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

Mahmoud Mahfouz<sup>\*1</sup>, Ethan Callanan<sup>\*2</sup>, Mathieu Sibue<sup>\*1</sup>, Antony Papadimitriou<sup>\*1</sup>, Zhiqiang Ma<sup>1</sup>, Xiaomo Liu<sup>1</sup>, and Xiaodan Zhu<sup>2</sup>

## <sup>1</sup> JPMorgan AI Research, <sup>2</sup> Queen's University



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.

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 D	Datas	set			Overall Performance										
Topic area	Level I	Level II	Level III	Provider	Model	Parameters	Architecture	Level I	Level II		Level III	0 11	Table 2: 1S-CoT overall accuracy (in		
Ethical Standards	16%	11%	9%							MCQ	Essay	Overall	percent) of different LLMs on CFA Level		
Investment Tools	39%	43%	0%	OpenAI	GPT-3.5 Turbo	-	-	$63.8 \pm 1.1$	$52.3 \pm 1.7$	$44.2 \pm 6.0$	$17.4 \pm 2.1$	$31.4 \pm 2.2$	I, II and III		
Corporate Finance	$\frac{39\%}{5\%}$	43% 10%	070	Open ii	GPT-4 Turbo	—	—	$\underline{84.6} \pm 0.5$	$76.7 \pm 0.7$	$52.5 \pm 3.3$	$\underline{42.4} \pm 4.4$	$49.2 \pm 3.1$	questions.		
Economics	10%	7%	-		GPT-40	_	—	$88.1 \pm 0.3$	$76.7 \pm 0.7$	$63.4 \pm 4.2$	$46.2 \pm 3.3$	$55.0 \pm 2.8$			
Financial Reporting	14%	16%	-	Anthropic	c Claude 3 Opus	—	_	$82.7\pm0.2$	$77.8 \pm 2.9$	$65.8 \pm 3.3$	$6.8\pm1.4$	$36.0\pm2.2$	<ul> <li>Essay questions are percentage of</li> </ul>		
Quantitative Methods	10%	10%	-		Mixtral-8x7B	46.7 <b>B</b>	Mixture of Experts	$63.6 \pm 1.0$	$49.4\pm0.8$	$43.3\pm5.3$	$18.9 \pm 1.3$	$31.8 \pm 2.2$	total marks.		
Asset Classes	38%	37%	32%	Mistral	Mixtral-8x22B	141 <b>B</b>	Mixture of Experts	$69.1 \pm 1.7$	$61.4 \pm 1.4$	$52.5 \pm 3.3$	$28.8 \pm 2.9$	$39.8 \pm 1.4$	<ul> <li>Proprietary LLMs are highlighted in</li> </ul>		
Alternative Investments Derivatives	${9\% \over 3\%}$	${3\% \over 7\%}$	-		Mistral Large	-	-	$69.0 \pm 1.4$	$63.1 \pm 2.3$	$47.5\pm5.5$	$6.8 \pm 0.8$	$28.0 \pm 2.8$	grey, others are		
Equity Investments	16%	14%	-	Casala	Gemma 2B	2.5B	Decoder-only	$38.9 \pm 1.4$	$35.2 \pm 2.4$	$43.0 \pm 3.7$	$6.1 \pm 1.0$	$24.6 \pm 2.3$	open source models.		
Fixed Income	10%	13%	-	Google	Gemma 7B	8.5B	Decoder-only	$46.0 \pm 1.7$	$39.8 \pm 3.3$	$43.3 \pm 6.2$	$7.6 \pm 1.8$	$24.2 \pm 3.8$	- The bold font		
Portfolio Management	7%	9%	59%	Mata	LLaMA 3 8B	8B	Decoder-only	$51.1 \pm 0.8$	$54.0 \pm 1.8$	$52.1 \pm 3.0$	$12.9 \pm 2.2$	$31.8 \pm 1.5$	marks the best results in the		
#Mock exams	5	2	2	Meta	LLaMA 3 70B	69 <b>B</b>	Decoder-only	$68.3 \pm 0.5$	$58.0 \pm 1.2$		$18.9 \pm 2.2$	$34.5 \pm 2.0$	corresponding columns and the		
#Questions per exam	180	88	44	Cohere	Command R+	104 <b>B</b>	Decoder-only	$51.8 \pm 1.9$	$45.5\pm3.6$	$35.4 \pm 4.7$	$3.0 \pm 1.1$	$18.2 \pm 2.4$	underline marks the second best		
"Questions per exam	100			Microsof	t Phi-3-mini	3.8B	Decoder-only	$60.6 \pm 1.9$	$27.3 \pm 4.8$	$22.9 \pm 3.5$	$1.5 \pm 2.6$	$12.9 \pm 1.5$			
Table 1: CFA mock example Level III uses a different				Ai2	OLMo 7B	6.9 <b>B</b>	Decoder-only	$46.7\pm2.0$	_	_	_	_			

