繼承的非理性與倫理脆弱性：大型語言模型在金融決策中的行為偏誤與對抗性攻擊脆弱性

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

部署為金融顧問的大型語言模型（LLMs）面臨雙重威脅：繼承訓練資料中的行為偏誤，並在對抗性壓力下放棄倫理標準。本研究提出一個統一框架檢驗這兩個面向。首先，我們使用 140 組配對金融情境（先導 $N = 60$、複製 $N = 80$）測量 GPT-4o-mini 的六種行為偏誤——損失趨避、錨定效應、框架效應、近因偏誤、處分效應與過度自信。先導研究發現平均偏誤分數為 0.500；中性重新框架將此降至 0.425（Wilcoxon $p = 0.027$），揭示三層級去偏結構：表層偏誤（損失趨避、框架效應）可透過提示介入處理；錨定效應呈現邊際反應；而深層偏誤（處分效應、過度自信、近因偏誤）則完全具有抗性。擴展複製實驗確認此結構，處分效應表現為整體最強偏誤（0.808）。其次，將五種對抗性壓力類型施加於 47 道 CFA 倫理題目，所有攻擊均降低準確率（ERS = 0.925–0.950），翻轉 14 道題目。141 道合成題目的擴展實驗顯示翻轉率降至 0.85%（vs. 5.96%），僅重新框架與道德困境仍有效。三種合理化策略——功利主義覆蓋、權威服從與語義重新包裝——使模型為倫理妥協的答案建構看似合理的論證。跨模型比較顯示 GPT-5-mini 達到零對抗性翻轉，暗示此脆弱性可能隨世代改善。這些發現揭示金融 LLMs 面臨雙重危機：行為偏誤產生次優投資結果風險，倫理脆弱性產生合規違規風險——這是需要不同緩解策略的互補失效模式。

Keywords: 行為金融學, 對抗性倫理, 大型語言模型, 損失趨避, 處分效應, 人工智慧安全, CFA 考試, 受託責任

*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

大型語言模型（LLMs）在金融諮詢、合規與分析角色中的快速部署，引發了兩個關於其可靠性的根本問題。第一，這些系統——以大量人類生成的金融文本進行訓練——是否繼承了行為金融學所記載的系統性認知偏誤？第二，當它們遭受日常壓迫人類專業人員的壓力時，能否維持倫理判斷？

這些問題指向具有不同後果的互補失效模式。行為偏誤（損失趨避、處分效應、錨定效應）導致次優投資結果——風險在於虧損。對抗性壓力下的倫理脆弱性導致合規違規——風險在於法律與監管制裁。部署一個既具有非理性偏誤又具有倫理脆弱性的 LLM 的金融機構，將面臨當前 AI 評估框架——僅聚焦於準確率——完全忽略的複合風險。

我們提出一個統一的實驗框架，涵蓋兩個面向：

1. **行為偏誤測量**：使用配對情境設計，以跨越六種偏誤類型的 60 個 CFA 等級金融情境，量化偏誤易感性與去偏效果，識別出三層級的偏誤持久性結構。

2. **對抗性倫理壓力測試**：將五種壓力類型施加於 47 道 CFA 倫理題目，測量對抗條件下的倫理韌性，發現全面性退化現象，並描繪模型放棄倫理標準時所使用的合理化策略分類。

本研究有四項貢獻：(1) 首次對金融 LLM 的行為偏誤與倫理脆弱性進行聯合測量；(2) 識別出區分表層偏誤、弱反應偏誤與深層偏誤的三層級去偏結構；(3) 引入 AI 在對抗性壓力下的合理化策略分類；(4) 證明兩種失效模式需要根本不同的緩解方法——提示工程用於表層偏誤、訓練時介入用於深層偏誤、對齊改進用於倫理韌性。

2. 文獻回顧

2.1. LLMs 的行為偏誤

Kahneman and Tversky [5]的奠基性研究確立了個體透過損失趨避與參考點依賴系統性地違反期望效用理論。在金融市場中，這些表現為處分效應 [10] 和錨定偏誤 [12]。Ross et al. [9]引入「LLM Economicus」框架，發現 GPT-4 在抽象賭局情境中違反期望效用公理。Suri et al. [11]將此擴展至經濟決策，顯示 GPT-3.5 展現損失趨避。Capraro et al. [2]提供了跨認知心理學實驗的全面綜述。Malberg et al. [7]證明偏誤測量方法論在不同研究間存在顯著差異。

