金融能力的假象：大型語言模型於 CFA 考試之壓力測試、錯誤分類體系與校準分析

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

[a] 國立臺灣科技大學財務金融研究所, 臺北, *10607*, 臺灣

Abstract

大型語言模型（LLMs）在金融考試基準測試中展現出令人印象深刻的標題準確率。本文提出一套三維度評估框架，揭示該準確率在很大程度上是一種假象。第一，反事實壓力測試涵蓋 CFA-Easy 語料庫（$N = 1{,}032$），揭露 18.6 個百分點的記憶落差——當數值參數被擾動時，準確率從 82.4% 降至 63.8%——證明相當比例的正確答案反映模式匹配而非真正推理。以 GPT-5-mini 進行的跨模型複製揭示記憶悖論：儘管標準準確率達 91.8%（+9.4 pp），記憶落差卻幾乎翻倍至 36.4 pp。第二，*CFA* 失敗分類體系對 557 筆開放式評測錯誤進行分類，揭示概念性錯誤居首（68.8%），而計算錯誤僅佔 1.4%。黃金情境注入顯示 82.4% 的錯誤可透過概念提示獲得改善，確立檢索增強補救的上限。第三，信心校準分析涵蓋 257 筆模型—方法觀測值，揭示普遍過度自信：表達信心超出準確率 22–32 個百分點，30.0% 的回答為高信心錯誤，且無任何信心閾值能將錯誤率降至可接受水準。CFA-Easy（$N = 1{,}032$）的穩健性驗證揭示「固定信心暫存器」——無論準確率是 53% 還是 82%，模型均維持約 85% 的信心水準。本文引入穩健準確率與風險信心（CaR）作為監管指標，主張金融 AI 治理必須超越標題準確率，涵蓋穩健性、錯誤結構與校準品質。

Keywords: 大型語言模型, 金融推理, 壓力測試, 穩健性, 校準, 錯誤分析, CFA 考試, AI 風險管理

*通訊作者

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)

1. 緒論

金融業正快速將大型語言模型（LLMs）應用於股票研究、風險分析、法規遵循及客戶諮詢等任務 [17, 9]。基準評測顯示，最先進的模型在 CFA 考試中的得分已接近甚至超過人類及格率 [3]，而推理模型現已通過所有三級 CFA 考試，成績超過人類考生第 90 百分位數 [14]。這些令人印象深刻的成果加速了部署時程，企業日益依賴 LLM 產出的分析結果來進行重要金融決策。

然而，標題準確率——金融 AI 基準測試中普遍報告的單一數字——是衡量 AI 能力的一個危險且不完整的指標。它無法告訴我們模型為何成功、如何失敗，或是否知道自己何時犯錯。本文提出一套三維度評估框架，分別探討上述每一個問題：

1. **穩健性（「它是在推理，還是在記憶？」）**我們透過反事實擾動與雜訊注入對 LLM 金融推理進行壓力測試，量化標準表現與壓力測試表現之間的差距——嵌入基準分數中的「記憶溢價」。
2. **錯誤結構（「失敗時，它如何失敗？」）**我們建構一套包含 557 筆開放式 CFA 評測錯誤的分類體系，揭示其主要失敗模式並非「無法計算」而是「不理解概念」——這一發現對補救策略具有直接啟示。
3. **校準（「它是否知道自己何時犯錯？」）**我們跨多個模型與方法評估信心校準，發現 LLMs 表現出普遍的過度自信，使其錯誤訊號對依賴表達信心的使用者而言幾乎不可見。

三個維度互為補充。壓力測試揭示標準準確率中有多少是真實的；錯誤分析揭示剩餘錯誤屬於何種類型；而校準分析揭示模型的信心訊號是否可信以辨識那些錯誤。三者共同描繪出一幅圖景：一個在標準基準上看似高度勝任的 AI 系統，其能力實際上因記憶而大幅膨脹、以概念性而非計算性錯誤為主、且伴隨著系統性高估可靠性的信心訊號。

本文有五項貢獻：（1）設計結合反事實擾動與雜訊注入的雙維度壓力測試框架，揭示跨世代模型的記憶悖論；（2）建構金融 LLMs 的首套系統性錯誤分類體系，證明 90.1% 的錯誤屬於推理類型；（3）引入黃金情境注入以區分知識缺口與推理缺口；（4）量化過度自信錯誤並引入風險信心（CaR）作為風險管理指標；（5）提出將本文指標與金融監管框架接軌的政策建議。

