Advanced Financial Reasoning at Scale: A Comprehensive Evaluation of Large Language Models on CFA Level III

Pranam Shetty[1] , **Abhisek Upadhayaya**[3] , **Parth Mitesh Shah**[2] , **Shilpi Nayak**[2] , **Anna Joo Fee**[2] , **Srikanth Jagabathula**[3] ,

[1]Rochester Institute of Technology, [2]GoodFin, Inc, [3]New York University

ps9960@rit.edu, {parth, shilpi, anna}@goodfin.com, au2216@nyu.edu, sjagabat@stern.nyu.edu

Abstract

As financial institutions increasingly adopt Large Language Models (LLMs), rigorous domain-specific evaluation becomes critical for responsible deployment. This paper presents a comprehensive benchmark evaluating 23 state-of-the-art LLMs on the Chartered Financial Analyst (CFA) Level III exam—the gold standard for advanced financial reasoning. We assess both multiple-choice questions (MCQs) and essay-style responses using multiple prompting strategies including Chain-of-Thought and Self-Discover. Our evaluation reveals that leading models demonstrate strong capabilities, with composite scores such as 79.1% (o4-mini) and 77.3% (Gemini 2.5 Flash) on CFA Level III. These results, achieved under a revised, stricter essay grading methodology, indicate significant progress in LLM capabilities for high-stakes financial applications. Our findings provide crucial guidance for practitioners on model selection and highlight remaining challenges in cost-effective deployment and the need for nuanced interpretation of performance against professional benchmarks. [1]

Keywords: Large Language Models, Financial Reasoning, CFA Level III Benchmark, Chain-of-Thought Prompting, Self-Consistency Prompting, Self-Discover Prompting

1 Introduction

Large Language Models (LLMs) are rapidly transforming financial services, with institutions increasingly deploying these systems for tasks ranging from research and education to trading and advisory roles [5, 6]. As models demonstrate increasing proficiency on general reasoning tasks, a critical question emerges: can they handle the specialized, high-stakes analytical reasoning required for professional financial decision-making? [3, 2]

The integration of LLMs into financial environments requires robust, professionally relevant benchmarks that assess complex cognitive skills beyond simple recall. Traditional benchmarks often fail to capture the nuanced reasoning required in financial analysis, where practitioners must synthesize quantitative data, regulatory knowledge, and market dynamics to make investment decisions [9].

The Chartered Financial Analyst (CFA) Level III exam represents an ideal evaluation framework, being the culminating assessment for investment management professionals. Level III is the final CFA exam, taken after the candidates have passed Levels I and II, and is centered portfolio management and wealth planning, assessing candidates' ability to implement financial knowledge in real-world scenarios [4, 7]. It uses a mixed format, comprising 11 item set or multiple choice questions (MCQs) and 11 essay or constructed response questions. Each question set begins with a narrative scenario and data (the vignette), which must be carefully analyzed to answer the questions. This dual format comprehensively tests higher-order cognitive skills including analysis, synthesis, and professional judgment [12] over rote memorization. The exam's rigorous standards, reflected in typical pass rates of 50-60%, make it an excellent benchmark for assessing LLM capabilities in advanced financial reasoning.

Previous studies have shown that while state-of-the-art models at the time of the analysis perform reasonably well on CFA Levels I and II, they struggle significantly with Level III's advanced reasoning requirements, particularly on the essay questions [9]. This performance gap highlights critical limitations in current LLMs' ability to handle sophisticated financial analysis—precisely the capabilities needed for real-world deployment in investment management.

However, prior evaluations have been limited in scope (typically 8-12 models), lack systematic prompting strategy comparison, and provide insufficient cost-effectiveness analysis for deployment decisions. Moreover, the rapid emergence of reasoning-focused models (o3-mini, DeepSeek-R1) and latest frontier models remains largely unevaluated on complex financial reasoning tasks.

This study addresses these critical gaps by providing the first comprehensive evaluation of 23 diverse LLMs on CFA Level III exam, including frontier, reasoning models and specialized financial models with open-source alternatives. Our contributions include:

- **First Demonstration of Professional-Grade LLM Performance:** We show that frontier models achieve composite scores of 79.1% (o4-mini) and 75.9% (Gem-

[1]Accepted at FinLLM@IJCAI 2025

ini 2.5 Pro) on CFA Level III, substantially exceeding the estimated 63% passing threshold [1]—representing a significant milestone in financial Artificial Intelligence (AI) capabilities.

- **Cost-Effectiveness Framework for Financial AI:** We provide a systematic analysis of accuracy-efficiency trade-offs in financial reasoning, which is essential for operationalizing these models in practice. For instance, we find that advanced prompting improves MCQ accuracy by 7.8 percentage points but at 3-11x cost increases. For essays, Chain-of-Thought strategies yield the best Essay Scores under strict evaluation, but also entail higher costs. Specialized models like Palmyra-fin achieve respectable MCQ accuracy (68.3%) and essay scores (58.39% with Zero Shot) at lower computational costs—delivering actionable guidance for deployment decisions.
- **Reproducible benchmarking framework for financial applications:** We establish a comprehensive assessment framework combining MCQ analysis and automated essay evaluation using semantic similarity and expert-level LLM scoring, providing the research community with reproducible benchmarks for tracking progress in financial AI reasoning capabilities.

