.8$ | $24.2 \pm 3.8$ |
| Meta | LLaMA 3 8B | 8B | Decoder-only | $51.1 \pm 0.8$ | $54.0 \pm 1.8$ | $52.1 \pm 3.0$ | $12.9 \pm 2.2$ | $31.8 \pm 1.5$ |
| | LLaMA 3 70B | 69B | Decoder-only | $68.3 \pm 0.5$ | $58.0 \pm 1.2$ | $50.4 \pm 2.9$ | $18.9 \pm 2.2$ | $34.5 \pm 2.0$ |
| Cohere | Command R+ | 104B | Decoder-only | $51.8 \pm 1.9$ | $45.5 \pm 3.6$ | $35.4 \pm 4.7$ | $3.0 \pm 1.1$ | $18.2 \pm 2.4$ |
| Microsoft | Phi-3-mini | 3.8B | Decoder-only | $60.6 \pm 1.9$ | $27.3 \pm 4.8$ | $22.9 \pm 3.5$ | $1.5 \pm 2.6$ | $12.9 \pm 1.5$ |
| Ai2 | OLMo 7B | 6.9B | Decoder-only | $46.7 \pm 2.0$ | – | – | – | – |

Table 2: 1S-CoT overall accuracy (in percent) of different LLMs on CFA Level I, II & III questions. Essay questions are percentage of total marks. Proprietary LLMs are highlighted in grey, others are open source models. The bold font marks the best results in the corresponding columns and the underline marks the second best.

answer...", the larger proprietary models adhere to this exact format, starting with their chain of thought and concluding with their answer. In contrast, the open-source models are inconsistent and often begin by stating an answer before giving their reasoning. We believe this deviation impacts their overall performance, as they are not really using the CoT procedure to inform the answer but rather to justify it. Furthermore, it is indicative of an overall weaker capacity to follow instructions carefully, which may lead to misinterpretations or missing critical nuance in exam questions.

### 4.2 Performance by CFA Levels and Topics

**Level I.** Breaking the results down by topic on the level I exams (Figure 3) shows that performance is relatively uniform. The top proprietary models all score roughly the same across each of the topics. There is more variation in the open-source models, with the smaller models struggling more on topics that frequently require multi-step calculations such as Alternative Investments and Fixed Income. Overall, they perform best on Derivatives and Economics, for which questions are most often either simple one-step calculations or straightforward knowledge questions. A clear trend emerges where the smaller models are more prone to small mistakes that propagate when questions require multi-step calculation or reasoning.

**Level II.** On the more challenging level II exams, there is far more variation in performance across the topics (Figure 4). Each of the three top proprietary models (GPT-4 Turbo, GPT-4o, and Claude-3 Opus) is able to ace Portfolio Management, which is especially notable since these questions are meant to evaluate real-world financial analysis and decision making. However, they struggle a bit more in some of the knowledge-based topics like Ethics, Fixed Income, and Alternative Investments. In general, most models perform relatively well on Portfolio Management, making it one of the easier topics for LLMs on the level II exams. The open-source models perform well on Alternative Investments relative to their other scores, but tend to once again struggle on the complex math-heavy sections like Quantitative Methods and Financial Reporting & Analysis. Alongside compounding calculation errors, all models suffer to varying degrees from interpretation and knowledge application errors. As noticed looking at overall results, it is common for a model to state and correctly define a relevant concept, but then miss the nuance in applying it correctly to the situation at hand. The frequency of these issues is consistent with a model's overall performance, and exacerbated on questions in levels II and III with more complex question context.

**Level III.** Following the trend observed between level I and level II, the performance of each model across topics is far more varied in level III. Once again, the models surprisingly perform marginally better on the management-focused topics than the knowledge-based ones. These questions all require a deep understanding of financial concepts and a strong ability to apply them to a highly specific context, which was identified in the previous sections as a challenge for the LLMs. In general, due to the complexity of the case studies and the focus on evaluating real-world decision making in all topics, the difficulty is far less determined by the topic and more so by the question specifics.

**Model Comparison.** To further investigate the error modes and differences between models, we inspect questions that GPT-4o answered correctly across all five 1S-CoT seeds but other models got wrong in at least one seed. We particularly look at errors from the top proprietary competitor Claude 3 Opus and one of the top open-source competitors LLaMA 3 70B. A few trends are observed from math or numerical analysis topics such as Quantitative Methods, Financial Statement Analysis, Fixed Income, Alternative Investments, Derivatives and Equity. One of the most common differences between other wrong models and GPT-4o is simple calculation error — a well known limitation of LLMs (Frieder et al., 2023). In some CFA questions requiring multiple formulas with relatively complex terms, errors are compounded and then lead to incorrect final answers. Our results show LLaMA 3 70B is more prone to these simple calculation errors and often appears to randomly select one of the candidate answers and hallucinate it as the result of an equation. For the larger and "smarter" Claude 3 Opus model, its rarer errors on math questions more often result from incorrect application of key knowledge, leading to the wrong formula. For example, Claude 3 Opus might correctly calculate an intermediate result but fail to recognize additional steps implied by the question, leading to incorrect final answers.

