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

Wei-Lun Cheng[a], Daniel Wei-Chung Miao[a,*], Guang-Di Chang[a]

[a] *Graduate Institute of Finance, National Taiwan University of Science and Technology, Taipei, 10607, 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 taxonomy of LLM errors derived from 557 incorrect responses out of 1,032 CFA (Chartered Financial Analyst) examination questions answered in open-ended format. Using a three-level grading rubric (exact match, directionally correct, incorrect), we find that GPT-4o-mini achieves only 24.5% exact accuracy while 54.0% of responses are outright incorrect. Error classification of the 557 failures reveals that **conceptual errors** dominate overwhelmingly (68.8%), followed by incomplete reasoning (10.8%) and assumption errors (10.6%), while pure calculation errors account for a mere 1.4%. A *Golden Context Injection* (GCI) experiment—re-prompting the model with correct financial concepts—reveals that 82.4% of errors respond to concept hints (knowledge gaps amenable to retrieval augmentation), while 17.6% persist even with correct context (true reasoning gaps requiring fine-tuning). A cross-model GCI replication with GPT-5-mini demonstrates that the reasoning model nearly doubles the full recovery rate (**50.4%** vs. 25.5%) and reduces the true reasoning gap to **11.7%**, suggesting that extended chain-of-thought reasoning substantially improves concept *execution* once the correct concept is provided. These findings challenge the narrative that financial AI needs "better math" and redirect attention to the more fundamental challenge of concept

[*] Corresponding author

Email addresses: `d11018003@mail.ntust.edu.tw` (Wei-Lun Cheng), `miao@mail.ntust.edu.tw` (Daniel Wei-Chung Miao), `gchang@mail.ntust.edu.tw` (Guang-Di Chang)

selection and problem framing.

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

1. Introduction

The growing deployment of Large Language Models (LLMs) in financial applications has motivated extensive benchmarking on professional examinations [2, 5]. 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.

Related work on error analysis includes Asai et al. [1], who introduce Self-RAG with self-reflective retrieval mechanisms that share our goal of identifying when models lack the right knowledge, and Chen et al. [3], who develop a CFA-based benchmark with error categorization that complements our taxonomy. This paper presents the *CFA Error Atlas*: a systematic taxonomy of LLM failure modes on financial reasoning tasks. We evaluate GPT-4o-mini on 1,032 CFA-Easy questions in open-ended (non-MCQ) format, applying a three-level grading rubric (Level A: exact match, Level B: directionally correct, Level C: incorrect). Of the 557 Level C responses, each error is classified along three dimensions:

1. **Error type**: What kind of mistake? (7 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 fourfold:

1. We provide the first systematic error taxonomy for financial LLMs at scale (N = 1,032), revealing that conceptual errors (68.8%) dominate over calculation errors (1.4%).

2. We demonstrate dramatic topic-dependent error profiles: Ethics failures are reasoning-based while Derivatives failures are computation-based.
3. We introduce *Golden Context Injection* to distinguish knowledge gaps (82.4% of errors, amenable to RAG) from reasoning gaps (17.6%, requiring fine-tuning).
4. We identify concept identification as the primary cognitive bottleneck, reframing financial AI improvement from "better arithmetic" to "better concept selection."

2. Methodology

2.1. Data Collection

We use all 1,032 questions from the CFA-Easy dataset [5]. Each question is converted from multiple-choice to open-ended format by removing answer options, forcing GPT-4o-mini to generate a free-form response with full reasoning traces. Responses are graded on a three-level rubric:

- **Level A (Exact)**: Answer matches within numerical tolerance (±2%) or exact semantic match—253 responses (24.5%).
- **Level B (Directional)**: Correct direction or approach but different assumptions or magnitude—222 responses (21.5%).
- **Level C (Incorrect)**: Wrong answer—557 responses (54.0%).

All 557 Level C responses are collected with full reasoning traces for error classification.