2 Methodology

2.1 Dataset Construction and Validation

As official CFA Level III questions post-2018 are not publicly available due to intellectual property protections, we constructed our dataset using mock exam materials–including questions, answers, and grading rubrics–from AnalystPrep, a reputable CFA preparation provider with over 100,000 candidates, following similar approaches in recent financial AI research [9]. This paywall protection significantly reduces the likelihood of test data contamination in model training corpora. Our dataset comprises two components reflecting the actual exam structure:

Multiple-Choice Questions (MCQs) dataset, consisting of 60 questions, organized into 10 vignettes with 6 questions each, covering all major Level III curriculum areas.

Essay Questions dataset, comprising 11 unique vignettes with 43 total questions (149 total points) covering major Level III curriculum areas: Private Wealth Management (2 vignettes), Portfolio Management (2 vignettes), Private Markets (2 vignettes), Asset Management, Derivatives and Risk Management, Performance Measurement, Portfolio Construction, and Ethical and Professional Standards. Each vignette presents realistic financial scenarios followed by 2-5 open-ended questions requiring synthesis of multiple concepts and clear articulation of investment reasoning.

Data preprocessing involved Optical Character Recognition(OCR) extraction from PDF sources, conversion to structured JavaScript Object Notation(JSON) format, and expert review to confirm alignment with the official CFA curriculum standards and appropriate difficulty levels.

2.2 Model Selection and Categorization

Our benchmark includes 23 state-of-the-art LLMs, selected to represent a diverse range of capabilities, architectural designs, and provider ecosystems. These models are categorized as follows:

Frontier Models: This category includes highly capable, general-purpose models, often representing the cutting edge from major providers: GPT-4o, GPT-4.1 series (including base, mini, and nano variants), Grok 3, Claude-3.5-Sonnet, Claude-3.5-Haiku, Claude-3.7-Sonnet, Claude-Opus-4, Claude-Sonnet-4, Gemini 2.5 Flash, and Mistral Large Official.

textbfReasoning-Enhanced Models: These models are specifically designed or configured to improve reasoning capabilities, often involving increased computational resources at inference time: o3-mini, o4-mini, Deepseek-R1, and Grok-3-mini-beta (evaluated with high and low reasoning effort settings, considered as seperate models).

Reasoning-Enhanced Models: These models are specifically designed or configured to improve reasoning capabilities, often involving increased computational resources at inference time: o3-mini, o4-mini, Deepseek-R1, Grok-3-mini-beta (evaluated with high and low reasoning effort settings, considered as separate models), Gemini 2.5 Pro and Gemini 2.5 Flash, Claude Opus 4, Claude Sonnet 4, and Claude 3.7 Sonnet.

Open-Source Models: This group comprises prominent open-source LLMs that offer accessible and adaptable alternatives. Models tested: Llama-3.1-8B instant, Llama-3.3-70B, Llama-4-Maverick, Llama-4-Scout and Deepseek-R1.

Specialized Models: These models have been trained or fine-tuned for specific domains or tasks. Models tested: Palmyra-fin (finance-domain specialized).

2.3 Prompting Strategy Design

We evaluated three carefully designed prompting approaches to assess their impact on financial problem-solving performance on both MCQs and essays:

Zero-Shot: Standard prompts requesting direct answers (selecting an option for MCQs, generating text for essays) without explicit reasoning structures. This establishes baseline performance and tests intrinsic model capabilities for solving problems and generating financial analyses directly.

Chain-of-Thought (CoT) with Self-Consistency (SC): Prompts explicitly instruct models to output a step-by-step reasoning process before providing the final answer (for MCQs, this involves selecting an option; for essays, generating the full text). Self-consistency [13] was employed by generating N=3 and N=5 distinct reasoning paths and corresponding answers. For MCQs, the final answer was determined by majority vote over these samples; for essays, each of the model generated essays was self-graded using the same model on a 10-point grading rubric and the essay with the highest score was selected as the final answer, with a fallback to the first valid sample if selection failed. The goal of this implementation was to more closely mimic the behavior of human candidates in the CFA exam, who typically articulate their reasoning process before arriving at a final answer.

By simulating multiple independent reasoning paths and aggregating the results, this approach aims to reflect authentic exam-taking strategies and provide a more robust assessment of model capabilities.

Self-Discover: An adaptive prompting technique where models first devise their own reasoning structure for the problem (e.g., select relevant reasoning modules, adapt them to the specific problem, and outline a structured approach) before constructing the final answer/response [14]. This tests the model's ability to engage in explicit metacognitive planning for complex financial problem-solving.

2.4 Evaluation Metrics and Framework

MCQ performance was evaluated by comparing generated answer with the correct answer for accuracy. Essay evaluation was more challenging because we had to compare model generated response with the suggested response on various qualitative factors, including semantic similarity, reasoning quality, and completeness. In line with established financial AI practices [9], we report scores from three different evaluation strategies: cosine similarity, Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [8], and LLM-as-a-judge. Cosine similarity measures the similarity between model generated answer and the correct answer as the cosine of the angle between the two TF-IDF vectors. Similarly, ROUGE-L measures lexical similarity by comparing the longest subsequence overlap between model generated and correct answers.

