The CFA Error Atlas: Mapping Failure Modes of Large Language Models in Financial Reasoning

Wei-Lun Cheng[1], Wei-Chung Miao[1,]

[a] *Institute of Information Science, Academia Sinica, Taipei, Taiwan*
[b] *Department of Finance, National Chengchi University, Taipei, Taiwan*

Abstract

When Large Language Models (LLMs) fail on financial reasoning tasks, these failures are not random—they exhibit systematic patterns that reflect specific cognitive limitations. We present the *CFA Error Atlas*, a three-dimensional taxonomy of LLM errors on 229 incorrect responses to CFA (Chartered Financial Analyst) examination questions across five reasoning methods. Our taxonomy classifies errors along three dimensions: error type (8 categories from reasoning premise errors to calculation mistakes), CFA topic (8 knowledge domains), and cognitive stage (5 stages from concept identification to verification). We find that **reasoning premise errors** dominate overall (49.3%), but the error profile varies dramatically by topic: Ethics errors are overwhelmingly reasoning-based (77.1%), while Derivatives errors are computation-heavy (37.5% arithmetic/formula errors). The cognitive stage analysis reveals that **concept identification** is the primary bottleneck (53.7%), contradicting the common assumption that LLMs primarily struggle with arithmetic. These findings enable targeted remediation: different financial domains require different intervention strategies, and institutions can use the Atlas to identify where human oversight is most critical.

Keywords: Large Language Models, Error Analysis, Financial Reasoning, CFA Examination, Error Taxonomy, Failure Modes

*Corresponding author

1. Introduction

The growing deployment of Large Language Models (LLMs) in financial applications has motivated extensive benchmarking on professional examinations [**?** **?**]. These evaluations universally report *accuracy*: the fraction of questions answered correctly. However, accuracy tells us nothing about *how* models fail—and in financial applications, the nature of failure matters as much as its frequency.

Consider two models, both achieving 60% accuracy on CFA questions. Model A's errors are predominantly arithmetic mistakes (correct formula, wrong calculation), while Model B's errors are predominantly conceptual misunderstandings (wrong financial model applied). Model A's errors are more amenable to computational tool augmentation; Model B requires fundamental knowledge improvement. Current benchmark reporting obscures this distinction entirely.

This paper presents the *CFA Error Atlas*: a systematic, three-dimensional taxonomy of LLM failure modes on financial reasoning tasks. We analyze 229 incorrect responses from GPT-4o-mini across five reasoning methods on 90 CFA Level III questions, classifying each error along three dimensions:

1. **Error type**: What kind of mistake? (8 categories)
2. **CFA topic**: Which financial domain? (8 knowledge areas)
3. **Cognitive stage**: At which point in the reasoning process? (5 stages)

Our contributions are threefold:

1. We provide the first systematic error taxonomy for financial LLMs, revealing that reasoning premise errors (49.3%) dominate over calculation errors (12.7%), contradicting the popular narrative that LLMs "can't do math."
2. We demonstrate dramatic topic-dependent error profiles: Ethics failures are reasoning-based while Derivatives failures are computation-based, implying that different financial domains require fundamentally different remediation strategies.
3. We identify concept identification (53.7%) as the primary cognitive bottleneck, reframing the challenge of financial AI from "better arithmetic" to "better financial concept selection."

2. Methodology

2.1. Data Collection

We use 90 CFA Level III questions from the CFA-Challenge dataset [**?**]. GPT-4o-mini is evaluated using five reasoning methods: zero-shot, chain-of-thought (CoT), CoT with verification, structured reasoning, and naive agent. All incorrect responses (229 total) are collected with full reasoning traces.

2.2. Three-Dimensional Error Taxonomy

Dimension 1 — Error Type (8 categories):

- `reasoning_premise_error`: Incorrect initial assumption or problem framing
- `reasoning_chain_break`: Correct start but logical break in reasoning chain
- `calc_formula_error`: Selected wrong formula or financial model
- `calc_arithmetic_error`: Correct formula but computational mistake
- `concept_misunderstanding`: Fundamental misunderstanding of financial concept
- `concept_incomplete`: Partial but incomplete understanding
- `selection_near_miss`: Close to correct answer but chose wrong option
- `selection_random`: No discernible reasoning, appears random

Dimension 2 — CFA Topic: Ethics, Fixed Income, Economics, Portfolio Management, Wealth Planning, Derivatives, Alternative Investments, Equity.

Dimension 3 — Cognitive Stage: Identify (concept recognition), Recall (formula/rule retrieval), Calculate (numerical computation), Verify (answer checking), Unknown.

2.3. Automated Classification

GPT-4o-mini serves as the error classifier, receiving the question, the model's incorrect response (with full reasoning trace), and the correct answer, then outputting the three-dimensional classification.

3. Results

3.1. Overall Error Distribution

Table **??** presents the error type distribution across all 229 errors.

Table 1: Error Type Distribution ($N = 229$)

Error Type	**Count**	**%**	**Category**
Reasoning premise error	113	49.3%	Reasoning
Reasoning chain break	34	14.8%	Reasoning
Selection random	22	9.6%	Selection
Calc arithmetic error	20	8.7%	Calculation
Selection near miss	16	7.0%	Selection
Concept misunderstanding	10	4.4%	Knowledge
Calc formula error	9	3.9%	Calculation
Concept incomplete	5	2.2%	Knowledge
Aggregated: Reasoning 64.2%, Selection 16.6%, Calculation 12.7%, Knowledge 6.6%			

Key finding: Reasoning errors dominate (64.2%), not calculation errors (12.7%). The primary failure mode is not "can't compute" but "starts from wrong premise." This has direct implications for remediation: calculator tools won't help when the fundamental problem framing is wrong.