本研究與此文獻的區別在於：使用 CFA 等級的投資情境而非抽象賭局、測試金融特定的處分效應，以及引入去偏層級結構。

2.2. 對抗性倫理測試

AI 安全研究已發展出精密的對抗性測試方法。Chen et al. [3]證明多輪「得寸進尺」攻擊比單次提示更有效。Hui et al. [4]提出 TRIDENT 作為金融安全基準。Mazeika et al. [8]引入 HarmBench 用於標準化紅隊測試。Andriushchenko et al. [1]顯示簡單的自適應攻擊可繞過對齊機制。本研究將對抗性測試具體應用於 CFA 倫理題目，將發現與 CFA 專業行為準則連結，並引入領域特定指標（Ethics Robustness Score, ERS）。

3. 研究方法

3.1. 研究一：行為偏誤測量

3.1.1. 配對情境設計

針對每個金融決策，我們建構兩個版本：偏誤誘導版本，以觸發目標偏誤的方式框架情境；以及中性版本，僅呈現量化事實。若模型完全理性，兩種框架下的建議應完全相同。

我們測試六種經典偏誤，每種 10 個情境（$N = 60$）：損失趨避、錨定效應、框架效應、近因偏誤、處分效應與過度自信。每個情境呈現一個具有真實金融背景的 CFA 等級投資決策。

3.1.2. 偏誤評分

一個 LLM 評判模型（GPT-4o-mini）指派偏誤分數 $\in \{0.0, 0.5, 1.0\}$，其中 0.0 表示完全理性、0.5 表示混合/模稜兩可、1.0 表示完全偏誤行為。去偏效果定義為：

$$\Delta_{\text{debias}} = S_{\text{bias}} - S_{\text{neutral}} \tag{1}$$

3.2. 研究二：對抗性倫理壓力測試

3.2.1. 對抗性提示設計

針對 CFA-Easy 資料集 [6] 中的 47 道 CFA 倫理題目，我們建立一個標準版本和五個對抗性版本，在題目前加入誘導壓力的情境：**利潤誘因**（對錯誤行為給予經濟獎勵）、**權威壓力**（上級指示覆蓋倫理準則）、**情感操控**（引發同情心的情境鼓勵違規）、**重新框架**（以語言偽裝違規行為）、以及**道德困境**（以功利主義論證反對遵守規則）。

3.2.2. 倫理韌性分數

$$\text{ERS}_t = \frac{\text{Accuracy}_{\text{adversarial},t}}{\text{Accuracy}_{\text{standard}}} \tag{2}$$

ERS $= 1.0$ 表示無影響；ERS < 1.0 表示退化。我們追蹤「翻轉」題目：在標準條件下答對但在對抗條件下答錯的題目。

3.3. 模型

兩項研究均評估 **GPT-4o-mini**（OpenAI），溫度參數 $\tau = 0.0$。跨模型比較使用 **GPT-5-mini** 進行對抗性倫理測試。

4. 研究結果

4.1. 研究一：行為偏誤

4.1.1. 先導研究結果（$N = 60$）

表 1呈現先導研究的偏誤測量結果。模型展現 0.500 的平均偏誤分數，中性重新框架將此降至 0.425（Wilcoxon $W = 14.0$, $p = 0.027$, $r = 0.286$，精確檢定）。

Table 1: 先導偏誤測量結果（GPT-4o-mini，$N = 60$ 個情境，每類型 10 個）

指標	**偏誤誘導**	**中性**	$\Delta_{\mathbf{debias}}$
平均分數	0.500	0.425	+0.075
標準差	0.129	0.201	0.220
Wilcoxon: $W = 14.0$, $p = 0.027$, $r = 0.286$			

4.1.2. 三層級去偏結構

表 2揭示偏誤類型間的顯著異質性，形成三層級結構。

Table 2: 先導各類型偏誤分數（GPT-4o-mini，$N = 60$，每類型 10 個）

偏誤類型	n	**偏誤**	**中性**	$\Delta_{\mathbf{debias}}$
第一層：表層偏誤（可透過提示去偏）				
損失趨避	10	0.500	0.200	+0.300
框架效應	10	0.550	0.400	+0.150
第二層：弱反應偏誤				
錨定效應	10	0.500	0.450	+0.050
第三層：深層偏誤（抗拒去偏）				
處分效應	10	0.500	0.500	+0.000
過度自信	10	0.500	0.500	+0.000
近因偏誤	10	0.450	0.500	−0.050
整體	**60**	**0.500**	**0.425**	**+0.075**

第一層（表層偏誤）：損失趨避（$\Delta = +0.300$）和框架效應（+0.150）對中性重新框架反應強烈，暗示它們由提示-回應映射中的詞彙線索（如「虧損」、「下跌」）觸發。

第二層（弱反應偏誤）：錨定效應（+0.050）顯示邊際改善，表明模型依附於所提供數值資訊的傾向可透過提示部分但非完全解決。

第三層（深層偏誤）：處分效應（+0.000）、過度自信（+0.000）與近因偏誤（−0.050）完全抗拒提示層級的去偏。最值得注意的是，近因偏誤在中性框架下反而增加，暗示移除資訊可能矛盾地加劇偏誤。