For LLM-as-a-judge, we used GPT-4.1 configured with a temperature of 0.0 (to ensure consistency) as the judge for all models and prompting strategies. For each of the prompting strategies, the model output (direct answer in case of Zero-shot and answer and complete reasoning process in case of Self-Discover and CoT-SC (N=3/5) prompting strategies), the correct answer, the question along with the complete vignette, the grading rubric, and the minimum and maximum possible scores were given to GPT-4.1 for grading. The rubric detailed the specific criteria for awarding points. The prompt was designed to enforce strict criteria-based scoring, disallowing subjective partial credit and requiring integer-only scores, with validation for score range adherence. GPT-4.1 then assigned an integer score based on technical accuracy, reasoning quality, relevance, communication, and rubric adherence. This score was subsequently weighted by the question's point value and normalized to a 0-100 scale for direct comparison with MCQ accuracy and composite scoring.

Overall Score Calculation: To reflect the CFA Level III exam structure [4], we compute overall scores by equally weighting MCQ and essay performance, consistent with the exam's 12-point allocation across 11 item sets per type. This aligns with Mahfouz et al. [9], who combine components using percentage-based accuracy. This method is sound as it (1) mirrors official exam weighting, (2) supports comparison with prior work, and (3) enables intuitive interpretation of balanced performance.

Efficiency and Cost Metrics: For all model runs, we tracked key operational and economic indicators: average API latency, input/output token usage (including internal reasoning tokens where available), estimated cost per evaluation (USD), average essay answer length, and total run time per model-strategy pair.

3 Results

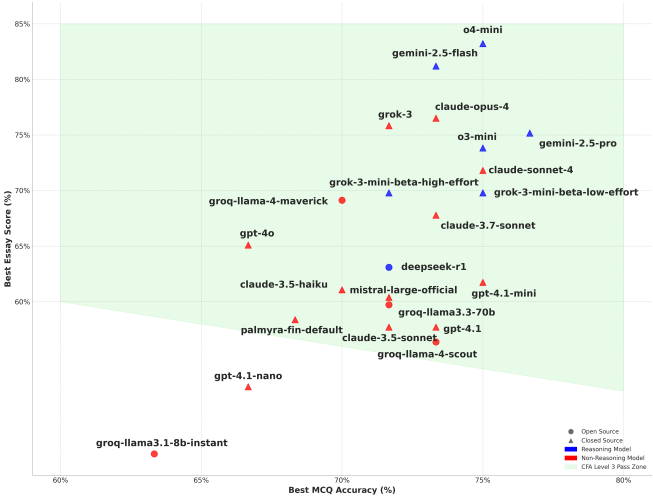

Figure 1: MCQ vs. Essay performance correlation across all evaluated models, plotting best MCQ accuracy (x-axis) against best essay Essay Score (y-axis). Key: Blue: Reasoning, Red: Non-reasoning; Triangle: Closed-source, Circle: Open-source. The green hue indicates models achieving the Minimum Passing Score (MPS) of 60-70 for the CFA Level III examination. The plot reveals distinct performance clusters and varying architectural strengths across task types.

This comprehensive evaluation of 23 LLMs on CFA Level III reveals that leading models exhibit strong financial reasoning capabilities. Top performers achieve notable composite scores, such as 79.1% (o4-mini) and 77.3% (Gemini 2.5 Flash); see Table 1. To summarize key findings, Figure 1 visualizes model performance across both Multiple-Choice Question (MCQ) accuracy and best Essay Score, with the CFA Level III minimum passing standard (MPS) [1] shaded in green. This composite view offers a holistic understanding of advanced financial task performance, revealing that high performance in one dimension does not always translate to the other.

3.1 Performance leaders

Our evaluation reveals a clear performance hierarchy among the 23 evaluated models.

For MCQ, Gemini 2.5 Pro achieved the highest accuracy at 76.7% with both Zero Shot (Single Pass) and Self-Discover prompting strategies. Following closely, 'o4-mini' (Self-Discover), 'o3-mini' (Self-Consistency CoT N=5), and 'gemini-2.5-pro' (Self-Consistency CoT N=3) all achieved 75.0% accuracy. Among specialized financial models, 'Palmyra-fin' from Writer achieved 68.3% accuracy with Self-Discover prompting, demonstrating competitive performance at a moderate cost of $0.70 per run.

For essay questions, while 'o4-mini' led in Essay Score (83.22), 'Gemini 2.5 Flash' ranked highest in semantic