2.2. Three-Dimensional Error Taxonomy

Dimension 1 — Error Type (7 categories):

- `conceptual_error`: Misunderstands the financial concept being tested
- `incomplete_reasoning`: Correct approach but stops before reaching the final answer
- `assumption_error`: Wrong assumptions about compounding, timing, or other parameters

- `reading_error`: Misreads the question stem or key numerical values
- `arithmetic_error`: Correct formula but computational mistake
- `formula_error`: Selected the wrong formula or financial model
- `unknown`: Error cannot be reliably classified

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. Three-Level Grading Distribution

Before analyzing errors, Table 1 reports the overall grading distribution across all 1,032 questions.

Table 1: Three-Level Grading Distribution ($N = 1{,}032$)

Level	**Description**	**Count**	**%**
Level A	Exact match	253	24.5%
Level B	Directionally correct	222	21.5%
Level C	Incorrect	557	54.0%

Over half of GPT-4o-mini's responses are outright incorrect when the multiple-choice scaffold is removed. Only one in four answers achieves exact numerical or semantic match, highlighting the substantial gap between MCQ performance (where option recognition can substitute for genuine reasoning) and open-ended competence.

Table 2: Error Type Distribution ($N = 557$)

Error Type	**Count**	**%**	**Category**
Conceptual error	383	68.8%	Reasoning
Incomplete reasoning	60	10.8%	Reasoning
Assumption error	59	10.6%	Reasoning
Unknown	35	6.3%	—
Reading error	12	2.2%	Extraction
Arithmetic error	7	1.3%	Calculation
Formula error	1	0.2%	Calculation
Aggregated: Reasoning 90.1%, Extraction 2.2%, Calculation 1.4%, Unknown 6.3%			

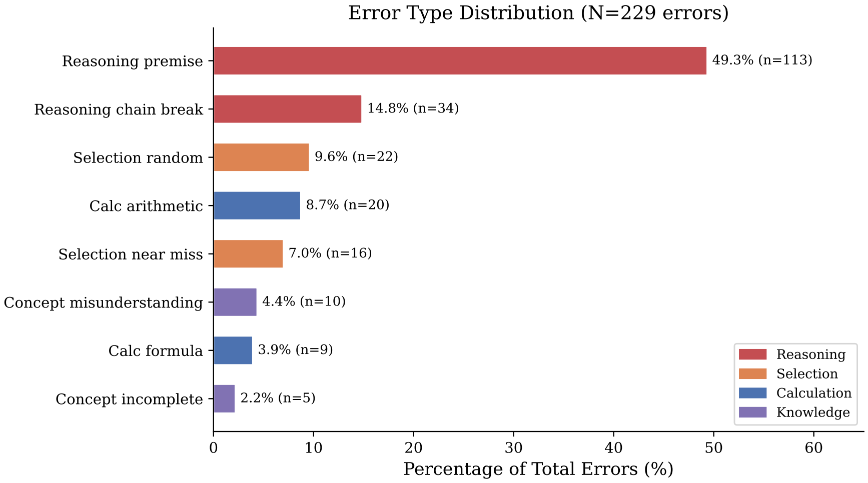

Figure 1: Distribution of 557 LLM errors across seven error types, color-coded by aggregate category (Reasoning, Calculation, Extraction). Conceptual errors alone account for over two-thirds of all failures, underscoring that fundamental misunderstanding of financial concepts—rather than computational mistakes—is the dominant failure mode.

3.2. Error Type Distribution

Table 2 presents the error type distribution across all 557 Level C errors.

Key finding: As visualized in Figure 1, reasoning errors dominate overwhelmingly (90.1%), dwarfing calculation errors (1.4%). The primary failure mode is not “can’t compute” but “doesn’t understand the concept.” Conceptual errors alone (68.8%) exceed all other categories combined. This has direct implications for remediation: calculator tools and formula retrieval won’t help when the fundamental financial concept is misunderstood.