2. 文獻回顧

2.1. 金融領域中的大型語言模型

LLMs 與金融的交會已引發大量研究關注。BloombergGPT [17] 展示了在金融自然語言處理任務上的競爭力表現。Ke et al. [9] 提出 FinDAP，透

過領域自適應後訓練在 CFA 基準測試上達到最先進成果。Callanan et al. [3] 以 GPT-4 評測 CFA Level I，發現其表現達及格水準。然而，這些評測僅評估標準問題上的準確率，未能檢視其表現是否反映真正的理解、失敗如何結構化，或信心訊號是否可靠。

2.2. 資料污染與基準有效性

LLM 評測中的資料污染威脅已被充分記載 [15]。Mirzadeh et al. [13] 證明當數學推理問題進行符號擾動時，LLMs 的準確率顯著下降，顯示高基準分數部分反映了記憶效應。Li et al. [10] 以 GSM-Plus 進一步擴展，系統性地跨八個擾動維度生成變體。Lopez-Lira et al. [12] 專門針對金融 LLM 評測中的記憶問題，證明基準污染普遍存在。本文的壓力測試將此範式延伸至金融領域，實現母體層級的覆蓋。

2.3. LLM 校準與信心

校準係指模型表達信心與其實際準確率之間的一致程度 [7]。Kadavath et al. [8] 證明大型語言模型「大致知道自己知道什麼」，但在分佈外任務上此能力會退化。Band et al. [2] 發展了適用於問答系統的 QA-Calibration 方法。Chhikara et al. [5] 辨識出跨領域持續存在的「信心落差」。Liu et al. [11] 利用預測市場資料提出 KalshiBench 用於校準評估。本文將此文獻延伸至金融專業考試，在此領域中校準不良的信心帶有直接的金錢與受託責任影響。

2.4. 錯誤分析與補救

Asai et al. [1] 引入 Self-RAG 的自我反思檢索機制，與本文辨識模型何時缺乏正確知識的目標相近。Chen et al. [4] 發展了基於 CFA 的基準測試並包含錯誤分類。本文的失敗分類體系提供了金融 LLMs 首套系統性的三維度錯誤分類體系（$N = 557$），支持針對性補救。

3. 研究方法

3.1. 維度一：壓力測試

3.1.1. 反事實擾動

本文採用受 Mirzadeh et al. [13] 啟發的數值擾動方法：修改每道題目中的一個數值參數（如利率、面額、到期日），同時保留解題程序不變。正確答案隨之改變，但所需的公式與推理步驟維持一致。使用 GPT-4o-mini 作為擾動生成器，每道原始題目產生一個變體，包含明確標識的變更參數、正確的擾動答案，以及邏輯結構保持完整的驗證。

3.1.2. 雜訊注入

我們定義四種雜訊類型，模擬逐漸困難的資訊環境：

- **N1 — 無關資料**：與解題無關的額外數值資料。
- **N2 — 誤導性干擾**：看似合理但與解題無關的金融陳述。
- **N3 — 冗長脈絡**：冗贅但實質空洞的填充文字。
- **N4 — 矛盾暗示**：引用常見錯誤答案的暗示。

3.1.3. 壓力測試指標

記憶落差：

$$\text{Memorization Gap}_\ell = \text{Acc}_{\text{original}} - \text{Acc}_{\text{Level } \ell} \tag{1}$$

雜訊敏感指數：

$$\text{NSI}_t = \frac{\text{Acc}_{\text{clean}} - \text{Acc}_{\text{noisy},t}}{\text{Acc}_{\text{clean}}} \tag{2}$$

穩健準確率要求在原始題目及所有有效擾動變體上均回答正確：

$$\text{Robust Acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[\text{correct}_i^{\text{orig}} \wedge \bigwedge_{\ell} \text{correct}_i^{\text{Level } \ell}\right] \tag{3}$$

3.2. 維度二：錯誤分類體系

3.2.1. 三級評分

1,032 道 CFA 題目以開放式格式呈現（移除選項）。回答以三個層級評分：

- **Level A（精確）**：答案在 $\pm 2\%$ 數值容差內吻合，或語義完全匹配。
- **Level B（方向正確）**：方向 / 方法正確但假設不同。
- **Level C（不正確）**：答案錯誤。