## **Performance Breakdown by Topic**

OpenAl / GPT-3.5 Turbo	- 65.6 ± 3.2	68.3 ± 4.1	65.2 ± 4.8	62.0 ± 0.9	58.4 ± 6.8	63.6 ± 4.2	71.6 ± 4.0	79.8 ± 8.9	56.7 ± 4.2	62.9 ± 3.5	- 45.8 ± 7.6	65.0 ± 8.8	45.0 ± 4.1	36.5 ± 2.7	50.0 ± 5.1	52.1 ± 6.3	50.0 ± 14.3	33.3 ± 8.5	62.5 ± 10.3	66.7 ± 13.3	Anthropic/Claude-3 Opus -	75.00 ± 10.29	50.00 ± 12.25	50.00 ± 10.00	75.00 ± 20.00	50.00 ± 9.35	75.00 ± 10.00	50.00 ± 10.00
OpenAl / GPT-4 Turbo	- 93.1 ± 1.9	85.0 ± 1.2	86.3 ± 1.3	86.4 ± 1.3	78.2 ± 2.1	83.9 ± 1.6	87.4 ± 2.5	92.7 ± 4.2	85.8 ± 2.6	82.1 ± 1.4	- 70.8 ± 4.9	$100.0 \pm 0.0$	76.7 ± 4.5	71.9 ± 1.8	60.0 ± 6.8	70.7 ± 3.2	75.0 ± 6.2	75.0 ± 3.3	87.5 ± 4.6	66.7 ± 0.0	Cohere/Command R+ -	37.50 ± 12.12	50.00 ± 12.25	$0.00 \pm 10.00$	25.00 ± 25.50	0.00 ± 18.37	75.00 ± 29.15	25.00 ± 20.00
OpenAl / GPT-4o	- 89.5 ± 1.9	90.0 ± 2.4	91.2 ± 1.9	86.9 ± 1.7	87.0 ± 0.4	89.5 ± 1.4	89.1 ± 0.8	92.7 ± 2.8	85.8 ± 2.5	83.3 ± 1.3	- 58.3 ± 8.5	100.0 ± 2.4	71.7 ± 2.0	69.8 ± 4.8	70.0 ± 4.0	79.3 ± 2.4	75.0 ± 4.1	75.0 ± 3.3	87.5 ± 3.1	83.3 ± 0.0	Google/Gemma 2B -	50.00 ± 11.86	50.00 ± 0.00	50.00 ± 15.81	25.00 ± 10.00	75.00 ± 5.00	25.00 ± 10.00	25.00 ± 20.00
Anthropic / Claude-3 Opus	- 90.8 ± 2.2	80.0 ± 1.2	81.3 ± 2.4	84.2 ± 1.5	77.0 ± 1.4	82.4 ± 1.5	84.3 ± 2.0	92.7 ± 1.3	80.4 ± 3.0	81.9 ± 1.5	- 70.8 ± 5.0	100.0 ± 2.0	57.5 ± 3.3	84.4 ± 10.1	70.0 ± 4.1	77.9 ± 0.6	83.3 ± 6.7	75.0 ± 11.3	95.8 ± 1.7	66.7 ± 0.0	Google/Gemma 7B -	50.00 ± 10.16	50.00 ± 10.00	25.00 ± 10.00	50.00 ± 10.00	37.50 ± 11.18	50.00 ± 12.25	50.00 ± 18.71
Mistral / Mixtral-8x7B	- 59.6 ± 6.9	60.0 ± 5.8	62.0 ± 2.7	61.4 ± 4.7	67.3 ± 1.7	56.7 ± 3.9	71.7 ± 2.2	82.0 ± 5.5	59.4 ± 4.4	65.3 ± 4.0	- 29.2 ± 6.8	71.7 ± 7.2	55.0 ± 8.4	32.3 ± 4.5	60.0 ± 5.8	47.1 ± 11.3	41.7 ± 10.0	41.7 ± 6.7	70.8 ± 15.2	66.7 ± 13.3	Meta/LLaMA 3 70B -	50.00 ± 9.35	50.00 ± 10.00	25.00 ± 0.00	75.00 ± 0.00	25.00 ± 0.00	100.00 ± 0.00	50.00 ± 10.00
											667+41	76.7 ± 4.0	50 8 ± 4 6	55 2 + 4 4	55.0 + 2.4	570+36	58.3 ± 4.1	58 2 + 4 1	75.0 + 6.1	66.7 + 14.0	Meta/LLaMA 3 8B -	75.00 ± 6.12	25.00 ± 10.00	50.00 ± 0.00	25.00 ± 18.71	50.00 ± 5.00	75.00 ± 10.00	50.00 ± 12.25
Mistral / Mixtral-8x22B	- 69.7 ± 5.3	70.0 ± 4.5	68.7 ± 2.0	62.4 ± 3.4	68.7 ± 2.9	65.6 ± 5.6	73.0 ± 3.9	83.1 ± 2.0	71.3 ± 5.4	$68.1 \pm 1.6$	- 00.7 ± 4.1	70.7 ± 4.0	50.8 ± 4.0								Microsoft/Phi-3 Mini -	$12.50 \pm 10.83$	$0.00 \pm 20.00$	$25.00 \pm 18.71$	$25.00 \pm 18.71$	0.00 ± 29.15	75.00 ± 12.25	$50.00 \pm 15.81$
Mistral / Mistral Large	- 69.9 ± 3.7	75.0 ± 4.1	69.7 ± 2.6	67.2 ± 2.6	70.9 ± 2.4	65.7 ± 3.8	70.7 ± 3.9	86.5 ± 3.1	68.6 ± 4.6	69.9 ± 2.7	- 45.8 ± 4.9	76.7 ± 4.9	59.2 ± 6.4	49.0 ± 4.5	60.0 ± 6.8	60.7 ± 3.3	66.7 ± 10.5	58.3 ± 9.7	79.2 ± 3.7	66.7 ± 12.5	Mistral/Mistral Large -	56.25 ± 18.37	$50.00 \pm 15.81$	50.00 ± 0.00	25.00 ± 20.00	37.50 ± 5.00	50.00 ± 0.00	75.00 ± 0.00