Provider	Model	MCQ Score (Strat.)	Essay Score (Strat.)	Human Essay Score	Cosine Sim.	ROUGE-L	Overall Score
OpenAI	o4-mini	75.00 (Self-Discover)	**83.22** (CoT-SC N=3)	79.9	0.4939	<u>0.1491</u>	**79.1**
	o3-mini	75.00 (CoT-SC N=5)	73.83 (Self-Discover)	-	0.5532	0.0881	74.4
	o3-mini	75.00 (CoT-SC N=5)	73.83 (CoT-SC N=5)	-	0.5493	0.1179	74.4
Google	Gemini 2.5 Flash	73.33 (Zero Shot)	<u>81.21</u> (CoT-SC N=3)	78.5	**0.5793**	0.0847	<u>77.3</u>
	Gemini 2.5 Pro	**76.67** (Zero Shot)	75.17 (CoT-SC N=3)	83.2	<u>0.5695</u>	0.0774	75.9
	Gemini 2.5 Pro	75.00 (CoT-SC N=3)	75.17 (CoT-SC N=3)	83.2	0.5695	0.0774	75.1
	Gemini 2.5 Pro	76.67 (Self-Discover)	61.74 (Self-Discover)	-	0.5691	0.0704	69.2
Anthropic	Claude Opus 4	73.33 (CoT-SC N=3)	76.51 (CoT-SC N=3)	81.9	0.5465	0.1256	74.9
	Claude Sonnet 4	75.00 (CoT-SC N=3)	71.81 (CoT-SC N=3)	73.8	0.5238	0.1241	73.4
	Claude 3.7 Sonnet	73.33 (Self-Discover)	67.79 (CoT-SC N=3)	77.8	0.5410	0.1172	70.6
xAI	Grok-3	71.67 (CoT-SC N=3)	75.84 (CoT-SC N=5)	-	0.5648	0.0945	73.8
	Grok-3 Mini Low Effort	75.00 (CoT-SC N=3)	69.80 (CoT-SC N=3)	71.8	0.5403	0.0738	72.4
DeepSeek	DeepSeek-R1	71.67 (Self-Discover)	63.09 (CoT-SC N=3)	62.4	0.5383	0.1132	67.4
Writer	Palmyra-fin	68.33 (Self-Discover)	58.39 (Zero Shot)	-	0.5376	**0.1652**	63.4

Table 1: CFA Level III top model performance summary (subset of 23 evaluated models shown for space). MCQ and Essay scores reflect best performance across all strategies tested (strategy noted in parentheses), with all scores reported as percentages. Models may appear multiple times when different strategy combinations yield distinct overall scores. Human Essay scores (graded by a CFA expert) reflect evaluation of responses generated using the best-performing essay strategy: Self-Consistency with 3 samples (CoT-SC N=3). Cosine Similarity and ROUGE-L scores correspond to the essay strategy associated with the listed Essay Score. Overall Score is the average of MCQ accuracy and normalized Essay Score. Bold indicates best performance, underlined indicates second-best. Gray highlighting denotes non-reasoning models. Claude models were evaluated with thinking mode disabled. Human evaluation was limited in scope, focusing on the most promising approach (CoT-SC N=3) for comprehensive assessment.

similarity (0.5793), and 'Palmyra-fin' topped lexical overlap (0.1652). These results highlight that models vary in strengths across evaluation dimensions, underscoring the task-specific nature of performance.

3.2 Model Type Analysis

Table 2 shows that reasoning models outperformed non-reasoning models across all performance metrics. For MCQ tasks, reasoning models achieved 70.6% accuracy compared to 65.8% for non-reasoning models. In essay evaluation, reasoning models scored 73.71 on LLM grading versus 61.87 for non-reasoning models, representing a 19.1% improvement. Notably, human evaluators showed similar patterns, rating reasoning model outputs at 75.97 compared to 62.52 for non-reasoning models, validating the LLM grading consistency and confirming the superior quality of reasoning model responses.

This demonstrates a clear quality-efficiency trade-off where reasoning capabilities deliver accuracy improvements at substantial computational cost, requiring careful consideration for deployment scenarios based on performance requirements and resource constraints.

Model Type	MCQ Performance		Essay Performance		
	Accuracy	Time(s)	LLM	Human	Time(s)
Reasoning	**0.706**	**59.2**	**73.71**	**75.97**	87.4
Non-Reasoning	0.658	17.9	61.87	62.52	**42.6**

Table 2: Performance comparison by architecture type across MCQ and Essay tasks (avg. values for Self consistency n = 3). MCQ accuracy is on a 0-1 scale, Human Score and Essay Score on a 0-100 scale. Human Scores are on Self-Consistency CoT n=3. Time shows average processing seconds. Bold indicates superior performance.

3.3 Prompting Strategy Effectiveness

Advanced prompting strategies show mixed but generally positive results across the evaluated models. Table 3 summarizes the aggregate performance across both MCQ and essay tasks, along with efficiency metrics.

Strategy	MCQ Performance			Essay Performance		
	Acc.	Time	Cost	Score	Cos.Sim.	ROUGE-L
Zero Shot	61.8	6.0	0.255	57.35	0.5151	**0.1522**
Self-Discover	68.6	14.9	0.686	46.32	0.5012	0.0977
CoT-SC (N=3)	69.1	37.9	1.701	<u>60.40</u>	<u>0.5353</u>	0.1227
CoT-SC (N=5)	**69.6**	63.1	2.840	**61.03**	**0.5376**	<u>0.1228</u>

Table 3: Combined MCQ and essay prompting strategy comparison. MCQ metrics: Accuracy (%), processing time (seconds), and cost (USD). Essay metrics: LLM-assessed score (0-100 scale), cosine similarity (0-1 scale), and ROUGE-L F1 (0-1 scale). All values represent means aggregated across evaluated models. Bold indicates best performance, underlined shows second best.

The combined analysis reveals consistent patterns across both task types, with advanced prompting strategies delivering performance improvements at substantial computational costs. For MCQs, Chain-of-Thought with Self-Consistency (N=5) achieves the highest accuracy (69.6%), representing a 7.8 percentage point improvement over Zero Shot prompting (61.8%), but requires 10.5x longer processing time (63.1s vs 6.0s) and 11.1x higher cost ($2.84 vs $0.255) compared to Zero Shot. Self-Discover provides the most balanced trade-off, delivering a substantial 6.8 percentage point accuracy improvement over Zero Shot while increasing cost by only 2.7x.