To explore the differences between various LLMs' relative performance across the levels, we also compare Gemma 7B and LLaMA 3 70B. The Gemma models break the consistent pattern of decreasing scores as the level increases with outsized performance on level III MCQs, while LLaMA 3 70B is representative of the standard decrease in score at higher exam levels. The most evident correlation is in their respective handling of prompt length. By weighting the questions by prompt length (in tokens), LLaMA 3 70B's score on level III MCQs drops 3.1 percentage points from 50.4% down to 47.3%, while Gemma 7B drops less than a percent from 43.3% to 42.5%. This suggests that the Gemma models are better at handling longer prompts for their size than other models, in line with the emphasis put on long context performance in subsequent models from Google (Kilpatrick et al., 2024). Considering CFA exam questions tend to get longer and provide more context at higher levels, this might explain a majority of the discrepancy in performance observed. Other less pronounced differences in performance are more difficult to attribute, though we suspect they may come down to the presence and quality of related financial topics in the models' respective private training data.

### 4.3 Open Book Evaluation

Experiments in Sections 4.1 and 4.2 exclusively relied on the internal knowledge of LLMs and concrete question examples via 1S-CoT prompting. In this section, we measure the benefits of providing external theoretical financial knowledge by implementing a RAG pipeline. For this purpose, we leverage textbooks from the same provider as the mock exams. Each CFA Level has its own dedicated textbook, structured into chapters comprising multiple readings (or subchapters) — themselves composed of posts. Table 7 in Appendix D contains statistics about the textbooks. Due to the significant length of chapters and readings, we index the textbooks at the post-level for retrieval. Figure 2 in Appendix D shows a public example post. Each MCQ in the mock exams is already paired with a post from the textbooks discussing concepts that should help answer the question — which we refer to as the oracle post.

**Retrieval Experiments.** To first assess the difficulty of retrieving posts given an MCQ, we benchmark two retrievers using the oracle annotations. We select one popular lexical model, BM25+ (Robertson et al., 1994), and one competitive semantic model of moderate size, gte-large-en-v1.5 (gte) (Li et al., 2023b). We compute their Recall@K for $K \in \{1, 3, 5, 10, 50\}$ on MCQs from levels I, II, and III. Table 8 in Appendix D compiles results. We observe that the semantic model outperforms the lexical one on all levels, with wider margins in levels I and III. We also notice that Level III MCQs are harder to match to textbook passages, despite a smaller number of posts to choose from.

**Generation Experiments.** We leverage posts retrieved by BM25+, gte, as well as oracle anno-

| Model | Retriever | Level I | | | Level II | | | Level III | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | K=1 | K=3 | K=5 | K=1 | K=3 | K=5 | K=1 | K=3 | K=5 |
| LLaMA 3 8B | 1S-CoT | 51.1 | – | – | **54.0** | – | – | **52.1** | – | – |
| | oracle | 63.0 | – | – | 49.1 | – | – | 41.2 | – | – |
| | BM25+ | **63.5** | 59.4 | 60.6 | 50.3 | 45.5 | 48.9 | 39.0 | 42.5 | 41.9 |
| | gte-large-en-v1.5 | 63.0 | 60.9 | 58.0 | 52.8 | 40.6 | 49.7 | 46.0 | 47.9 | 41.9 |
| LLaMA 3 70B | 1S-CoT | 68.3 | – | – | 58.0 | – | – | 50.4 | – | – |
| | oracle | 77.6 | – | – | 61.4 | – | – | 51.5 | – | – |
| | BM25+ | 79.2 | 79.0 | 76.7 | **62.5** | 61.9 | 56.5 | 45.4 | 39.2 | **56.5** |
| | gte-large-en-v1.5 | 79.4 | 79.9 | **80.0** | 59.7 | 56.8 | 59.9 | 43.3 | 51.2 | 48.1 |

Table 3: End-to-end RAG results. Numbers reported are obtained by averaging two runs, one with the retrieval results ordered by relevance, and another with the results presented in the reverse order. The bold font marks the best results of each language model at the corresponding level and the underline marks the second best results.

tations to augment the generation of two LLMs: LLaMA 3 8B and LLaMA 3 70B.[1] In order to understand the influence of LLM size as well as the influence of the quality, quantity, and ordering of the retrieved passages, we run a total of 28 trials. Each trial features a unique combination of the following parameters:

- retriever $\in$ {oracle, BM25+, gte};
- $\text{K} \in \{1, 3, 5\}$, which designates the number of retrieved passages fed to the LLM;[2]
- order $\in \{\text{relevance}, \text{relevance}_{\text{reversed}}\}$, used to order passages and average predictions;
- reader $\in$ {LLaMA 3 8B, LLaMA 3 70B}.