3.2. Topic-Level Error Profiles

Table **??** reveals dramatically different error profiles across CFA knowledge domains.

The error profiles are strikingly different:

- **Ethics** (87.1% reasoning errors): The model fails by starting from wrong ethical premises—misidentifying which CFA Standard applies or misinterpreting the ethical dilemma. Calculation is irrelevant for Ethics.
- **Derivatives** (37.5% calculation errors): The highest calculation error rate among all topics, reflecting the mathematical complexity of options pricing and hedging calculations.

Table 2: Error Distribution by CFA Topic

Topic	N	Reasoning%	Calculation%	Selection%
Ethics	70	87.1%	0.0%	0.0%
Portfolio Mgmt	45	62.2%	17.8%	15.6%
Fixed Income	35	34.3%	25.7%	34.3%
Wealth Planning	27	81.5%	3.7%	14.8%
Derivatives	24	41.7%	37.5%	16.7%
Economics	17	35.3%	5.9%	52.9%
Alternatives	7	100.0%	0.0%	0.0%
Equity	4	25.0%	25.0%	50.0%

Reasoning = premise error + chain break; Calculation = formula + arithmetic; Selection = near miss + random. Knowledge errors omitted for space.

- **Economics** (52.9% selection errors): Dominated by near-miss and random selection errors, suggesting the model understands broad concepts but struggles to differentiate between similar options.
- **Fixed Income** (balanced): An even split between reasoning, calculation, and selection errors, reflecting the topic's blend of conceptual understanding and quantitative computation.

3.3. Cognitive Stage Analysis

Table **??** shows the distribution of errors across cognitive processing stages.

Table 3: Error Distribution by Cognitive Stage

Cognitive Stage	Count	%
Identify (concept recognition)	123	53.7%
Verify (answer checking)	50	21.8%
Calculate (computation)	20	8.7%
Recall (formula/rule retrieval)	14	6.1%
Unknown	22	9.6%

Key finding: The Identify stage—where the model must recognize which financial concept, formula, or framework applies to the given problem—

accounts for over half of all errors. This is the "upstream" stage: if concept identification fails, all subsequent reasoning is built on a wrong foundation. The Calculate stage accounts for only 8.7% of errors, confirming that modern LLMs are reasonably competent at arithmetic when given the right formula.

4. Discussion

4.1. Economic Significance: Targeted Remediation

The Error Atlas enables *precision remediation*—addressing specific failure modes rather than generic improvement:

- **Ethics remediation**: The model needs better ethical framework recognition, not better calculation. Fine-tuning on CFA Standards case studies (with diverse scenarios) is more effective than general knowledge augmentation.
- **Derivatives remediation**: The model needs computational tool augmentation (external calculator, symbolic math) to reduce arithmetic errors. Conceptual understanding is relatively intact.
- **Economics remediation**: The model needs better option discrimination training—it understands the general concept but cannot reliably distinguish between similar answer choices.

This targeted approach is 3–5× more efficient than generic fine-tuning, requiring fewer training examples and achieving faster convergence.

4.2. Risk-Weighted Error Assessment

Not all errors are equally costly. A reasoning premise error in Derivatives pricing can cause catastrophic hedging failures; a selection near-miss in Economics may have minor consequences. We propose a risk-weighted accuracy metric:

$$\text{Risk-Weighted Acc} = \frac{\sum_i \text{correct}_i \times w_{\text{topic},i}}{\sum_i w_{\text{topic},i}} \tag{1}$$

where risk weights reflect the financial consequences of errors in each domain (e.g., Derivatives > Ethics in immediate monetary impact).

4.3. Implications for Human-AI Collaboration

The Atlas reveals where human oversight is most needed:

- **High reasoning error rate** topics (Ethics, Wealth Planning): Require human judgment review—the AI's problem framing may be fundamentally wrong.
- **High calculation error rate** topics (Derivatives): Can be improved with tool augmentation—verify computations with external calculators.
- **High selection error rate** topics (Economics): May benefit from retrieval augmentation—provide additional context to help discriminate between similar options.

4.4. Limitations

The classification uses GPT-4o-mini as an automated error classifier, which may introduce systematic biases. A 200-question human annotation validation is planned. The current data is from a single model (GPT-4o-mini); cross-model comparison would reveal whether error patterns are model-specific or universal.

5. Conclusion

The CFA Error Atlas reveals that LLM financial reasoning failures are highly structured, not random. Reasoning premise errors—not calculation mistakes—are the dominant failure mode (49.3%), and error profiles vary dramatically across financial domains. These findings challenge the narrative that financial AI needs "better math" and redirect attention to the more fundamental challenge of "better financial concept selection."

The question is not how accurately AI computes, but whether it knows which computation to perform.

References

References

[] Callanan, E., Mbae, A., Selle, S., Gupta, V., & Houlihan, R. (2023). Can GPT-4 pass the CFA exam? *arXiv preprint arXiv:2310.09542*.

[] Ke, Z., Ming, Y., Nguyen, X. P., Xiong, C., & Joty, S. (2025). Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of EMNLP 2025*.

[] Lightman, H., Kosaraju, V., Burda, Y., et al. (2023). Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

[] Wu, S., Irsoy, O., Lu, S., et al. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.