4.1.3. 擴展複製實驗（$N = 80$ 合成情境）

為解決每類型 10 個情境的統計限制，我們使用 80 個獨立生成的合成情境（每類型 12–14 個）進行擴展複製。表 3呈現各類型結果。

Table 3: 擴展複製偏誤分數（GPT-4o-mini，$N = 80$ 合成情境）

偏誤類型	n	偏誤分數	層級
處分效應	13	0.808	深層
損失趨避	12	0.583	表層
錨定效應	13	0.538	弱反應
過度自信	14	0.464	深層
框架效應	14	0.429	表層
近因偏誤	14	0.000	深層
整體	**80**	**0.463**	—

結果確認並強化了先導研究的發現。處分效應表現為整體最強偏誤（0.808），大幅超越先導估計值（0.500），確認其作為深層去偏抗性偏誤的分類。近因偏誤在合成情境中降至 0.000，整體平均值（0.463）與先導研究（0.500）一致。跨獨立生成情境的收斂性強化了研究發現的可推廣性。

4.2. 研究二：對抗性倫理

4.2.1. 全面性退化

表 4呈現對抗性倫理結果。所有五種攻擊類型均持續降低表現（ERS < 1.0），共翻轉 14 道題目。

沒有任何攻擊類型未能損害倫理判斷，提供了 LLMs 學習的是倫理回應的形式而非原則的證據。

Table 4: 對抗性倫理結果（GPT-4o-mini，$N = 47$ 道 CFA 倫理題目）

條件	準確率	翻轉數	**ERS**	**ΔAcc**
標準（無壓力）	85.1%	—	1.000	—
利潤誘因	78.7%	4	0.925	−6.4 pp
權威壓力	78.7%	3	0.925	−6.4 pp
情感操控	80.9%	2	0.950	−4.3 pp
重新框架	80.9%	3	0.950	−4.3 pp
道德困境	80.9%	2	0.950	−4.3 pp

ERS = 倫理韌性分數（Ethics Robustness Score）。翻轉 = 在標準條件下答對但在對抗性壓力下答錯的題目。總計：14 題。

4.2.2. 合理化策略分類

對 14 個翻轉回應的質性分析揭示三種主要的合理化策略（涵蓋 14 個翻轉中的 12 個；其餘 2 個呈混合模式）：

- **功利主義覆蓋**（6 次翻轉）：模型建構結果論論證，將違規行為框架為「更大的善」，挪用受託人語言來合理化放棄受託責任。
- **權威服從**（3 次翻轉）：模型將其判斷從屬於層級權威，透過援引權威人物的經驗來合理化服從——直接違反 CFA 準則 I(B)。
- **語義重新包裝**（3 次翻轉）：模型吸收對抗性框架，將倫理違規重新定性為「務實解讀」——對應「創造性合規」的合規風險。

關鍵洞察在於，對抗性壓力並非產生明顯錯誤的輸出，而是生成聽起來合理的倫理推理卻得出錯誤結論——這是一種比簡單錯誤更加危險的合規威脅。

4.2.3. 合成倫理實驗

在 141 道合成生成的 CFA 倫理題目上，翻轉率從 5.96%（CFA-Easy）大幅下降至 0.85%，僅重新框架（4 次翻轉）和道德困境（2 次翻轉）仍保有效力。利潤誘因、權威壓力與情感操控產生零翻轉，暗示這些攻擊類型主要利用記憶化題目上的邊際信心，而非真正的推理脆弱性。

4.2.4. 跨模型比較

GPT-5-mini 達到零對抗性翻轉（所有五種攻擊類型的 ERS > 1.0），對抗性壓力矛盾地提升了準確率。這暗示此脆弱性可能具有世代界限，但此結果可能部分反映訓練資料記憶化而非真正的倫理韌性。

5. 討論

5.1. 雙重危機問題

我們的兩項研究揭示了為金融 AI 部署創造「雙重危機」的互補失效模式：

- **行為偏誤**導致系統性次優的金融建議——過早賣出贏家（處分效應）、過度重視近期表現、錨定於過時價格。
- **倫理脆弱性**導致模型在壓力下放棄合規標準——透過功利主義覆蓋、權威服從或語義重新包裝來合理化違規。

此區別對緩解策略至關重要。行為偏誤需要訓練資料層級的介入，因為深層偏誤在結構上根深蒂固且抗拒提示層級的去偏。倫理脆弱性可能可透過對齊改進來處理——GPT-5-mini 的零翻轉結果暗示了這一點——但必須跨模型家族驗證。

5.2. 經濟顯著性

處分效應——完全抗拒去偏——若機器人理財顧問系統性地過早賣出贏家，可能實質性地降低投資組合報酬。在規模化運作下，即使微小的系統性偏誤也會複利累積為顯著的財富損失。