Table 3 shows the end-to-end RAG results across all CFA levels. We first observe that RAG mainly benefits Level I exams, with more modest gains in levels II and III. This could be due to the increased abstraction required in vignette-based MCQs and the challenge for LLMs to apply theoretical knowledge contextually.

Additionally, providing the oracle post to the reader does not yield perfect accuracy, suggesting that answers are not easily found in textbook posts. Interestingly, passages retrieved by BM25+ and gte sometimes outperform the oracle post. While counterintuitive, this can be explained by the fact that the LLaMA 3 models are prompted to think step by step in the RAG experiments; it is possible that certain posts better steer the reasoning of the LLMs than the oracle. Similarly, the retrieval performance advantage of gte over BM25+ does not consistently lead to higher MCQ accuracy.

[1]We pick the LLaMA 3 models because of their popularity and room for improvement on the CFA exams in 1S-CoT.

[2]K is fixed to 1 when retriever = oracle and capped to 5 due to the length of textbook posts and to the limited context window of LLaMA 3 models.

Finally, RAG helps reduce the gap between open source and proprietary LLMs. Indeed, with just $\text{K} = 5$ passages from gte, LLaMA 3 70B achieves 97% of Claude 3 Opus's performance in Level I. Nonetheless, it seems that LLaMA 3 8B benefits less from textbook data than its larger variant. While Table 3 shows that, for each CFA level, at least one LLaMA 3 70B RAG configuration surpasses 1S-CoT, LLaMA 3 8B RAG is outperformed by 1S-CoT in levels II and III – with no advantage gained from retrieving more passages. This suggests that larger models have an edge in understanding and applying theoretical financial knowledge in context.

## 4.4 LLMs as Certified CFA Professionals?

**No model successfully passes all three levels of the examinations.** The CFA Institute does not disclose the official Minimum Passing Score (MPS), which varies from exam to exam. According to estimates (Kaplan Schweser, 2024), the MPS ranges between a lower bound of 60% and an upper bound of 70%. Based on these thresholds, GPT 4 models and Claude 3 Opus passed levels I and II in both lower and upper bounds. The open-source model LLaMa 3 70B with the help of open book setting (RAG) can pass levels I and II using the lower bound score. None of the models can reliably pass level III to obtain the CFA certification, as there is still a significant gap between LLMs and professionals in essay writing. The best performing GPT-4o received 46.2 in essay score and thus brought down the overall level III score to 55.0. A limitation is that our essay grading method is not exactly the same as actual grading. The complete pass/fail comparison is provided in Table 4.

| Provider | Model | Level I | | Level II | | Level III | |
|---|---|---|---|---|---|---|---|
| | | L | U | L | U | L | U |
| OpenAI | `GPT-3.5 Turbo` | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `GPT-4 Turbo` | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| | `GPT-4o` | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| Anthropic | `Claude 3 Opus` | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| Mistral | `Mixtral-8x7B` | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `Mixtral-8x22B` | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| | `Mistral Large` | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| Google | `Gemma 2B` | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `Gemma 7B` | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Meta | `LLaMA 3 8B` | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `LLaMA 3 70B` | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `LLaMA 3 8B + RAG` | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| | `LLaMA 3 70B + RAG` | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| Cohere | `Command R+` | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Microsoft | `Phi-3-mini` | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Ai2 | `OLMo 7B` | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |

Table 4: LLMs' ability to pass each CFA level using `1S-CoT` or RAG, with the lower bound score L ($\geq 60\%$) and upper bound score U ($\geq 70\%$). ✓ indicates the LLM should pass the exam according to the corresponding bound, while ✗ indicates it should fail.

## 5 Related Work

**LLMs for Finance.** As highlighted by 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 knowledge, poses significant challenges. Li et al. (2023a) has shown that generalist LLMs like `ChatGPT` are able to reach excellent performance on simple financial NLP tasks like sentiment analysis, but still cannot outcompete professionals on more complex tasks requiring math computation and financial knowledge like question answering. Enhancement 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 (Lewis et al., 2020) have been proposed to use domain-specific knowledge from other sources to address these challenges.

**LLMs on Professional Exams.** Recent work (Callanan et al., 2023) has started to study CFA but is inherently limited by *only* evaluating on two models, `ChatGPT` and `GPT-4`, and *only* on MCQs from levels I and II — thus lacking a complete view of the state of the art of LLMs on the entirety of the CFA program. There also emerges various studies of scrutinizing LLMs in other professional exams such as the United States medical licensing exam (Kung et al., 2023), free-text response clinical reasoning exams (Strong et al., 2023), college-level scientific exams (Wang et al., 2023), and the Bar exam (Katz et al., 2023). Benchmarking LLMs on professional exams plays a fundamental role to understand the advances of AI in various areas.