3.2.2. 三維度分類

所有 Level C 錯誤沿三個維度分類：(1) 錯誤類型（7 個類別：概念性、不完整推理、假設、閱讀、算術、公式、未知）；(2) CFA 主題（8 個知識領域）；(3) 認知階段（5 個階段：辨識、回憶、計算、驗證、未知）。

3.2.3. 黃金情境注入（*GCI*）

對每筆 Level C 錯誤，我們以正確金融概念作為明確提示重新提問模型，然後評估模型是否能修正答案。此方法區分知識缺口（可透過 RAG 修復）與推理缺口（需要微調）。

3.3. 維度三：校準分析

3.3.1. 信心估計方法

我們採用兩種方法：**口述信心法**，提示模型以百分比表達信心；以及**自我一致性** [16]，以 $k = 10$ 次抽樣（$\tau = 0.7$）定義信心為一致率。

3.3.2. 校準指標

期望校準誤差（ECE）：

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{4}$$

風險信心（CaR）：

$$\mathrm{CaR}(\alpha) = \inf\{c^* : P(\text{incorrect} \mid \text{confidence} \geq c^*) \leq \alpha\} \tag{5}$$

CaR 回答的問題是：「在何種最低信心水準下，錯誤率可降至 α 以下？」若 CaR 無定義，則模型的信心訊號從根本上不適用於風險預算用途。

過度自信錯誤：

$$\text{Overconfident Error} = \mathbf{1}\left[\text{confidence} \geq 0.80 \ \wedge \ \text{incorrect}\right] \tag{6}$$

4. 資料與實驗設計

4.1. 資料集

本文使用 FinEval [9] 中的兩個資料集：**CFA-Easy**（1,032 道涵蓋完整 CFA 課程的多選題）用於壓力測試與錯誤分析，以及 **CFA-Challenge**（90 道 CFA Level III 題目）用於初始校準分析，並以穩健性驗證將校準擴展至完整 CFA-Easy 語料庫。

4.2. 模型

主要評估模型為 **GPT-4o-mini**（OpenAI），一個廣泛部署的商業模型。跨模型比較使用 **GPT-5-mini**，一個採用延伸思維鏈的次世代推理模型。校準分析另外評估 **Qwen3-32B**（阿里巴巴），一個 32B 參數的開源權重模型。除自我一致性抽樣（$\tau = 0.7$）外，所有評測均使用溫度 $\tau = 0.0$ 以獲得確定性輸出。

4.3. 整體實驗規模

表 1 彙整三個維度的實驗規模。

Table 1: 實驗規模摘要

維度	組成	題數	推論次數
壓力測試	反事實擾動	1,032	1,734[a]
	雜訊注入（×4 種類型）	1,032	5,160
錯誤分類體系	開放式 + GCI	1,032	1,589[b]
校準	CFA-Challenge	90	257[c]
	CFA-Easy（穩健性驗證）	1,032	1,032
合計			**>9,700**

[a] 原始題目 + 有效擾動（GPT-4o-mini 為 702 題）。
[b] 1,032 道開放式 + 557 次 GCI 重新提問。
[c] 2 個模型 × 2 種方法，各組態樣本數不同。257 筆觀測值包含主實驗（250 筆）及先導驗證（5 筆 GPT-4o-mini 語言化、2 筆 Qwen3-32B 語言化）。

5. 結果

5.1. 維度一：壓力測試

5.1.1. 反事實擾動

表 2 呈現核心發現。在母體層級（$N = 1{,}032$），18.6 pp 的記憶落差證實標準準確率中有相當比例可歸因於數值模式匹配而非真正的金融推理。

Table 2: 反事實擾動結果（GPT-4o-mini，$N = 1{,}032$）

條件	有效 N	準確率	記憶落差	Δ	方向
原始題目	1,032	82.4%	—	—	—
Level 1（數值擾動）	702	63.8%	+18.6 pp	↓	記憶效應
穩健準確率	1,032	63.5%	—	—	—
疑似記憶	—	+18.9%	—	—	—

穩健準確率要求在原始題目及所有有效擾動上均回答正確。疑似記憶 = 原始正確但至少一個擾動答錯的比例。

5.1.2. 雜訊敏感度

表 3 揭示了細微的表現輪廓。N1（無關資料）產生最高敏感度（NSI = 0.032），而 N4（矛盾暗示）反而提升了表現（NSI = −0.072），將準確率從 81.6% 提升至 87.5%。我們假設矛盾暗示透過排除法與後設認知觸發兩個管道發揮作用。