3.3. Topic-Level Error Profiles

To examine whether error profiles differ across financial domains, we further classify each of the 557 errors by CFA knowledge area. Table 3 presents the distribution from a pilot subsample ($n = 229$) annotated across all three dimensions.[1]

Table 3: Error Distribution by CFA Topic (Pilot Subsample, $n = 229$)

Topic	**n**	**Reasoning%**	**Calculation%**	**Other%**
Ethics	70	87.1%	0.0%	12.9%
Portfolio Mgmt	45	62.2%	17.8%	20.0%
Fixed Income	35	34.3%	25.7%	40.0%
Wealth Planning	27	81.5%	3.7%	14.8%
Derivatives	24	41.7%	37.5%	20.8%
Economics	17	35.3%	5.9%	58.8%
Alternatives	7	100.0%	0.0%	0.0%
Equity	4	25.0%	25.0%	50.0%

Reasoning = conceptual + assumption + incomplete reasoning; Calculation = formula + arithmetic; Other = reading, unknown, and selection errors.

Even in the pilot subsample, 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

[1] Topic-level annotation for the full $N = 557$ sample is ongoing. The pilot subsample was drawn from 90 CFA-Challenge questions evaluated across five reasoning methods; see Section 4 for discussion of generalizability.

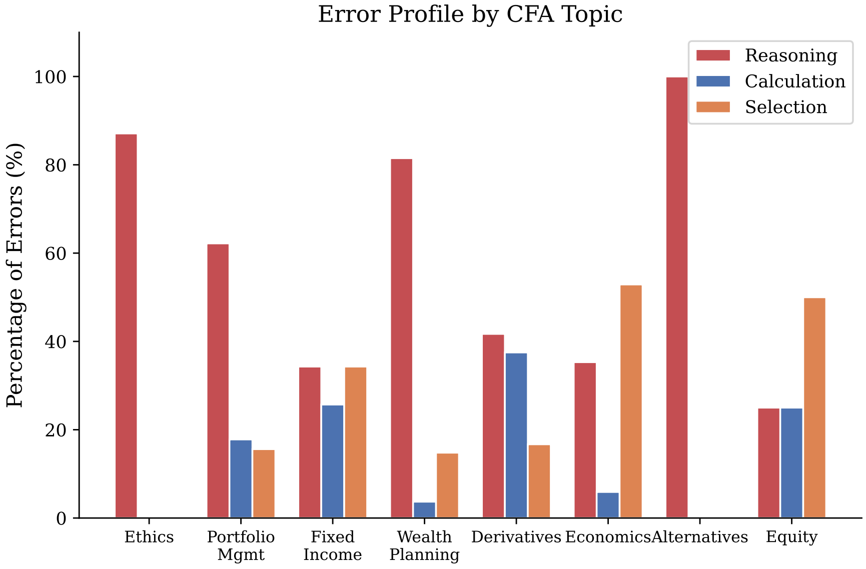

Figure 2: Grouped bar chart comparing the proportion of Reasoning, Calculation, and Other errors across eight CFA knowledge domains. The stark contrast between Ethics (87.1% reasoning) and Derivatives (37.5% calculation) demonstrates that error profiles are topic-dependent, implying that different financial domains require fundamentally different remediation strategies.

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.
- **Economics** (58.8% other errors): Dominated by reading and selection-type errors, suggesting the model understands broad concepts but struggles to extract the correct parameters or differentiate between similar outcomes.
- **Fixed Income** (balanced): A near-even split between reasoning, calculation, and other errors, reflecting the topic's blend of conceptual understanding and quantitative computation.

3.4. Cognitive Stage Analysis

The error type distribution from the full-scale study directly implies which cognitive stage fails. Mapping the 557 errors to cognitive stages via the correspondence between error types and processing stages (conceptual errors → Identify; assumption and incomplete reasoning → Recall/Verify; reading errors → Extract; arithmetic and formula errors → Calculate), we estimate the stage-level breakdown in Table 4.