## 6 Conclusion

In this paper, we benchmark the performance of 14 LLMs on the CFA exams, revealing that closed-source models like `GPT-4o` and `Claude 3 Opus` consistently outperform their open-source counterparts. These models not only demonstrated superior accuracy across all three CFA levels, but also highlighted the importance of model architecture and training data quality over sheer size. Our detailed analysis of topic-wise performance and error modes underscores the complexities LLMs face in financial tasks, particularly in math-heavy sections. This research advances our understanding of LLM capabilities in high-stakes financial environments and identifies areas for improvement in their application to domain-specific challenges. We hope this work will serve as a point of reference for the evaluation of future models as steps forward are made, and hope the insights will inform future work developing financial domain-specific models.


## 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


Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AnalystPrep. 2024. Study materials for cfa®, frm®, actuarial, and mba admission exams. https://www.analystprep.com [Accessed: 18 July 2024].

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

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

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.

Ethan Callanan, A. Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *ArXiv*, abs/2310.08678.

CFA Institute. 2024a. Cfa institute. https://www.cfainstitute.org [Accessed: 18 July 2024].

CFA Institute. 2024b. Level iii cfa exam structure. https://www.cfainstitute.org/en/programs/cfa/exam/level-iii [Accessed: 18 July 2024].

Cohere. 2024. Introducing command r+: A scalable llm built for business. *Cohere*.

Benjamin Curry and Michael Adams. 2022. Chartered financial analysts are the rock stars of finance. *Forbes*.

Gus De Franco and Yibin Zhou. 2009. The performance of analysts with a cfa® designation: The role of human-capital and signaling theories. *The Accounting Review*, 84(2):383–404.

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

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and other. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kaplan Schweser. 2023. Cfa exam sample questions. https://www.schweser.com/cfa/blog/how-to-pass-the-cfa-exam/cfa-exam-sample-questions [Accessed: 18 July 2024].

Kaplan Schweser. 2024. Understand cfa® scores: The grading process. https://www.schweser.com/cfa/blog/how-to-pass-the-cfa-exam/cfa-exam-grading [Accessed: 18 July 2024].

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

Logan Kilpatrick, Shrestha Basu Mallick, and Kofman Ronen. 2024. Gemini 1.5 pro 2m context window, code execution capabilities, and gemma 2 are available today. https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio [Accessed: 09 September 2024].

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.

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, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023a. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 408–422.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available gllm to date. *Meta AI*.

Mistral AI. 2024. Mistral Large Blog. https://mistral.ai/news/mistral-large/.

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*.

Open AI. 2024. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf.

OpenAI. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC-3*, pages 0–.

Ravi Shukla and Sandeep Singh. 1994. Are cfa charterholders better equity fund managers? *Financial Analysts Journal*, 50(6):68–74.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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 Dataset Details

For CFA Level I, the dataset includes five mock exams, each consisting of 180 multiple-choice questions. These questions cover a range of topics, including quantitative methods, economics, and portfolio management. The Level II dataset comprises two mock exams, each featuring 22 item sets with four multiple-choice questions per set, based on detailed vignettes, resulting in a total of 176 questions. These questions address topics such as financial reporting & analysis, fixed income securities, and alternative investments. Finally, for CFA Level III, the dataset includes two mock exams, each containing a mix of item sets and essay questions, totaling 88 questions. Topics for Level III exams span areas like derivatives & currency management, capital markets, and wealth management.

### B Evaluation Details

We chose to use one-shot learning for our main experiments instead of few-shot as some of the models have smaller context windows that would not fit many CoT examples. In addition, it was previously found that increasing the number of shots does not appear to have a large impact on performance (Callanan et al., 2023) – though we investigate this in Appendix C.

During experiments, each model is presented with a single example question along with the correct answer and explanation of the reasoning from a question bank. The prompt then asks the model to solve a different question from the mock exams. The example question is selected to ensure it covers the same topic and is not part of the mock exams utilized for evaluation. We repeat each question with five different examples to account for variation in model responses. To get overall scores for each model, we compute the mean score for each exam, then take the median of means as the score for the model on that exam level. We also report the standard deviation of scores across the five attempts to capture the variability in model performance. Experiment costs are reported in Tables 9 and 10.