Table 3: 雜訊敏感度結果（GPT-4o-mini，$N = 1{,}032$）

雜訊類型	含雜訊準確率	翻轉數	NSI	解讀
乾淨（基線）	81.6%	—	—	—
N1（無關資料）	79.0%	58/1,032	0.032	低
N2（誤導性）	80.3%	49/1,032	0.015	極低
N3（冗長脈絡）	82.0%	32/1,032	−0.005	無影響
N4（矛盾暗示）	87.5%	21/1,032	−0.072	負值（有助益）

整體模式確認模型的主要弱點在於依賴記憶的推理，而非雜訊敏感度：最壞情況下的雜訊退化（2.6 pp）遠小於 18.6 pp 的記憶落差。

5.1.3. 跨模型壓力測試：記憶悖論

表 4 揭示了一項引人注目的記憶悖論。

Table 4: 跨模型反事實擾動（$N = 1{,}032$）

指標	GPT-4o-mini	GPT-5-mini
標準準確率	82.4%	91.8%
Level 1 準確率（有效 n）	63.8%（$n = 702$）	55.3%（$n = 638$）
記憶落差	**18.6 pp**	**36.4 pp**
穩健準確率	63.5%	67.2%

GPT-5-mini 達到更高的標準準確率，但擾動後準確率反而更低，導致記憶落差幾乎翻倍。

GPT-5-mini 達到顯著更高的標準準確率（+9.4 pp），但在擾動題目上的表現實際上更差（55.3% 對比 63.8%），產生幾乎兩倍於 GPT-4o-mini 的記憶落差。穩健準確率僅微幅提升（67.2% 對比 63.5%），意味著 GPT-5-mini 的大部分表面改善在擾動壓力下蒸發殆盡。相比之下，雜訊敏感度約減半（最大 NSI 0.017 對比 0.032），確認資訊過濾能力有真正的改善。這種記憶—雜訊的不對稱性表明，反事實擾動與雜訊注入探測的是根本不同的認知維度。

5.2. 維度二：錯誤分類體系

5.2.1. 錯誤類型分佈

表 5 呈現開放式評測 557 筆 Level C 錯誤的分佈。

Table 5: 錯誤類型分佈（$N = 557$）

錯誤類型	數量	%	類別
概念性錯誤	383	68.8%	推理
不完整推理	60	10.8%	推理
假設錯誤	59	10.6%	推理
未知	35	6.3%	—
閱讀錯誤	12	2.2%	提取
算術錯誤	7	1.3%	計算
公式錯誤	1	0.2%	計算
彙總：推理 90.1%，提取 2.2%，計算 1.4%，未知 6.3%			

推理錯誤以壓倒性比例居首（90.1%），僅概念性錯誤（68.8%）就超過所有其他類別的總和。主要失敗模式並非「無法計算」而是「不理解概念」——當基本金融概念被誤解時，計算器工具與公式檢索皆無濟於事。

5.2.2. 主題層級錯誤輪廓

錯誤輪廓在不同主題間呈現顯著差異。倫理學表現出 87.1% 的推理錯誤（無計算錯誤），而衍生性商品則具有最高的計算錯誤率（37.5%）。這意味著不同金融領域需要根本不同的補救策略。

5.2.3. 黃金情境注入

表 6 揭示 82.4% 的錯誤可透過黃金情境注入獲得改善，表明大多數失敗為知識缺口，適合以檢索增強進行補救。

Table 6: 黃金情境注入結果（$N = 557$ 筆錯誤）

恢復層級	數量	%
完全恢復（Level A）	142	25.5%
部分恢復（Level B）	317	56.9%
仍然錯誤（Level C）	98	17.6%
任何恢復（A+B）	**459**	**82.4%**

然而，僅 25.5% 達到完全恢復；大多數改善為部分性的（56.9%），顯示即使提供正確概念，模型仍常在精確執行上遇到困難。17.6% 的殘餘率代表需要訓練時期介入的真正推理缺口。

以 GPT-5-mini 進行的跨模型 GCI 複製實驗使完全恢復率幾乎翻倍（50.4% 對比 25.5%），同時將真正推理缺口降至 11.7%，證明延伸思維鏈推理在提供正確概念後能顯著改善概念執行能力。