For essays, CoT-SC (N=5) achieves the highest mean Essay Score (61.03%) and also the highest mean Cosine Similarity (0.5376), indicating strong semantic alignment on average. The CoT-SC N=3 variant follows closely in both Essay Score (60.40%) and mean Cosine Similarity (0.5353). For lexical overlap, Zero Shot prompting yields the highest mean ROUGE-L F1 score (0.1522), suggesting closer textual matching to reference answers on average with this simpler strategy. CoT-SC (N=5) however requires 8.6x (130.48s vs 15.18s) longer processing times and 8.8x ($3.390 vs $0.387) evaluation cost compared to Zero-shot. Self-Discover, while offering a balance in some scenarios, results in the lowest mean scores across all three essay metrics under the current strict evaluation.

These findings highlight that while Self-Consistency approaches generally provide the highest LLM-assessed quality and semantic similarity, the simpler Zero Shot strategy can lead to better lexical overlap on average. The substantial performance decrease for Self-Discover across all metrics under stricter grading suggests its metacognitive planning may not align well with requirements for precise, rubric-adherent answers.

Strategy	MCQ Performance		Essay Performance	
	Latency (s)	Cost ($)	Latency (s)	Cost ($)
Zero Shot	6.0	0.255	15.18	0.387
Self-Discover	14.9	0.686	25.27	0.703
CoT-SC (N=3)	37.9	1.701	69.83	2.089
CoT-SC (N=5)	63.1	2.840	130.48	3.390

Table 4: Comprehensive efficiency comparison across MCQ and essay tasks. MCQ metrics show average latency and cost per task. Essay metrics show average latency per task and total cost per model evaluation. Values aggregated across all 23 models for MCQ and all models for essays.

Open-Source versus Closed-Source Model Comparison

The aggregated data in Table **??** suggests that the group of closed-source models evaluated in this benchmark, on average, achieved slightly higher scores in MCQ accuracy (72.69 versus 70.00) and LLM-graded essay quality (67.80 versus 58.93) compared to the group of open-source models. Conversely, open-source models exhibited a marginal lead in mean best Cosine Similarity (0.5413 versus 0.5405) and ROUGE-L F1 scores (0.1512 versus 0.1499).

4 Human vs LLM Grade Analysis

This section presents a comprehensive analysis comparing human evaluator grades with LLM-generated grades for the same set of responses using the self-consistency Chain-of-Thought strategy with n=3 iterations. Unlike previous analyses that aggregated across multiple strategies, this comparison ensures methodological consistency by examining identical response sets evaluated by both human expert and LLM grader.

4.1 Methodology & Metrics

The study examines 23 distinct language models across 43 CFA Level 3 questions using a single, consistent evaluation strategy: self-consistency Chain-of-Thought with three iterations (n=3). The experimental design ensures complete methodological consistency through 100% question coverage across both grading methods, resulting in 989 question-answer pairs (43 questions × 23 models) evaluated by both human expert and LLM grader. This comprehensive coverage eliminates sampling bias and provides sufficient depth for rigorous statistical analysis, forming the central empirical foundation for validating reasoning model performance on professional financial assessments.

We assess alignment using:

- **Performance Gap** (Δ): Human - LLM grade difference
- **Variance**: Consistency of differences across questions
- **Agreement Rate**: % of exact matches
- **Statistical Significance**: p-values and confidence intervals

4.2 Overall Results

Table 5 presents the comprehensive grading analysis across 23 language models evaluated on 43 CFA Level 3 questions. The results reveal systematic patterns in human-LLM grading alignment, with human evaluators consistently assigning higher scores than the automated LLM grader across the majority of models.

The aggregate analysis reveals three key empirical findings. First, human evaluators systematically assign higher scores than the LLM grader, with an average performance gap of +5.6 percentage points (68.4% vs. 62.8%). This human-favored bias manifests in 20 of 23 models, suggesting potential systematic differences in evaluation criteria or leniency between human and automated assessment approaches. Second, grading alignment between human and LLM evaluators achieves moderate consistency, with an overall agreement rate of 70.4% across all question-model pairs. Third, grading variance exhibits substantial heterogeneity across models, ranging from 0.543 (Claude Sonnet 4) to 2.010 (o4 mini), indicating differential stability in assessment consistency.