### C Ablations

Tables 5 and 6 show the overall performance across levels I, II and III with different numbers of shots and temperatures respectively. We choose two proprietary models (`GPT-4 Turbo`, `GPT-4o`) and two open-source models (`LLaMA 3 8B`, `LLaMA 3 70B`) for our ablations. We observe that increasing the number of shots generally has a mixed impact on the performance of the language models evaluated. For most models, there is a slight decrease in performance as the number of shots increases, as noted in Section 4.3. This suggests that providing more examples does not necessarily improve model performance and, in some cases, may even slightly hinder it, possibly due to the model becoming overwhelmed or distracted by too much context. As for increasing the temperature, we also observe that it results in a slight decrease in the performance of the language models on the CFA tasks. This indicates that higher temperatures, which introduce more randomness in model responses, can negatively affect the accuracy and consistency of the models' outputs in the context of CFA exam tasks.

| Model | Level I | | | Level II | | | Level III | | |
|---|---|---|---|---|---|---|---|---|---|
| | K=1 | K=2 | K=5 | K=1 | K=2 | K=5 | K=1 | K=2 | K=5 |
| GPT-4 Turbo | **84.6** | 82.3 | 82.1 | **76.7** | 72.8 | 68.2 | **55.4** | 51.9 | 50.2 |
| GPT-4o | 88.1 | **88.3** | 86.9 | **76.7** | 76.2 | 73.9 | 67.9 | **71.4** | 67.9 |
| LLaMA 3 8B | 51.1 | 55.0 | **56.9** | **54.0** | 48.9 | 41.5 | **44.6** | **44.6** | 41.0 |
| LLaMA 3 70B | 68.3 | 72.4 | **74.3** | **58.0** | 52.4 | 45.3 | 48.4 | **48.5** | 44.3 |

Table 5: Overall Performance with different numbers of shots K for CFA levels I, II, and III

| Model | Level I | | | Level II | | | Level III | | |
|---|---|---|---|---|---|---|---|---|---|
| | T=0 | T=0.7 | T=1 | T=0 | T=0.7 | T=1 | T=0 | T=0.7 | T=1 |
| GPT-4 Turbo | **84.6** | 84.1 | 84.0 | **76.7** | 73.9 | 73.3 | 55.4 | **59.0** | 53.7 |
| GPT-4o | **88.1** | **88.1** | 86.9 | 76.7 | **77.3** | 74.5 | 67.9 | **75.0** | 71.4 |
| LLaMA 3 8B | **51.1** | 46.0 | 45.0 | **54.0** | 50.6 | 46.6 | **44.6** | **44.6** | 41.0 |
| LLaMA 3 70B | **68.3** | 63.2 | 62.2 | **58.0** | 54.6 | 50.6 | **48.2** | **48.2** | 44.6 |

Table 6: Overall Performance with different temperatures T for CFA levels I, II, and III

### D RAG details

Table 7 shows textbook data characteristics and Table 8 passage retrieval results. Figure 2 shows a public example post from the level I textbook.

| Section | Level I | | Level II | | Level III | |
|---|---|---|---|---|---|---|
| | Count | Length | Count | Length | Count | Length |
| Chapter | 10 | 51 710 | 10 | 50 243 | 11 | 26 734 |
| Reading | 73 | 7 084 | 46 | 10 922 | 33 | 8 911 |
| Post | 572 | 904 | 409 | 1 228 | 252 | 1 167 |

Table 7: Textbook data characteristics: number of passages and average passage length per section type (in number of tokens returned by the `LLaMA 3` tokenizer).

| Level | Model | Recall | | | | |
|---|---|---|---|---|---|---|
| | | @1 | @3 | @5 | @10 | @50 |
| I | BM25+ | 34.7 | 48.7 | 55.1 | 63.7 | 84.3 |
| | gte | **40.9** | **59.6** | **66.3** | **73.6** | **90.5** |
| II | BM25+ | 22.7 | 39.3 | 44.7 | 54.7 | **77.3** |
| | gte | **24.7** | **43.3** | **51.3** | **60.7** | **77.3** |
| III | BM25+ | 12.5 | 22.5 | 32.5 | 47.5 | 72.5 |
| | gte | **17.5** | **35.0** | **40.0** | **57.5** | **80.0** |

Table 8: Passage retrieval results.

### Net Present Value (NPV)

The net present value (NPV) of a project is the potential change in wealth resulting from the project after accounting for the time value of money. The NPV for a project with one investment outlay made at the start of the project is defined as the present value of the future after-tax cash flows minus the investment outlay.

$$\text{NPV} = \sum_{t=1}^{n} \frac{CF_t}{(1+r)^t} - \text{Outlay}$$

Where:

$CF_t$ = After-tax cash flow at time $t$

$r$ = Required rate of return for the investment

Outlay = Investment cash flow at time zero

Many projects have cash flow patterns in which outflows occur not only at the start of the project (at time = 0) but also at future dates. In these instances, a better formula to use is:

- to invest in the project if NPV > 0;
- not to invest in the project if NPV < 0; and
- stay indifferent if NPV = 0.

Figure 2: Public level I textbook post excerpt from https://analystprep.com/cfa-level-1-study-notes/ (AnalystPrep, 2024).

## E Performance by Topic

Figures 3, 4, and 5 show the detailed breakdown of the performance by topics across levels I, II and III respectively. The full analysis of the results is outlined in Section 4.2 in the paper.

