文獻回顧-

應用於股票交易的新型 DAPO 演算法

一、研究動機

在近年來，將強化學習（Reinforcement Learning, RL）應用到金融投資的自動化決策中越來越受到重視。透過 FinRL 框架，AI Agent 能夠自行決定股票或期貨等資產的買進、賣出時機，並在模擬的市場環境中進行學習。然而，現有的金融強化學習模型，像是 CPPO（Conditional Proximal Policy Optimization）和 FinRL-DeepSeek，雖然在投資策略上表現良好，但面臨了不少挑戰：它們需要長時間的訓練（通常超過 7 小時），並且需要佔用大量記憶體，對於硬體的要求較高，也讓系統的開發與維護變得更困難。此外，這些模型的決策過程也比較像「黑箱」，解釋性有限，無法讓使用者輕易理解 AI 為什麼會做出某個投資決定。

針對這些問題，透過進一步結合大型語言模型（LLMs），讓 AI 不只從市場數據學習，也能分析從新聞或財經報告中抽取的「情緒」和「風險」等訊號，提升決策品質。為了實現這些目標，本文提出了一種全新的強化學習演算法 DAPO-GRPO，透過多動作組內的獎勵標準化（Group Relative Policy Optimization）、避免過度壓縮探索行為（Decoupled Clipping）以及動態挑選有意義的學習樣本（Dynamic Sampling），在縮短訓練時間與減少記憶體開銷的同時，也能在 Nasdaq-100 等真實金融市場指數上，相較於過往強化學習模型取得更好的表現。

二、研究背景

在過去的交易決策領域，強化學習已經被廣泛應用，尤其是像 PPO（Proximal Policy Optimization）這樣的演算法，常被用來處理股票市場的自動化交易決策。然而，這類模型往往面臨大規模回撤的風險，導致表現不穩定。為了降低這種尾部風險，CPPO（Conditional PPO）在原本的 PPO 基礎上引入了條件風險價值（CVaR）限制。接著，FinRL-DeepSeek 更進一步，將來自金融新聞中透過大型語言模型（LLMs）生成的「情緒」和「風險」訊號，整合到 CPPO 的交易決策中。雖然這些模型在強化學習與交易結合上取得了顯著成果，但因為高維度的狀態空間、價值函數估計以及繁雜的超參數調整，通常需要大量的計算資源和訓練時間。

- PPO(Proximal Policy Optimization): 是一種在策略空間進行優化的演算法，

用於強化學習。它的核心思想是在保證新策略與舊策略不會差異太大的前提下，尋找一個性能更好的策略。這個特性通過一個被稱為「信賴區域(Trust region)」的概念來實現，這使得每一步更新都不會讓策略偏離太遠，從而避免了訓練過程中的不穩定現象。PPO 的目標函數 L(θ)結合了策略的性能以及新舊策略之間的差異表示：

$$L(\theta) = \hat{E}\left[\min\left(r(\theta)\widehat{A_t}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right]$$

其中，

- $\hat{E}$表示對樣本的期望
- $r(\theta)$是機率比率$\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$，表明了在新策略下選擇動作與舊策略下選擇動作機率的比例
- $\widehat{A_t}$是優勢函數(advantage function)，用來估計在狀態s_t下採取動作 a_t比平均更好多少
- Clip函數將 $r(\theta)$的值限制在$[1-\epsilon, 1+\epsilon]$的範圍內，這樣就避免了過大的策略更新

- CPPO (Conditional PPO): CPPO 是在 PPO 基礎上，引入了 CVaR（條件風險價值）來控制「風險」。它的核心目的是在 PPO 的策略學習目標中，加入對「尾部損失（特別大的損失）」的限制或考量。避免模型在追求高平均獎勵的同時，忽略可能出現的嚴重虧損。

為了改善這些問題，近來提出了 Group Relative Policy Optimization（GRPO），這是一種透過計算組內相對優勢（而不需要價值函數）的方法，能夠減少記憶體需求並且讓模型更新更加穩定。進一步地，Dynamic sAmpling Policy Optimization（DAPO）對 GRPO 進行了改進，尤其針對大型語言模型的場景。DAPO 包含兩個關鍵技術：一是非對稱裁剪（Decoupled Clipping），利用不對稱的範圍（由 ϵlow 和 ϵhigh 決定）取代對稱裁剪，讓模型在高獎勵情境下仍能靈活探索，同時控制風險；二是動態採樣（Dynamic Sampling），它能夠過濾掉沒有新訊號的樣本，讓模型專注於更有價值的學習資料，促進收斂更快且更穩定。這些技術的核心思想也被應用到本研究的 FinRL 案例中，特別是針對每天從 LLMs 取得的情緒和風險訊號，進一步優化交易決策的表現。