此處有一項重要注意事項，涉及知識缺口與注意力缺口的區分。當 GCI 恢復了一筆錯誤時，模型可能本來就「知道」該概念但未能檢索——提示僅是引導了注意力。一個確定性的測試應注入無關概念作為對照條件。我們將此留待未來研究，並指出目前的結果可能反映了兩種機制的混合。

5.3. 維度三：校準

5.3.1. 整體校準

表 7 呈現 CFA-Challenge 題目上所有模型—方法組合的校準指標。

Table 7: 各模型與方法之校準指標（CFA-Challenge，$N = 257$）

模型	方法	N	準確率	信心	ECE	Brier	AUC	OC 落差
GPT-4o-mini	自我一致性	90	.522	.829	.307	.334	.639	+.307
GPT-4o-mini	口述信心	95	.526	.841	.315	.340	.586	+.315
Qwen3-32B	口述信心	72	.611	.836	.247	.226	.787	+.225

ECE = 期望校準誤差；AUC = ROC 曲線下面積；OC 落差 = 平均信心 − 準確率。

所有組態均呈現顯著的過度自信，過度自信落差範圍從 +22.5% 至 +31.5%。模型表達平均約 84% 的信心，但實際僅達到 52–61% 的準確率。單樣本 t 檢定對逐觀測過度自信落差的結果為 $t = 9.70$（$p < 0.0001$）。

5.3.2. 過度自信錯誤分析

在所有 257 筆觀測中，77 筆為過度自信錯誤（30.0%），顯著超過 20% 的基線（$z = 3.99$，$p < 0.0001$）。在不正確的回答中，66.4% 伴隨信心 $\geq 80\%$，意味著大多數錯誤為高信心錯誤——對依賴表達信心的使用者而言，錯誤訊號幾乎不可見。

5.3.3. 主題層級校準偏差

倫理與標準展現最高的過度自信錯誤率（43.5%）與最低準確率（47.8%），而衍生性商品的過度自信錯誤率較低（22.2%）。這種達克效應（Dunning-Kruger effect）模式——模型在最不擅長的領域恰恰最為過度自信——對 AI 治理有直接啟示，顯示校準失敗與任務難度呈反比。

5.3.4. 風險信心

對 GPT-4o-mini 而言，CaR(5%) 無定義——沒有任何信心閾值能達到 5% 的錯誤率。即使在最高自我一致性信心（1.0）下，錯誤率為 32.4%（12/37 筆觀測值）；若放寬至 ≥ 0.9 信心區間，錯誤率達 41.7%（20/48）。對 Qwen3-32B 而言，信心 $\geq 95\%$ 時的錯誤率為 19.6%，仍遠超可接受的風險容忍度。現有 LLM 信心訊號從根本上不適用於金融風險管理。

5.3.5. 穩健性驗證：*CFA-Easy*（$N = 1{,}032$）

為評估可推廣性，我們在完整 CFA-Easy 資料集上複製了口述信心協議。

Table 8: 校準比較：CFA-Challenge 與 CFA-Easy（GPT-4o-mini）

指標	**CFA-Challenge** ($N = 95$)	**CFA-Easy** ($N = 1{,}032$)
準確率	52.6%	81.7%
平均信心	84.1%	86.0%
ECE	0.315	0.073
AUROC	0.586	0.671
OC 落差	+31.5 pp	+4.3 pp
OC 錯誤率	40.0%	15.1%

OC 落差 = 平均信心 − 準確率。OC 錯誤率 = 高信心（$\geq 80\%$）不正確回答的比例。

三項發現浮現。第一，在較簡單的問題上校準顯著改善（ECE：0.315 → 0.073）。第二，改善幾乎完全由準確率上升（+29.1 pp）所驅動，而非信心調整（+1.9 pp）——無論實際表現如何，模型均維持約 85% 的信心，這是「固定信心暫存器」的證據，而非真正的後設認知能力。第三，AUROC 仍然平庸（0.671），表明模型無法可靠地區分正確與不正確的回答。

5.4. 三維度整合

表 9 呈現整合後的三維度評估。

三個維度匯聚至同一結論：模型 82.4% 的標題準確率大幅高估了真正的能力。大約每五個正確答案中就有一個反映的是記憶而非推理；當模型失敗時，失敗在於概念辨識而非計算；且其信心訊號無法可靠地區分正確與不正確的回答。