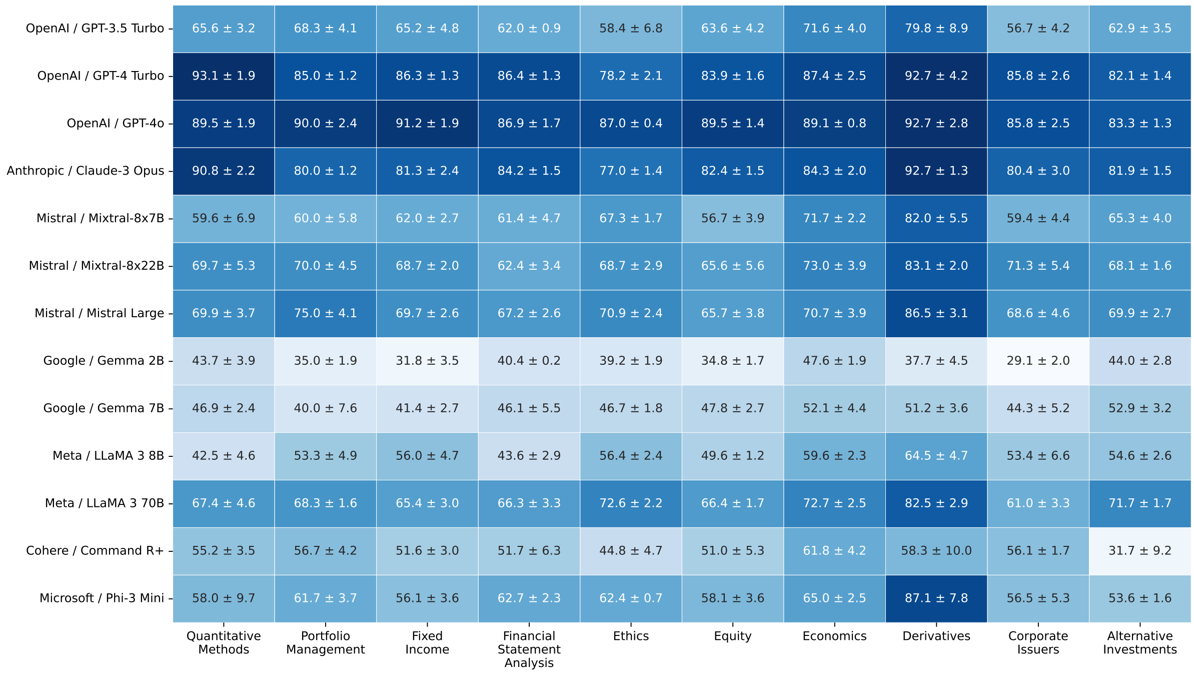


Figure 3: `1S-CoT` accuracy (in percent) of different LLMs on CFA **Level I** broken down by topics (Quantitative Methods, Portfolio Management, Fixed Income, Financial Statement Analysis, Ethics, Equity, Economics, Derivatives, Corporate Issuers, and Alternative Investments)