- GRPO(Group Relative Policy Optimization): GRPO 是 DeepSeek 使用的一種更新穎的算法，目的是簡化 PPO 中的流程，降低計算成本。它的主要改進在於去掉了評論家（Critic）這個額外的神經網絡，改為直接利用模型輸出的多個回應來比較它們的相對表現，對它們的「相對優勢」做學習。

三、研究方法

1. 資料集: 使用了 FNSPID 資料集，其中包含 1999-2023 年的 1570 萬條時間對齊的金融新聞記錄。
2. Stock Trading Environment:遵循標準的 FinRL 框架設定，定義了狀態、動作和獎勵。
   - ■ 狀態 (State)：代表當前交易環境的狀態，包括可用現金、股票價格、持股數量、技術指標以及 LLM 情緒和風險值。
   - ■ 動作 (Action)：交易代理做出的決策，包括買入、賣出或持有股票。
   - ■ 獎勵 (Reward)：採取動作後收到的反饋訊號，計算為總資產價值的變化。
3. GRPO with Exponentiated Sentiment-Risk Reward
   - ■ GRPO：公式定義了在狀態s_t下，對 候選動作$a_{t,i}$的「組內相對優勢值」A^G。

$$A^G(s_t, a_{t,i}) = \frac{r_{t,i} - \mu_t}{\sigma_t + \epsilon}$$

   - $r_{t,i}$：當執行動作$a_{t,i}$時得到的獎勵。
   - μ_t：在同一組的所有動作平均獎勵（組內平均值）。

$$\mu_t = \frac{1}{n}\sum_{j=1}^{n} r_{t,j}$$

   - σ_t：組內獎勵的標準差（衡量組內變異度）。

$$\sigma_t = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(r_{t,j} - \mu_t)^2}$$

   - ϵ：一個很小的常數，避免分母為 0。

   - ■ 結合風險和情緒訊號影響的獎勵設計:

$$r'_{t,i} = r_{t,i} \cdot \frac{(S_{t,i})^{\alpha}}{(R_{t,i})^{\beta} + 1 \times 10^{-8}}$$

$$S_{t,i} = \sum_{j=1}^{m} w_{t,i,j} f(S_{f,j}), R_{t,i} = \sum_{j=1}^{m} w_{t,i,j} f(R_{f,j})$$

   以 α、β 來代表情緒及風險因子的使用比重

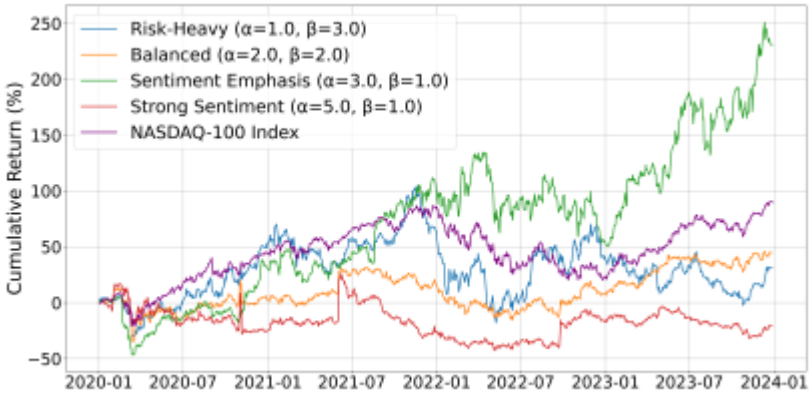
4. DAPO Policy Optimization

- 非對稱剪裁(Decoupled Clipping): 為了解決 GRPO 可能存在的探索受限和學習效率低下的問題，用兩個非對稱閾值 ϵlow 和 ϵhigh 取代了 GRPO 中的統一對稱裁剪參數 ϵ
- 動態採樣 (Dynamic Sampling)：過濾掉所有獎勵相同的狀態，以提高採樣效率

四、結果討論

1. 風險和情緒訊號的影響: 平衡配置下 (α=1, β=1) 實現了比單獨使用任一訊號更高的累積回報和改進的風險調整指標。結合兩個訊號使代理能夠偏好具有高情緒（表示市場樂觀）同時避免高風險情況的動作。