Table 9: 三維度整合評估（GPT-4o-mini）

指標	數值	涵義
標準準確率	82.4%	標題（具誤導性）
穩健準確率	63.5%	記憶修正後
記憶溢價	18.9 pp	「幻影能力」
推理錯誤	90.1%	概念性而非計算性
GCI 恢復率	82.4%	知識缺口，可透過 RAG 修復
真正推理缺口	17.6%	需要微調
過度自信落差	+31.5 pp	困難題目
OC 錯誤率	30.0%	不可見的錯誤
CaR(5%)	無定義	信心不可信

6. 討論

6.1. 經濟意涵

記憶溢價具有具體的經濟意涵。標準準確率（82.4%）暗示五次金融計算中有四次正確；穩健準確率（63.5%）揭示僅有五次中三次正確。18.9 pp 的落差代表「幻影能力」——AI 看似勝任但在現實變體上會失敗的題目。

過度自信問題放大了這一風險。30% 的過度自信錯誤率意味著，依賴 AI 信心訊號的投資組合經理會以為模型六次中有五次正確，但實際上幾乎每隔一次就會出錯。一個過度自信的存續期間估計（D_{error}）在 1,000 萬美元部位上，若利率衝擊 100 個基點，將產生與誤差幅度成正比的意外損失。

更廣泛地說，AI 建議的資訊價值與訊號精確度 $\tau = 1/\sigma^2$ 成正比。我們觀測到的 ECE 值 0.25–0.32，使訊號精確度比使用者在依據「85% 信心」建議行事時隱含假設的低 40 倍。

6.2. 記憶悖論

跨模型證據引入了一項對治理具有重要意涵的發現：GPT-5-mini 的記憶落差（36.4 pp）幾乎是 GPT-4o-mini（18.6 pp）的兩倍，儘管前者是更強大的模型。這表明**標準準確率的提升可能在相當程度上歸因於增強的記憶能力，而非增強的推理能力**。隨著 AI 模型的改進，標準與穩健準確率之間的差距可能擴大而非縮小。追蹤標準準確率作為能力代理指標的金融監管機構，可能觀察到穩定的改善，而底層的穩健性實則停滯不前。

6.3. 對市場效率的影響

根據效率市場假說 [6]，市場價格反映理性代理人處理過的資訊。當 AI 諮詢系統成為邊際價格設定者時，本文記錄的系統性錯誤模式將威脅市場效率：倫理學中的概念誤用（87.1% 推理錯誤）可能產生系統性的合規違規，而衍生性商品定價失敗（37.5% 計算錯誤）可能在 AI 輔助的投資組合間產生相關的避險錯誤。與隨機雜訊不同，結構化錯誤會造成方向性偏差。

6.4. CFA 倫理與受託責任

本文的發現涉及三項 CFA 專業行為準則：

- **準則 I(C) — 虛偽陳述**：30% 的回答為高信心錯誤——而這些錯誤的平均表達信心高達 89%——模型系統性地向依賴信心訊號的使用者虛偽陳述其可靠性。
- **準則 V(A) — 勤勉盡責**：當 66.4% 的錯誤為高信心錯誤時，依賴 AI 信心作為驗證替代無法達到「合理基礎」的標準。
- **準則 III(C) — 適合性**：與主題相關的校準偏差意味著 AI 恰恰在最需要專業判斷的領域最不可靠。

6.5. 監管意涵

借鑒計量金融學，本文的記憶落差類似於「delta」（對輸入擾動的敏感度），NSI 功能類似於「vega」（對資訊雜訊的敏感度），而 CaR 直接對應風險值（VaR）。我們提出分級部署標準：

- **第一級 (顧問性)**：ECE < 0.15，記憶落差 $< 10\%$，OC 錯誤率 $< 15\%$
- **第二級（研究性)**：ECE < 0.25，記憶落差 $< 20\%$
- **第三級（內部使用)**：ECE < 0.35，並附強制性免責聲明