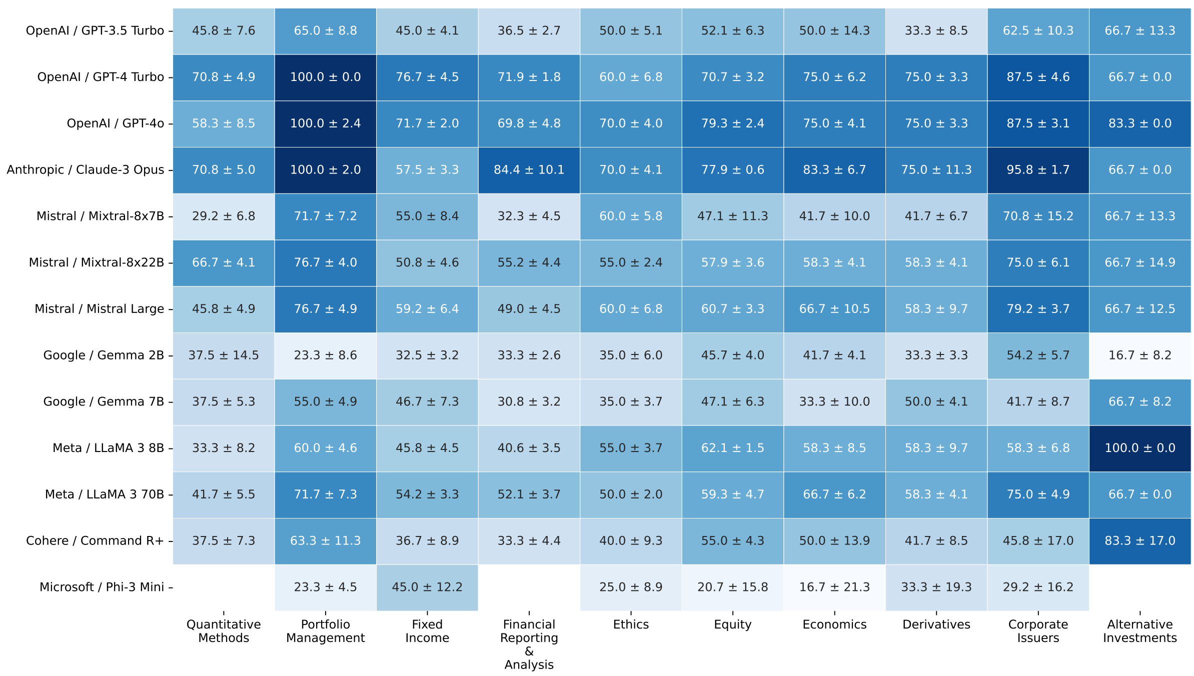


Figure 4: `1S-CoT` accuracy (in percent) of different LLMs on CFA **Level II** broken down by topics (Quantitative Methods, Portfolio Management, Fixed Income, Financial Reporting & Analysis, Ethics, Equity, Economics, Derivatives, Corporate Issuers, and Alternative Investments)

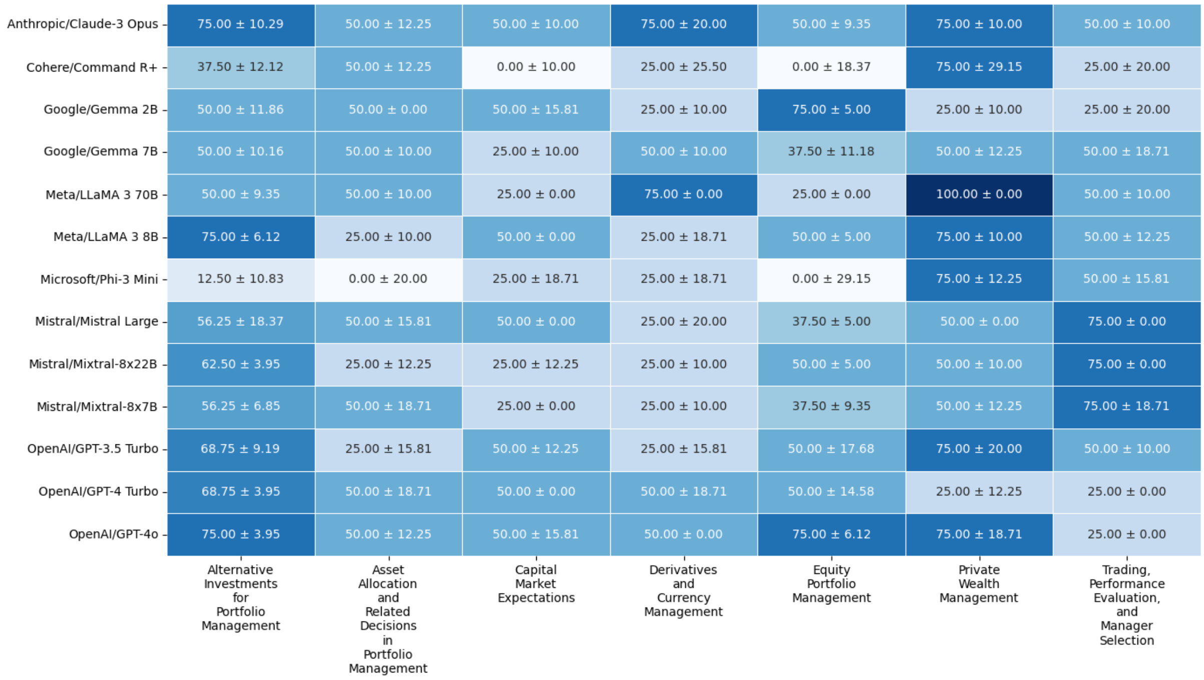


Figure 5: `1S-CoT` accuracy (in percent) of different LLMs on CFA **Level III** broken down by topics (Alternative Investments for Portfolio Management, Asset Allocation and Related Decisions in Portfolio Management, Capital Market Expectations, Derivatives and Currency Management, Equity Portfolio Management, Private Wealth Management, and Trading, Performance Evaluation, and Manager Selection)

| **Provider** | **Model** | **Tokens** | | **Cost per Token (¢)** | | **Cost ($)** | | |
|---|---|---|---|---|---|---|---|---|
| | | **Prompt Tokens** | **Completion Tokens** | **Prompt Cost** | **Completion Cost** | **Input** | **Output** | **Total** |
| OpenAI | `GPT 3.5 Turbo` | 5,207,711 | 1,166,090 | 0.0002 | 0.0002 | 10.42 | 2.33 | 12.75 |
| | `GPT 4 Turbo` | 5,207,711 | 1,665,269 | 0.001 | 0.003 | 52.08 | 49.96 | 102.03 |
| | `GPT-4o` | 5,207,711 | 1,826,928 | 0.0005 | 0.0015 | 26.04 | 27.40 | 53.44 |
| Anthropic | `Claude 3 Opus` | 5,207,711 | 1,773,782 | 0.0015 | 0.0075 | 78.12 | 133.03 | 211.15 |
| Mistral | `Mistral Large` | 5,207,711 | 1,547,536 | 0.0003 | 0.0009 | 15.62 | 13.93 | 29.55 |