Google / Gemma 2B	- 43.7 ± 3.9	35.0 ± 1.9	31.8 ± 3.5	$40.4 \pm 0.2$	$39.2 \pm 1.9$	34.8 ± 1.7	47.6 ± 1.9	37.7 ± 4.5	$29.1 \pm 2.0$	44.0 ± 2.8	- 37.5 ± 14.5	23.3 ± 8.6	32.5 ± 3.2	33.3 ± 2.6	35.0 ± 6.0	45.7 ± 4.0	$41.7 \pm 4.1$	33.3 ± 3.3	54.2 ± 5.7	16.7 ± 8.2	Mistral/Mixtral-8x22B -	62.50 ± 3.95	25.00 ± 12.25	25.00 ± 12.25	$25.00 \pm 10.00$	50.00 ± 5.00	50.00 ± 10.00	75.00 ± 0.00
Google / Gemma 7B	- 46.9 ± 2.4	40.0 ± 7.6	41.4 ± 2.7	46.1 ± 5.5	46.7 ± 1.8	47.8 ± 2.7	52.1 ± 4.4	51.2 ± 3.6	44.3 ± 5.2	52.9 ± 3.2	- 37.5 ± 5.3	55.0 ± 4.9	46.7 ± 7.3	30.8 ± 3.2	35.0 ± 3.7	47.1 ± 6.3	33.3 ± 10.0	$50.0 \pm 4.1$	41.7 ± 8.7	66.7 ± 8.2	Mistral/Mixtral-8x7B -	56.25 ± 6.85	50.00 ± 18.71	25.00 ± 0.00	25.00 ± 10.00	37.50 ± 9.35	50.00 ± 12.25	75.00 ± 18.71
Meta / LLaMA 3 8B	- 42.5 ± 4.6	53.3 ± 4.9	56.0 ± 4.7	43.6 ± 2.9	56.4 ± 2.4	49.6 ± 1.2	59.6 ± 2.3	64.5 ± 4.7	53.4 ± 6.6	54.6 ± 2.6	- 33.3 ± 8.2	$60.0 \pm 4.6$	45.8 ± 4.5	40.6 ± 3.5	55.0 ± 3.7	62.1 ± 1.5	58.3 ± 8.5	58.3 ± 9.7	58.3 ± 6.8	$100.0 \pm 0.0$	OpenAl/GPT-3.5 Turbo -	68.75 ± 9.19	25.00 ± 15.81	50.00 ± 12.25	25.00 ± 15.81	50.00 ± 17.68	75.00 ± 20.00	50.00 ± 10.00
Meta / LLaMA 3 70B	- 67.4 ± 4.6	68.3 ± 1.6	65.4 ± 3.0	66.3 ± 3.3	72.6 ± 2.2	66.4 ± 1.7	72.7 ± 2.5	82.5 ± 2.9	61.0 ± 3.3	71.7 ± 1.7	- 41.7 ± 5.5	71.7 ± 7.3	54.2 ± 3.3	52.1 ± 3.7	50.0 ± 2.0	59.3 ± 4.7	66.7 ± 6.2	58.3 ± 4.1	75.0 ± 4.9	66.7 ± 0.0	OpenAl/GPT-4 Turbo -	68.75 ± 3.95	50.00 ± 18.71	50.00 ± 0.00	50.00 ± 18.71	50.00 ± 14.58	25.00 ± 12.25	25.00 ± 0.00
Cohere / Command R+	- 55.2 ± 3.5	56.7 ± 4.2	51.6 ± 3.0	51.7 ± 6.3	44.8 ± 4.7	51.0 ± 5.3	61.8 ± 4.2	58.3 ± 10.0	56.1 ± 1.7	31.7 ± 9.2	- 37.5 ± 7.3	63.3 ± 11.3	36.7 ± 8.9	33.3 ± 4.4	40.0 ± 9.3	55.0 ± 4.3	50.0 ± 13.9	41.7 ± 8.5	45.8 ± 17.0	83.3 ± 17.0	OpenAl/GPT-4o -		50.00 ± 12.25	50.00 ± 15.81	50.00 ± 0.00	75.00 ± 6.12	75.00 ± 18.71	25.00 ± 0.00
Microsoft / Phi-3 Mini	- 58.0 ± 9.7	61.7 ± 3.7	56.1 ± 3.6	62.7 ± 2.3	62.4 ± 0.7	58.1 ± 3.6	65.0 ± 2. <u>5</u>	87.1 ± 7. <u>8</u>	56.5 ± 5.3	53.6 ± 1.6	-	23.3 ± 4.5	45.0 ± 12.2		25.0 ± 8.9	20.7 ± 15.8	16.7 ± 21.3	33.3 ± 19.3	29.2 ± 16.2			Alternative Investments for Portfolio	Asset Allocation and Related	Capital Market Expectations	Derivatives and Currency Management	Portfolio Management	Private Wealth Management	Trading, Performance Evaluation, and
	Quantitative Methods	Portfolio Management	Fixed Income	Financial Statement Analysis	Ethics	Equity	Economics	Derivatives	Corporate Issuers	Alternative Investments	Quantitative Methods	Portfolio Management	Fixed Income	Financial Reporting & Analysis	Ethics	Equity	Economics	י Derivatives	Corporate Issuers	Alternative Investments		Management	Decisions in Portfolio Management					Manager Selection