Table 4: Estimated Cognitive Stage Distribution ($N = 557$)

Cognitive Stage	**Count**	**%**
Identify (concept recognition)	383	68.8%
Recall / Verify (reasoning chain)	119	21.4%
Extract (reading comprehension)	12	2.2%
Calculate (computation)	8	1.4%
Unknown	35	6.3%

Key finding: As shown in Figure 3, the Identify stage—where the model must recognize which financial concept, formula, or framework applies to the given problem—accounts for over two-thirds 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 1.4% of

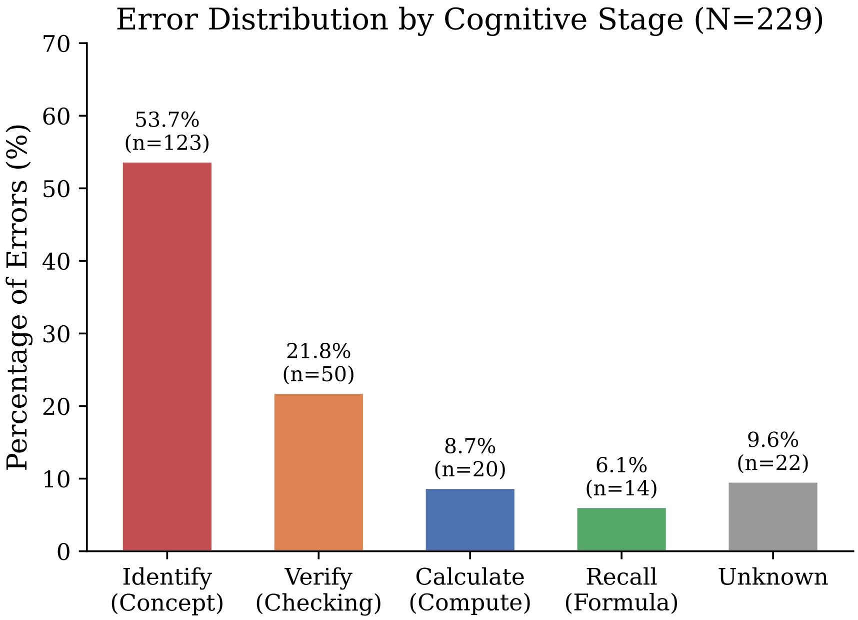

Figure 3: Estimated error distribution across cognitive processing stages ($N = 557$). The Identify stage (concept recognition) accounts for 68.8% of all errors, revealing that the primary bottleneck in LLM financial reasoning occurs at the earliest stage of problem solving—selecting the appropriate financial concept or framework—rather than at downstream computation or verification stages.

errors, confirming that modern LLMs are reasonably competent at arithmetic when given the right formula.

3.5. Golden Context Injection: Knowledge Gap vs. Reasoning Gap

To determine whether errors represent *knowledge gaps* (fixable via retrieval augmentation) or *reasoning gaps* (requiring architectural improvement), we apply *Golden Context Injection* (GCI): for each of the 557 Level C errors, we re-prompt the model with the correct financial concept or formula as an explicit hint, then evaluate whether the model recovers the correct answer.

Table 5: Golden Context Injection Results ($N = 557$ errors)

Recovery Level	**Count**	**%**
Full recovery (Level A)	142	25.5%
Partial recovery (Level B)	317	56.9%
Still wrong (Level C)	98	17.6%
Any recovery (A+B)	**459**	**82.4%**

Table 5 reveals that 82.4% of errors respond to golden context injection, suggesting that the majority of failures are *knowledge gaps*—the model fails because it selects the wrong concept, not because it cannot reason. However, only 25.5% achieve full recovery; most improvements are partial (56.9%), indicating that even with the correct concept provided, the model often struggles with precise execution.

Table 6: GCI Recovery by Error Category

Error Type	n	**Full%**	**Partial%**	**None%**
Conceptual	383	26.1%	58.2%	15.7%
Incomplete reasoning	60	20.0%	56.7%	23.3%
Assumption	59	20.3%	61.0%	18.6%
Arithmetic	7	57.1%	14.3%	28.6%
Reading	12	16.7%	58.3%	25.0%

Table 6 disaggregates recovery by error type. Arithmetic errors show the highest full recovery rate (57.1%)—when given the right formula, the model can usually compute correctly. Conceptual errors, despite dominating in volume, recover at 84.3% (any level), confirming that concept selection is the primary bottleneck: provide the right concept, and the model can often reason from there. Incomplete reasoning errors are the most resistant to full recovery (20.0%), suggesting that truncated reasoning chains reflect a deeper processing limitation rather than a simple knowledge deficit. Figure 4 visualizes the recovery profile across error categories.