依據這些閾值，在 CFA-Challenge 上測試的所有模型均不符合第一級或第二級的部署資格。

6.6. 研究限制

若干限制應予說明。第一，實驗流程存在 LLM-as-judge 循環性：GPT-4o-mini 生成擾動、分類錯誤，並判斷自身輸出的開放式回答。雖然這在文獻中為常見做法，但可能引入系統性偏差；重要的是，同一流程應用於 GPT-5-mini 時產生了質性不同的結果（如記憶化悖論），顯示該方法能捕捉真實的模型差異，而非被評判者的偏差所主導。第二，擾動生成依賴 GPT-4o-mini；僅 68.0% 的題目產生了通過自動驗證的有效擾動，限制了覆蓋範圍。在 702 筆有效擾動中，未偵測到的正確答案錯誤將導致記憶化缺口被高估。第三，CFA-Challenge 上的校準分析（$N = 90$）統計檢定力有限，儘管 CFA-Easy 的穩健性驗證（$N = 1{,}032$）增強了可推廣性。第四，主題層級分析的每主題樣本量有限（$N = 10$–46），應視為探索性分析。第五，口述信心可能對提示敏感。第六，跨模型比較僅限於同一供應商（OpenAI）的兩個模型；擴展至其他模型家族將增強可推廣性。最後，GCI 實驗在缺乏以無關概念提示作為對照條件的情況下，無法完全區分知識缺口與注意力缺口。

7. 結論

本文證明標準基準準確率顯著高估了大型語言模型的金融推理能力，且此高估程度隨模型能力的提升而加劇。本文的三維度評估揭示：

1. **穩健性**：GPT-4o-mini 的記憶落差為 18.6 pp，GPT-5-mini 幾乎翻倍至 36.4 pp——一個更強大模型反而更依賴記憶的記憶悖論。
2. **錯誤結構**：90.1% 的錯誤屬於推理類型，以概念性錯誤為主(68.8%)，計算錯誤僅佔 1.4%。黃金情境注入恢復了 82.4% 的錯誤，確立了 RAG 補救的上限。
3. **校準**：普遍的過度自信，30% 的高信心錯誤，5% 閾值下 CaR 無定義，以及無論實際表現如何均維持約 85% 信心的「固定信心暫存器」。

三個維度匯聚一致：標題準確率因記憶而膨脹，失敗屬於概念性而非計算性，且模型無法可靠地發出其犯錯時的訊號。金融 AI 治理必須超越標題準確率，涵蓋穩健性、錯誤結構與校準品質。

問題不在於 AI 能否通過 CFA 考試，而在於其通過是否反映了可轉移至新問題的能力、其失敗是否能被有效補救、以及其信心是否值得信賴。本文的證據對上述三個問題均給出否定答案——而記憶化悖論顯示，穩健性問題可能隨著模型能力的提升而惡化，而非改善。

資料可用性

CFA-Easy 與 CFA-Challenge 資料集可透過 HuggingFace 上的 FinEval 基準測試取得 [9]。實驗程式碼與原始結果可向通訊作者合理請求後取得。

利益衝突聲明

作者聲明無已知的競爭性財務利益或個人關係可能影響本文所報告之研究。

CRediT 作者貢獻

Wei-Lun Cheng：概念化、研究方法、軟體、正式分析、資料整理、撰寫——初稿、視覺化。**Daniel Wei-Chung Miao**：指導、撰寫——審閱與修訂。**Guang-Di Chang**：指導、撰寫——審閱與修訂。

致謝

計算資源由國立臺灣科技大學（NTUST）提供。

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 ICLR 2024*.

[2] Band, N., Rudner, T. G. J., Filos, A., et al. (2025). Calibration of natural language understanding models with vague concepts. In *Proceedings of ICLR 2025*.

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

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

[5] Chhikara, P., Gaur, N., & Kumaraguru, P. (2025). Mind the confidence gap: Evaluating probabilistic forecasting of large language models. *Transactions on Machine Learning Research*.

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

[7] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of ICML 2017* (pp. 1321–1330).

[8] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

[9] 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*.

[10] Li, Q., Zhu, Z., Wang, Z., et al. (2024). GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of ACL 2024*.

[11] Liu, M., Chen, Y., & Wang, J. (2025). KalshiBench: Evaluating LLM probabilistic calibration using prediction markets. *arXiv preprint*.

[12] Lopez-Lira, A., Kirtac, K., & Tang, Y. (2025). The memorization problem: When can we trust financial LLM benchmarks? *arXiv preprint*.

[13] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., et al. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

[14] Patel, R., Singh, A., & Torres, M. (2025). Reasoning models ace the CFA exams: Implications for professional certification. *arXiv preprint*.

[15] Shi, W., Ajith, A., Xia, M., et al. (2023). Detecting pretraining data from large language models. In *Proceedings of ICLR 2024*.

[16] Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR 2023*.

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