Notably, only three models demonstrate LLM-favored grading bias: o4 mini (-3.3 points), Gemini 2.5 Flash (-2.7 points), and DeepSeek r1 (-0.7 points). These exceptions warrant particular attention as they represent instances where automated grading potentially identifies performance characteristics that human evaluators may undervalue or assess differently. The performance distribution spans a wide range, from

Model	Human	LLM	Δ	Var.	Agree %
Gemini 2.5 pro	83.2	75.2	+8.0	1.635	69.8
Claude Opus 4	81.9	76.5	+5.4	1.441	76.7
o4 mini	79.9	83.2	-3.3	2.010	76.7
Gemini 2.5 Flash	78.5	81.2	-2.7	1.277	72.1
Grok 3	77.9	73.2	+4.7	1.520	79.1
Claude 3.7 Sonnet	77.9	67.8	+10.1	1.423	81.4
Grok 3 mini high effort	75.8	69.8	+6.0	1.503	72.1
o3 mini	76.0	72.5	+3.5	1.828	67.4
Grok 3 mini low effort	75.2	72.7	+2.5	0.733	74.4
Claude sonnet 4	73.8	71.8	+2.0	0.543	79.1
Claude 3.5 sonnet	70.5	57.7	+12.8	1.443	76.7
Claude 3.5 haiku	67.1	61.1	+6.0	1.169	69.8
GPT 4.1 mini	67.1	60.4	+6.7	0.992	69.8
Llama 4 Maverick*	67.8	57.7	+10.1	1.756	65.1
GPT 4.1	68.5	54.4	+14.1	1.684	67.4
GPT 4o	64.4	61.1	+3.3	1.677	65.1
Mistral Large	61.7	56.4	+5.3	1.250	67.4
Llama 4 Scout*	59.7	55.7	+4.0	0.885	65.1
DeepSeek r1*	62.4	63.1	-0.7	0.928	72.1
Llama 3.3 70b*	57.7	51.7	+6.0	1.027	69.8
GPT 4.1 nano	55.7	52.4	+3.3	1.772	67.4
Palmyra Fin	55.7	50.3	+5.4	1.060	74.4
Llama 3.1 8b*	38.9	31.5	+7.4	1.100	62.8
Overall Average	68.4	62.8	+5.6	1.344	70.4

Table 5: Comparison of Human and LLM Grades for All Models with Performance Gap (Δ), Variance (SC-CoT n=3), and Agreement Rate (%). Reasoning models are highlighted in gray. Open source models are marked with an asterisk (*).

Gemini 2.5 Pro achieving the highest human-assigned score (83.2%) to Llama 3.1 8b recording the lowest performance (38.9% human, 31.5% LLM), demonstrating substantial capability differences across the evaluated model landscape.

4.3 Model Category Analysis

This section examines performance patterns across distinct model categories, focusing on architectural differences and their impact on human-LLM grading alignment. We analyze two primary taxonomies: reasoning-enhanced versus standard models, and open-source versus proprietary architectures.

Reasoning Models Performance

Reasoning models, characterized by enhanced deliberative capabilities and multi-step inference architectures, demonstrate distinctive performance patterns in the dual-grading analysis. Ten models in our evaluation incorporate specialized reasoning mechanisms: Gemini 2.5 Pro, Claude Opus 4, o4 mini, Claude 3.7 Sonnet, Grok 3 mini variants, o3 mini, Claude Sonnet 4, and DeepSeek r1.

The reasoning model cohort exhibits superior absolute performance, with an average human-assigned score of 76.1% compared to 63.8% for standard models—a statistically significant difference of 12.3 percentage points. However, this performance advantage manifests differently in human versus LLM evaluation patterns. Reasoning models demonstrate greater variance in grading alignment, ranging from exceptional consistency (Claude Sonnet 4: variance = 0.543) to notable disagreement (o4 mini: variance = 2.010).

While the original claim suggests that reasoning-focused models (such as o4 mini, Gemini 2.5 Flash, and DeepSeek r1) receive higher scores from LLM graders because they demonstrate logical structure or mathematical precision undervalued by humans, this interpretation may not fully hold up under scrutiny.

The architectural heterogeneity within reasoning models yields distinct grading patterns. Hybrid reasoning models (Claude Opus 4, Claude Sonnet 4) demonstrate balanced performance with moderate variance and high agreement rates (76.7% and 79.1% respectively), while pure reasoning architectures exhibit more polarized outcomes. This finding has implications for the deployment of reasoning models in professional assessment contexts, where grading consistency may be as important as absolute performance.

Open-Source Models

Open-source models represent a critical category for understanding the democratization of advanced language capabilities in financial assessment. Table 6 presents the comprehensive analysis of five open-source models in our evaluation.

Open-source models demonstrate a performance gap relative to proprietary alternatives, with an average human-assigned score of 57.3% compared to 72.1% for closed-source models. This 14.8 percentage point difference reflects both computational resource constraints and the proprietary advantages in training data curation and architectural optimization. However, the open-source category exhibits notable heterogeneity, with Llama 4 Maverick achieving com-

Model	Human	LLM	Δ	Var	Agree %
Llama 4 Maverick	67.8	57.7	+10.1	1.756	65.1
DeepSeek r1	62.4	63.1	-0.7	0.928	72.1
Llama 4 Scout	59.7	55.7	+4.0	0.885	65.1
Llama 3.3 70b	57.7	51.7	+6.0	1.027	69.8
Llama 3.1 8b	38.9	31.5	+7.4	1.100	62.8
Category Average	57.3	51.9	+5.4	1.139	67.0

Table 6: Comparison of Human and LLM Grades for Open-Source Models with Performance Gap (Δ), Variance (n=3), and Agreement Rate (%).

petitive performance (67.8% human score) while Llama 3.1 8b significantly underperforms (38.9% human score).

The grading alignment patterns within open-source models reveal important insights for model selection in resource-constrained environments. DeepSeek r1, the sole reasoning-enhanced open-source model, demonstrates the most favorable characteristics: minimal grading bias (-0.7 points), lowest variance (0.928), and highest agreement rate (72.1%). This performance profile suggests that reasoning capabilities may be particularly valuable for achieving grading consistency in open-source architectures.