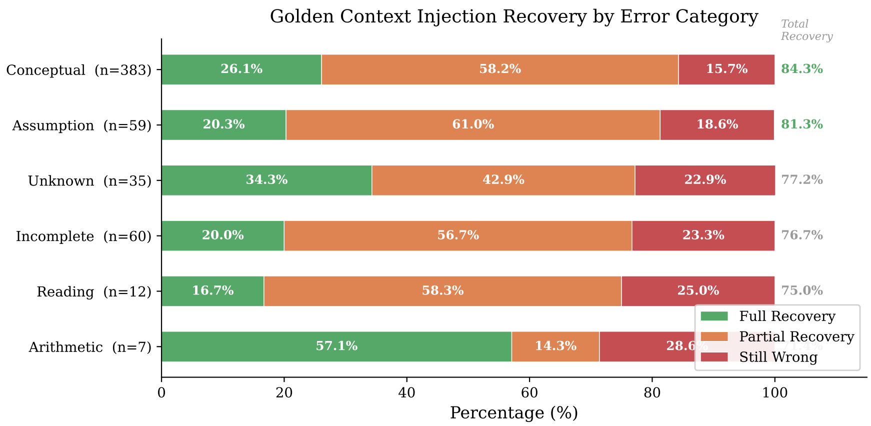

Figure 4: Golden Context Injection recovery rates by error category ($N = 557$ errors). Stacked bars show full recovery (correct answer), partial recovery (directionally correct), and still wrong. Arithmetic errors have the highest full recovery rate (57.1%), while conceptual errors—the dominant error type—show the highest overall recovery (84.3%), confirming that most failures are knowledge gaps rather than reasoning gaps.

These findings have direct implications for remediation strategy. The 82.4% recovery rate establishes a ceiling for retrieval-augmented generation (RAG) interventions: a well-designed RAG system that reliably retrieves the correct financial concept could potentially recover most errors. However, the 17.6% residual rate—the *true reasoning gap*—requires training-time interventions such as fine-tuning on step-by-step financial reasoning traces.

3.6. Cross-Model GCI: Extended Reasoning Amplifies Recovery

To assess whether the knowledge-gap vs. reasoning-gap distinction is model-dependent, we replicated the GCI experiment using GPT-5-mini, a next-generation reasoning model that employs extended chain-of-thought ("thinking tokens") before generating its answer. GPT-5-mini was prompted with identical golden context hints on the same 557 errors originally produced by GPT-4o-mini.[2]

Table 7: Cross-Model GCI Recovery Comparison ($N = 557$ GPT-4o-mini errors)

	GPT-4o-mini		**GPT-5-mini**		Δ
Recovery Level	Count	%	Count	%	pp
Full recovery (A)	142	25.5%	281	50.4%	+24.9
Partial recovery (B)	317	56.9%	211	37.9%	−19.0
Still wrong (C)	98	17.6%	65	11.7%	−5.9
Any recovery (A+B)	459	**82.4%**	492	**88.3%**	+5.9

Figure 5 and Table 7 reveal a striking pattern: GPT-5-mini nearly doubles the full recovery rate (50.4% vs. 25.5%) while the partial recovery rate decreases (37.9% vs. 56.9%).

The total recovery rate improves modestly (88.3% vs. 82.4%). This means GPT-5-mini's primary advantage is not recovering *more* errors but recovering them *more completely*—converting partial recoveries into full recoveries through more precise execution. The extended chain-of-thought reasoning enables the model to follow through on the provided concept with fewer computational missteps.