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

LLMs as CFA Professionals?	TL;DR
Provider Model Level I Level II Level III L U L U L U	
GPT-3.5 Turbo ✓ X X X X X	Proprietary models, such as GPT-40 and Claude 3 Opus, excel in CFA exam

OpenAI								-
OpenAI	GPT-4 Turbo	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	×	×	
	GPT-40	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	×	×	
Anthropic	Claude 3 Opus	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	×	×	Table 3: LLM's ability to pass each CFA level
	Mixtral-8x7B	$\checkmark$	×	×	×	×	×	using 1S-CoT or RAG,
Mistral	Mixtral-8x22B	$\checkmark$	×	$\checkmark$	×	×	×	with the lower bound score L (≥ 60%) and
	Mistral Large	$\checkmark$	X	$\checkmark$	X	X	×	upper bound score
Google	Gemma 2B	X	×	×	×	×	×	- U (≥ 70%).
Google	Gemma 7B	X	×	×	×	×	×	$\checkmark$ indicates the LLM
	LLaMA 3 8B	X	×	×	×	×	×	should pass the exam
Meta	LLaMA 3 70B	$\checkmark$	×	×	×	×	×	while X indicates it should fail.
	LLaMA 3 8B + RAG	$\checkmark$	×	×	×	×	×	_
	LLaMA 3 70B + RAG	$\checkmark$	$\checkmark$	$\checkmark$	×	×	×	_
Cohere	Command R+	X	×	×	×	×	×	-
Microsoft	Phi-3-mini	$\checkmark$	×	×	×	×	×	-
Ai2	OLMo 7B	¥	X	×	X	X	X	-

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.