Table 9: Proprietary models prompt and completion costs amounting to $408.9 in total. Note that inference costs from closed source providers are subject to change over time

| **Provider** | **Model** | **Inference Time (hours)** | **GPUs** | **Cost per Hour ($)** | **Total Cost ($)** |
|---|---|---|---|---|---|
| Mistral | `Mixtral-8x7B` | 6.99 | 2x Nvidia A100 | 8.0 | 55.93 |
| | `Mixtral-8x22B` | 12.05 | 4x Nvidia A100 | 16.0 | 192.75 |
| Google | `Gemma 2B` | 1.64 | 1x Nvidia L4 | 0.8 | 1.31 |
| | `Gemma 7B` | 2.30 | 1x Nvidia L4 | 0.8 | 1.84 |
| Meta | `LlaMA 3 8B` | 5.95 | 1x Nvidia L4 | 0.8 | 4.76 |
| | `Llama 3 70B` | 25.88 | 4x Nvidia A100 | 16.0 | 414.13 |
| Cohere | `Command R+` | 11.02 | 4x Nvidia A100 | 16.0 | 176.26 |
| Microsoft | `Phi-3-mini` | 3.10 | 1x Nvidia L4 | 0.8 | 2.481 |

Table 10: Open Source Models by Provider, Inference Time, GPUs, and Cost amounting to $849.5 in total. Note that external serverless LLM API providers could have been used to reduce inference costs

## F Prompt templates

Listing 1: Level I Prompt Template

```
SYSTEM: You are taking a test
   for the Chartered Financial
   Analyst (CFA) program
   designed to evaluate your
   knowledge of different topics
   in finance.
You will be given a question
   along with three possible
   answers (A, B, and C). Think
   step by step and respond with
   your thinking and the correct
   answer (A, B, or C) between
   square brackets.

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

Listing 2: Level II Prompt Template

```
SYSTEM: You are taking a test
   for the Chartered Financial
   Analyst (CFA) program
   designed to evaluate your
   knowledge of different topics
   in finance.
You will be given a question
   along with three possible
   answers (A, B, and C). Think
   step by step and respond with
   your thinking and the correct
   answer (A, B, or C) between
   square brackets.

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

Listing 3: Level III Prompt Template

```
SYSTEM: You are taking a test
   for the Chartered Financial
   Analyst (CFA) program
   designed to evaluate your
   knowledge of different topics
   in finance.
You will be given an open ended
   essay question. Think step by
   step and respond with your
   thinking and answer the
   question.

USER: Case:
{case}
Question:
{question}
Answer:
```

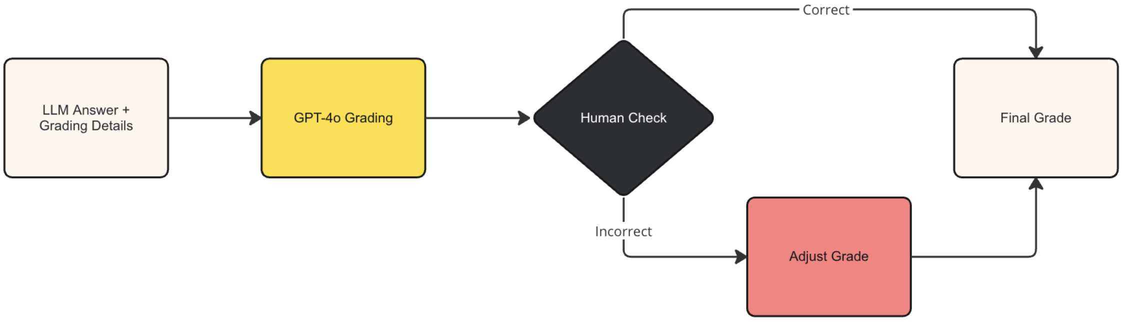


Figure 6: Level III Essay Grading Process

# G Level III Essay Grading

Listing 4: Level III Essay Grading

```
SYSTEM: You are tasked with
   grading essay answers from
   the CFA Level 3 examination.

You will be supplied with an
   explanation of the correct
   answer, the grading details
   (where to assign marks) and
   the student's answer.

Return a numeric value
   indicating the number of
   marks the student should
   receive and the explanation
   as to why the student did/did
   not receive the marks outline
   in the grading detail.

USER: Here are the answer
   grading details:
{answer_grading_details}

USER: Here is the student's
   answer:
{answer}
```