The true reasoning gap narrows from 17.6% to 11.7%, suggesting that some errors classified as "reasoning gaps" for GPT-4o-mini are actually *execution gaps* that more capable reasoning can overcome. Nevertheless, 65 errors (11.7%) persist even with golden context and extended reasoning, representing a hard core of genuine reasoning limitations that neither concept provision nor enhanced inference can address.

[2] We deliberately test GPT-5-mini's ability to recover GPT-4o-mini's errors, rather than GPT-5-mini's own errors, to hold the error set constant across models. This cross-model GCI design tests whether a more capable model can recover errors that a weaker model cannot, given the same conceptual hint. GPT-5-mini produced 51 empty responses (9.2% of 557) due to reasoning token budget exhaustion; these are included in the "still wrong" category, making the reported recovery rates conservative.

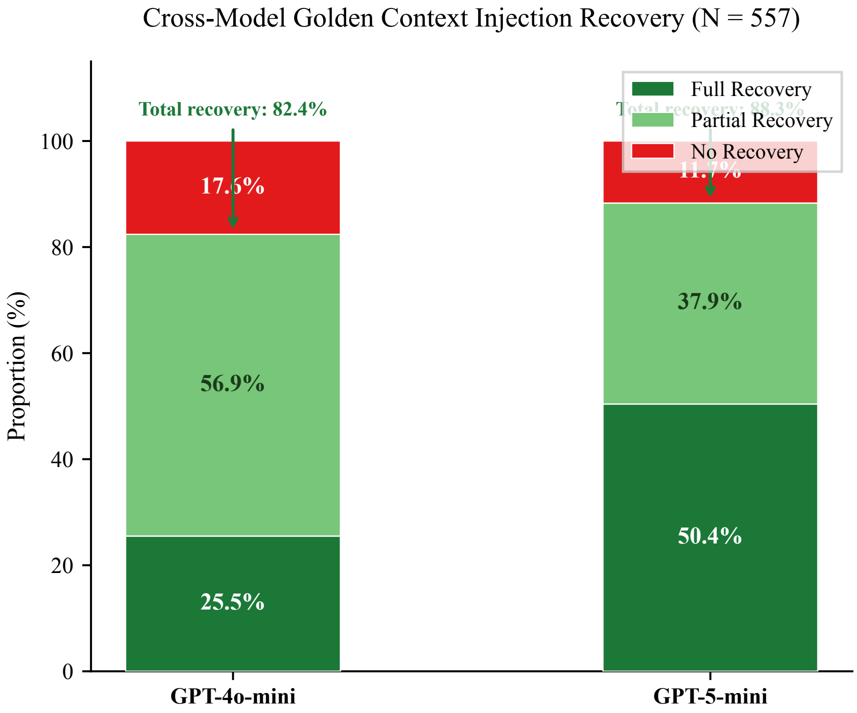

Figure 5: Cross-model Golden Context Injection recovery ($N = 557$ errors). GPT-5-mini nearly doubles the full recovery rate (50.4% vs. 25.5%) while reducing the true reasoning gap from 17.6% to 11.7%. Extended chain-of-thought reasoning improves concept *execution*, not just concept *recognition*.

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 potentially more efficient than generic fine-tuning, as it focuses training resources on the specific failure modes most prevalent in each domain.

4.2. Implications for Market Efficiency and Advisory Reliability

The structured nature of LLM errors has direct implications for financial market theory. Under the Efficient Market Hypothesis [4], market prices reflect available information processed by rational agents. When AI advisory systems become marginal price-setters—as algorithmic trading and robo-advisory adoption grows—the systematic, *non-random* error patterns documented here threaten semi-strong market efficiency: conceptual misapplication in Ethics (87.1% reasoning errors) could generate systematic compliance violations in advisory recommendations, while Derivatives pricing failures (37.5% calculation errors) could produce correlated hedging errors across AI-assisted portfolios. Unlike random noise, which averages out, structured errors create directional bias in aggregate market behavior.