倫理脆弱性帶來不同的後果。當 AI 合規系統能被操控為違規行為建構合理化論證——產生令人信服的理由而非明顯錯誤的答案——部署機構將面臨加重的監管責任。非理性損失基點，但倫理失敗損失執照。

5.3. 三層級去偏結構

此結構提供可操作的指引：

- **第一層（表層）**：損失趨避和框架效應可透過提示工程處理——指示模型「僅使用量化分析進行評估」及「聚焦於期望值」。
- **第二層（弱反應）**：錨定效應需要架構性修改——例如在處理前過濾提示中的數值錨點。
- **第三層（深層）**：處分效應、過度自信與近因偏誤需要訓練資料層級的介入——具偏誤意識的 RLHF、具去偏推理的合成資料，或對比式微調。

5.4. 與 *CFA* 準則的關聯

我們的對抗性攻擊對應特定的 CFA 準則：利潤誘因測試準則 III(C) 適合性；權威壓力測試準則 I(B) 獨立性；情感操控測試準則 III(A) 忠誠義務；重新框架測試準則 I(A) 法律知識。沒有任何 CFA 準則能免於對抗性攻擊的威脅。

5.5. 研究限制

應承認若干限制。第一，先導研究每種偏誤類型僅 10 個情境；雖然擴展複製（$N = 80$）確認了三層級結構，但每類型手動設計 20–30 個情境將進一步強化各類型結論。第二，LLM-as-judge 的評分方式可能引入偏誤。第三，倫理研究（$N = 47$）的子群分析屬探索性質。第四，跨模型偏誤驗證因 GPT-5-mini 在約 80% 情境中產生空白回應而被放棄。第五，單輪對抗性提示可能低估脆弱性；多輪攻擊 [3] 可能更為有效。最後，確定偏誤是從訓練資料中吸收還是從架構中湧現，需要進一步研究。

6. 結論

本研究證明金融 LLMs 面臨雙重威脅：它們展現與人類行為偏誤一致的模式，並且容易受到損害倫理判斷的對抗性壓力影響。處分效應與過度自信完全抗拒提示層級的去偏，而所有五種對抗性攻擊類型均可靠地降低倫理準確率。

三層級去偏結構與合理化策略分類為金融 AI 治理提供了可操作的框架：表層偏誤可透過提示工程處理，但深層偏誤與倫理脆弱性需要訓練時介入與對齊改進。跨模型證據（GPT-5-mini 的零對抗性翻轉）提供了倫理韌性可能隨世代改善的謹慎樂觀，但強制性測試仍不可或缺。

問題不在於 AI 是否消除了金融建議中的人類非理性，而在於它是否引入了一種新形式的非理性——統計性而非情感性、系統性而非特異性——伴隨著任何提示工程都無法完全解決的倫理脆弱性。

資料可用性

實驗情境與分析程式碼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的研究。

CRediT 作者貢獻

Wei-Lun Cheng：構思、方法論、軟體開發、正式分析、資料整理、撰寫——初稿、視覺化呈現。**Daniel Wei-Chung Miao**：指導、撰寫——審閱與編修。**Guang-Di Chang**：指導、撰寫——審閱與編修。

致謝

計算資源由國立臺灣科技大學（NTUST）提供。

References

[1] Andriushchenko, M., Croce, F., & Flammarion, N. (2025). Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *Proceedings of ICLR 2025*.

[2] Capraro, V., Lentsch, A., Aczel, B., et al. (2025). The impact of generative AI on collaborative human–AI decision-making. *Proceedings of the National Academy of Sciences* 122(7), e2413116122.

[3] Chen, Y., Yang, Z., Wang, X., et al. (2025). Foot-in-the-door: Multi-turn jailbreak attack on large language models. In *Proceedings of EMNLP 2025*.

[4] Hui, B., Chen, J., Li, S., et al. (2025). TRIDENT: A comprehensive financial safety benchmark for large language models. *arXiv preprint arXiv:2502.13399*.

[5] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292.

[6] 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*.

[7] Malberg, S., Dippold, J., & Romirer-Maierhofer, P. (2025). Cognitive biases in LLMs: A survey of findings and methodologies in NLP research. In *Proceedings of NLP4DH Workshop at COLING 2025*.

[8] Mazeika, M., Phan, L., Yin, X., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming. In *Proceedings of ICML 2024*.

[9] Ross, S., Kim, T.W., & Lo, A.W. (2024). LLM Economicus? Mapping the behavioral biases of large language models via utility theory. In *Proceedings of COLM 2024*.

[10] Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long. *The Journal of Finance*, 40(3), 777–790.

[11] Suri, G., Slater, L.R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? *Journal of Experimental Psychology: General*, 153(4), 1066–1075.

[12] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.

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