Variance analysis reveals that open-source models generally exhibit more stable grading patterns than their proprietary counterparts (average variance: 1.139 vs. 1.434), potentially reflecting more conservative response generation strategies. This consistency advantage, while accompanied by lower absolute performance, may be valuable in applications where predictable grading behavior is prioritized over peak performance.

The economic implications of these findings are substantial. Open-source models offer a viable pathway for implementing automated CFA assessment systems with acceptable grading alignment (67.0% average agreement rate) at significantly reduced computational and licensing costs. However, the performance trade-offs must be carefully evaluated against the specific requirements of professional certification contexts.

4.4 Key Findings

Grading Bias Patterns

Our analysis reveals a systematic human-favored grading bias across the model landscape, with an aggregate performance gap of +5.6 percentage points (human: 68.4

Reasoning models exhibit distinctive bias patterns, accounting for all LLM-favored cases and demonstrating reduced average bias magnitude (+2.8 points vs. +7.2 points for standard models). This suggests that advanced reasoning architectures may produce outputs that align more closely with automated evaluation criteria, potentially due to structured logical chains that LLM graders evaluate more favorably. In contrast, open-source models show amplified human-favored bias (+5.4 points average), possibly reflecting limitations in response sophistication that human evaluators compensate for through contextual leniency.

These bias patterns have significant implications for automated assessment systems in professional finance certification. The systematic positive delta indicates potential calibration needs for LLM graders to better approximate human expert judgment, particularly in domains requiring nuanced financial reasoning.

Consistency & Agreement Analysis

Grading consistency analysis reveals moderate overall agreement between human and LLM evaluators, with a 70.4

Variance patterns demonstrate notable heterogeneity, with model-level variance ranging from 0.543 (Claude Sonnet 4) to 2.010 (o4 mini). Reasoning models exhibit higher average variance (1.401 vs. 1.297 for standard models), suggesting that advanced architectures may introduce greater evaluative complexity. Conversely, open-source models display more stable patterns (average variance: 1.139), potentially due to conservative response generation strategies that yield predictable grading outcomes.

Agreement rates correlate positively with absolute performance (r = 0.62, p ¡ 0.01), with top-performing models achieving 76-81% agreement. This relationship implies that model capability serves as a proxy for grading reliability in CFA assessment contexts. The observed consistency metrics underscore the need for hybrid evaluation frameworks that combine human expertise with automated efficiency to achieve optimal reliability in professional certification processes.

5 Discussion

5.1 Key Findings and Implications

Frontier LLMs Show Strong Performance, but Passing Standards Require Nuance: Our evaluation shows that leading frontier models achieve notable results on CFA Level III, with composite scores such as 79.1% (o4-mini) and 77.3% (Gemini 2.5 Flash). While these exceed the estimated 63% passing threshold, comparing them directly to the CFA Level III MPS requires careful consideration given exam conditions differences and our revised conservative essay scoring.

Human Evaluation Reveals Distinct Performance Patterns: Human grading provides critical validation of model capabilities, with Gemini 2.5 Pro achieving the highest human essay score (83.2%), followed by Claude Opus 4 (81.9%) and o4-mini (79.9%). Notably, human evaluation sometimes diverges from automated scoring—Gemini 2.5 Pro's superior human performance (83.2%) contrasts with its automated essay score (75.17%), suggesting that human evaluators may recognize nuanced reasoning patterns that automated metrics miss.

Self-Assessment Capabilities Reveal Model Confidence Patterns: Model's self-grading capabilities show systematic patterns, with CoT-SC leading to higher final grades compared to Zero Shot or Self-Discover approaches. However, significant variation in essay scores under strict external grading indicates that internal model confidence may not align with rigorous external rubrics, crucial for risk management in financial decision-making.

Human-AI Performance Correlation Insights: The correlation between human and automated evaluation varies significantly across models. Claude models demonstrate consistent performance across both evaluation methods (Claude Opus 4: 76.51% automated vs. 81.9% human), while other models show larger discrepancies. Gemini 2.5 Flash achieves strong automated scores (81.21%) with correspondingly high human evaluation (78.5%), suggesting robust reasoning capabilities that translate well across evaluation paradigms.

Reasoning Models Show Promise but at High Cost: Reasoning-focused models generally outperform others on both MCQs and essays but require significantly higher computational time (3.3x for MCQs, 2.1x for essays). This reinforces the quality-efficiency trade-off requiring careful cost-benefit analysis for real-world deployment.

Domain-Specific Models Offer Strategic Value: Palmyra-fin achieves respectable results with 68.3% MCQ accuracy and 58.39% essay score (Zero Shot) at low computational cost. Financial institutions can strategically deploy such models for routine analysis, reserving computationally intensive frontier models for complex reasoning scenarios.

Prompting Strategy ROI is Context-Dependent: Advanced prompting improves MCQ accuracy by 7.8 percentage points but at 11x cost increases. For essays, CoT-SC provides the best scores (60-61% mean) while Self-Discover significantly underperforms (46.32%). The cost-benefit analysis shows strategies should be chosen based on specific task needs and acceptable cost limits.