2. α 和 β 不同值的影響: 文獻中比較了五種不同的配置，分別是風險偏重（α=1, β=3）、平衡（α=2, β=2）、情緒強調（α=3, β=1）、強情緒（α=5, β=1），以及直接使用 NASDAQ-100 指數作為基準。實驗結果顯示，情緒強調的配置（α=3, β=1）表現尤為突出，產生了明顯高於平衡或風險導向設置的累積回報。值得注意的是，平衡（α=1, β=1）和情緒強調（α=3, β=1）的配置都達到了約 230.49% 的累積回報，這也說明了在合理的 α 與 β 範圍內，模型的表現具有一定的穩健性。然而，風險偏重以及強情緒的極端配置則表現不如理想，顯示在這個模型中，α 與 β 需要取得適當的平衡才能達到最佳表現。除了上述累積回報的比較，研究也提供了模型在 2020-2023 年期間，與 CPPO-DeepSeek 10% 在關鍵指標上的比較表格；並且包含了 2019-2023 年的回測結果，讓實驗結果更具說服力和完整度。



Metric	Our Model	CPPO-DeepSeek 10%
Cumulative Return	**230.49%**	~215%
Max Drawdown	-49.11%	**~-35%**
Rachev Ratio[1]	**1.12**	0.9818
Information Ratio[2]	**0.37**	0.0078
CVaR (5%)	-5.64%	**-4.37%**
Outperformance Frequency	50.0%	Not reported

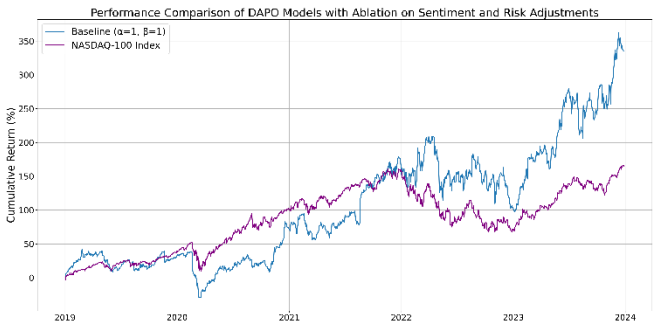
Metric	Our Model (2019–2023)
Cumulative Return	335.58%
Max Drawdown	-50.24%
Rachev Ratio	1.09
Information Ratio	0.30
CVaR (5%)	-5.50%
Outperformance Frequency	49.6%

3. 計算效率提升

Metric	Our Model	CPPO-DeepSeek 10%
RAM Usage (GB)	**15**	120
Training Time (100 Epochs)	**2.5 hours**	~7-8 hours

五、自行實作

1. 先安裝 Anaconda prompt，並執行 setup.bat 進行環境套件安裝
2. 使用 download_data.bat 下載預處理數據集(已加入 LLM 判斷訊號)
3. 執行 python train_dapo_llm_risk.py --adjustment_type both --alpha 1.0 --beta 1.0 進行模型訓練
4. 執行 python backtest_main_dapo.py 進行結果測試分析
5. 測試結果如下:



六、結論及心得

透過這次的程式模擬，我們驗證了 DAPO 模型在自動化投資組合決策上的強大表現。從累積回報率的走勢圖可以看出，無論是在市場波動劇烈的階段，還是長期的上漲趨勢中，DAPO 模型都能有效地超越傳統指數表現，展現出更高的累積報酬率和穩健性。特別是從 2022 年開始，DAPO 模型不僅跑贏了 NASDAQ-100 指數，還能在高點與回撤中保持穩定的增長，這顯示出模型在動態市場中的適應能力和抗風險能力。

此外，這次比較也讓我深刻體會到在強化學習模型中，情緒因子 α 和風險因子 β 的配置對最終表現的重要性。從論文結果來看，過於強調風險或情緒的極端設定反而導致表現下滑，而適度平衡的配置則能達到最好的報酬和穩健度。這提醒我在未來的模型調整中，需要更加謹慎考慮權重分配與超參數的選擇。

總體來說，這次模仿實作不僅讓我熟悉了如何運用 DAPO 模型結合金融市場訊號，還幫助我建立了對強化學習在金融交易應用的信心，未來也希望能進一步探索如何在更多市場情境中應用這些技術，持續優化模型表現。

七、引用文獻

Zha, R., & Liu, B. (2025). A New DAPO Algorithm for Stock Trading. 2025 IEEE 11th International Conference on Intelligent Data and Security (IDS). https://doi.org/https://doi.org/10.48550/arXiv.2505.06408