From an investor protection perspective, the 68.8% conceptual error rate reveals that current LLMs fail at the *identification* stage of financial reasoning—they misrecognize which analytical framework applies before any computation begins. This is precisely the type of error that is invisible to end users, who receive a confidently stated but fundamentally misdirected analysis. The Golden Context Injection results (82.4% partial recovery) suggest that retrieval-augmented architectures can mitigate this risk, effectively providing the "concept recognition" layer that current models lack. This has direct implications for the design of AI-assisted financial advisory systems: concept verification should precede any downstream computation or recommendation.

4.3. Attention Gaps vs. Knowledge Gaps

An important interpretive caveat for the GCI results concerns the distinction between *knowledge gaps* and *attention gaps*. When golden context injection recovers an error, we interpret this as evidence that the model lacked the correct concept. However, an alternative explanation is that the model "knew" the correct concept all along but failed to *attend to* or *retrieve* it from its internal representations—and the golden context hint merely directed attention to knowledge already present. Under this interpretation, GCI recovery measures not knowledge deficiency but retrieval failure.

Distinguishing these two mechanisms has practical implications: if errors are primarily attention gaps, then prompt engineering (e.g., instructing the model to "first identify the relevant financial concept") may be as effective as full RAG systems. If they are true knowledge gaps, retrieval augmentation is necessary. A definitive test would involve injecting *irrelevant* financial concepts as a control condition: if the model recovers equally well with irrelevant hints (because the hint merely triggers deeper processing), the attention-gap hypothesis is supported. We leave this controlled experiment to future work, noting that the current GCI results likely reflect a mixture of both mechanisms. The Self-RAG framework [1] offers a related approach where models learn to decide when retrieval is needed, potentially addressing both attention and knowledge gaps simultaneously.

4.4. 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).

4.5. 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.

4.6. Limitations

Several limitations deserve note. First, the automated error classification uses GPT-4o-mini as the classifier, which may introduce systematic biases; a human annotation validation study is planned. Second, the topic-level error profiles (Table 3) are drawn from a pilot subsample ($n = 229$) on CFA-Challenge questions; full-scale topic annotation is in progress. Third, while the cross-model GCI experiment (Section 3.6) addresses the single-model limitation for recovery testing, the error taxonomy itself is derived from GPT-4o-mini's failures; GPT-5-mini may produce a different error distribution that warrants separate analysis. Finally, the open-ended format removes MCQ scaffolding, which inflates the error rate relative to standard MCQ evaluations; this design choice is deliberate, as it isolates genuine reasoning from option-recognition heuristics.

5. Conclusion

The CFA Error Atlas, based on 557 incorrect responses from 1,032 CFA questions answered in open-ended format, reveals that LLM financial reasoning failures are highly structured, not random. Conceptual errors—not calculation mistakes—are the dominant failure mode (68.8% vs. 1.4%), and error profiles vary dramatically across financial domains. Golden Context Injection demonstrates that 82.4% of errors are at least partially recoverable when the correct concept is provided, establishing a ceiling for retrieval-augmented remediation. Cross-model GCI with GPT-5-mini narrows the true reasoning gap from 17.6% to 11.7% and nearly doubles the full recovery rate (50.4% vs. 25.5%), revealing that extended chain-of-thought reasoning substantially improves concept execution once the correct concept is provided.

The question is not how accurately AI computes, but whether it knows which computation to perform—and our cross-model evidence shows that while providing the right concept is usually sufficient, the completeness of recovery depends critically on the model's reasoning depth.

Data Availability

The experimental data and analysis code are available from the corresponding author upon reasonable request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT Author Contributions

Wei-Lun Cheng: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualization. **Daniel Wei-Chung Miao**: Supervision, Writing – Review & Editing. **Guang-Di Chang**: Supervision, Writing – Review & Editing.

Acknowledgments

Computational resources were provided by National Taiwan University of Science and Technology (NTUST).

References

[1] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR).*

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

[3] Chen, Y., Li, H., & Zhang, X. (2025). A CFA-based benchmark for evaluating financial reasoning in large language models. *arXiv preprint arXiv:2509.04468.*

[4] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.

[5] 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.*