5.2 Limitations and Constraints

Dataset Representativeness: Our reliance on mock exam questions rather than official CFA materials may introduce stylistic differences or difficulty variations from actual exams.

LLM Self-Assessment Reliability: Our reliance on LLM self-grading introduces inherent limitations, as models may exhibit systematic biases in self-evaluation or lack calibration between confidence and actual performance quality.

Evaluation Limitations: API-level restrictions and reliability issues in some models complicated uniform benchmarking and exposed limitations in instruction adherence. Additionally, automated metrics cannot fully capture nuances like argumentation quality and professional communication standards that human experts would assess.

ROUGE-L Considerations: Observed ROUGE-L F1 scores are modest (ranging from 0.106 to 0.199) even for capable models, indicating limited lexical overlap with reference answers and highlighting needs for improved alignment in complex financial essay tasks.

5.3 Future Research

Building on our findings, several high-priority research directions emerge:

Human Expert Validation: Incorporating evaluations by certified financial professionals is essential to assess nuanced reasoning and validate model-generated content against industry standards. This step would enhance credibility and reveal qualitative gaps not captured by automated metrics.

Self-Assessment Calibration: As self-grading strategies like CoT-SC demonstrate internal consistency, further research is needed to improve the alignment between model confidence and external performance quality. Developing methods to calibrate self-assessment in financial reasoning contexts could enhance decision reliability in professional settings.

Cross-Certification Evaluation: Expanding evaluation frameworks to include other financial certifications such as FRM, CAIA, and CPA would test the generalizability of LLMs across diverse financial domains, broadening insights into their utility for specialized knowledge assessments.

6 Conclusion

Our key findings provide crucial guidance for practitioners: advanced prompting strategies can improve MCQ accuracy by approximately 7.8 percentage points but at substantial cost increases of 3-11x. For essays, Chain-of-Thought with Self-Consistency (CoT-SC) yields the highest Essay Scores (around 60-61% mean), while Self-Discover's performance declined. Reasoning models offer good accuracy but require careful cost-benefit analysis. Specialized financial models like Palmyra-fin demonstrate competitive MCQ performance (68.3%) and reasonable essay scores (e.g., 58.39% with Zero Shot) at lower computational cost, suggesting domain-specific training value.

As LLMs continue evolving rapidly, benchmarks like ours will be essential for tracking progress toward professional-grade financial AI capabilities. Our open-source evaluation framework[10] [11] provides a foundation for continued research into making LLMs more capable and cost-effective tools for financial professionals.

Ethical Statement

This research evaluates publicly available language models using educational materials, presenting no direct ethical concerns. However, our findings have important implications for responsible AI deployment in finance. The performance gaps we identified underscore the critical need for continued human oversight in financial decision-making, particularly given the fiduciary responsibilities involved in investment management. We emphasize that current LLMs should be viewed as assistive tools rather than autonomous decision-makers for financial applications.

Acknowledgments

This research was supported in part by GoodFin and New York University (NYU) Faculty Research. We thank the CFA Institute for their educational materials and mission, AnalystPrep for access to mock examination content, and the broader financial education community for maintaining rigorous professional standards that enable meaningful AI evaluation.

References

[1] 300Hours. CFA passing score: Mps estimates to help your prep. https://300hours.com/cfa-passing-score/, 2025. Estimates and analysis of CFA exam minimum passing scores (MPS) for all levels, including methodology and recent trends.

[2] J. Benton, M. Wagner, E. Christiansen, C. Anil, E. Perez, J. Srivastav, E. Durmus, D. Ganguli, S. Kravec, B. Shlegeris, J. Kaplan, H. Karnofsky, E. Hubinger, R. Grosse, S. R. Bowman, and D. Duvenaud. Sabotage evaluations for frontier models. arXiv preprint arXiv:2410.21514, 2024.

[3] Ethan Callanan et al. Can GPT models be financial analysts? An evaluation of ChatGPT and GPT-4 on mock CFA exams. arXiv preprint arXiv:2310.08678, 2023.

[4] CFA Institute. CFA program level III exam. https://www.cfainstitute.org/programs/cfa/exam/level-iii, 2024.

[5] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, and W. Y. Wang. A survey on large language models for critical societal domains: finance, healthcare, and law. arXiv preprint arXiv:2405.01769, 2024.

[6] A. De La Cruz. Multi-agent large language models for traditional finance and decentralized finance. *Journal of Industrial Engineering and Applied Science*, 2025.

[7] Robert Johnson. The CFA designation and the finance curriculum: A survey of faculty. *Survey*, 1999.

[8] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[9] A. Mahfouz et al. The state of the art of large language models on chartered financial analyst exams. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, EMNLP 2024, 2024.

[10] pranam gf. CFA_ESSAY_REPRODUCER: Codebase for CFA essay question answer generation and LLM benchmark. https://github.com/pranam-gf/CFA_ESSAY_REPRODUCER, 2024. Accessed: December 3, 2024.

[11] pranam gf. CFA_MCQ_REPRODUCER: Codebase for CFA multiple-choice question evaluation and LLM benchmark, 2024. Accessed: December 3, 2024.

[12] Redefine AT Stirling and Elaine Calleja. *Investment analysis*. Managerial Economics, 2022.

[13] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[14] Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. Self-discover: Large language models self-compose reasoning structures, 2024.