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] **JPMorgan AI Research**, [2] **Queen's University**

The Chartered Financial Analyst (CFA) Exam

- The CFA exam is a globally recognized financial certification for investment professionals.
- It requires, on average, over a thousand hours of preparation and is known for its rigorous examination process.
- The CFA exam is divided into three levels, each increasing in difficulty and complexity.
  - **Level I**: focuses on basic financial analysis through multiple-choice questions covering 10 topic areas
  - **Level II**: involves vignette-based multiple-choice questions, requiring application of investment tools and concepts.
  - **Level III**: includes essay questions that simulate real-world financial scenarios for portfolio management and financial decision making
- The exams test mastery of financial concepts, contextual reasoning and case analysis skills.

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.

"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.

"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.

Figure 1: Public CFA example questions for **Level I** (top left), **Level II** (top right) and **Level III** (bottom).

Our Dataset

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 area breakdown

Overall Performance

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

Table 2: 1S-CoT overall accuracy (in percent) of different LLMs on CFA Level I, II and 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

Performance Breakdown by Topic

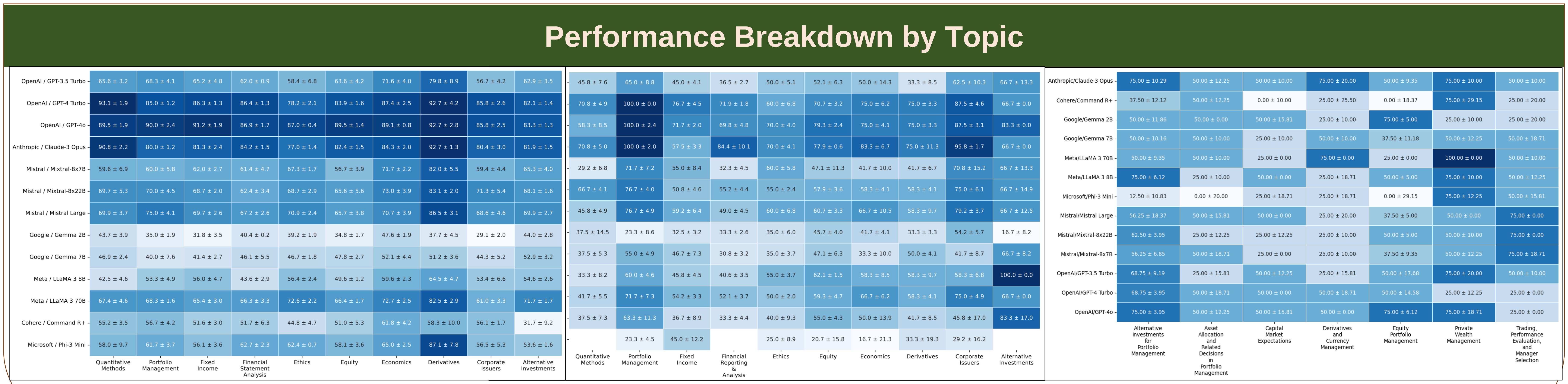

Figure 2: Performance breakdown by topic for **Level I** (left), **Level II** (middle) and **Level III** (right).

LLMs as CFA Professionals?

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 3: LLM's ability to pass each CFA level using 1S-CoT or RAG, with the lower bound score L (≥ 60%) and upper bound score U (≥ 70%).

✓ indicates the LLM should pass the exam while ✗ indicates it should fail.

TL;DR

- Proprietary models, such as GPT-4o and Claude 3 Opus, excel in CFA exam performance, particularly in Levels I and II.
- Mixtral and LLaMA 3 models offer competitive alternatives, balancing performance with reduced size and cost.
- Level III essay questions pose significant challenges for all models, highlighting limitations in complex reasoning and writing.
- Open-source models often miss nuanced details, impacting their overall accuracy and reliability.
- Despite advancements, no model has successfully passed all three levels of the CFA exam, indicating room for improvement in financial analysis capabilities.

* 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. 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.