A New DAPO Algorithm for Stock Trading

Ruijian Zha†
Department of Computer Science, Columbia University
New York, NY, USA
rz2689@columbia.edu

Bojun Liu†
Department of Computer Science, Columbia University
New York, NY, USA
bl3092@columbia.edu

***Abstract*—Recent advances in reinforcement learning, like Dynamic sAmpling Policy Optimization (DAPO) algorithm, have demonstrated strong performance when combined with Large Language Models (LLMs). Motivated by this, we investigate whether similar benefits can be achieved in financial trading. We develop a novel trading agent that integrates an *improved* Group Relative Policy Optimization (GRPO) algorithm, enhanced with insights from DAPO, and incorporates LLM-based signals of risk and sentiment from financial news. We evaluate our method on the NASDAQ-100 index using the FNSPID dataset. Our improved DAPO algorithm achieves a cumulative return of 230.49% and an Information Ratio of 0.37, outperforming the CPPO-DeepSeek baseline. Additionally, our approach reduces training time from approximately 8 hours to 2.5 hours over 100 epochs, while significantly decreasing RAM usage. Our approach addresses limitations of existing RL-LLM frameworks, offering a scalable solution for building financial trading agents. Code is available at: https://github.com/Ruijian-Zha/FinRL-DAPO-SR/**

***Index Terms*—Algorithmic trading, Reinforcement learning, Risk management, Sentiment analysis, DAPO, GRPO**

I. Introduction

Financial Reinforcement learning (FinRL) has gained traction in algorithmic trading for automating portfolio decisions under uncertainty [1], [2]. However, challenges like large drawdowns and limited interpretability persist. Extensions of PPO, e.g., CPPO [3] and FinRL-DeepSeek [5], incorporate tail-risk constraints and textual signals, yet often require lengthy training ($>$ 7 hours) and significant memory overhead.

The FinRL Contest 2025 [9] aims to advance the application of reinforcement learning in finance by evaluating intelligent trading strategies. *Task 1: FinRL-DeepSeek* [9] explores the integration of FinRL with large language models (LLMs), using DeepSeek's open-source models [4] to generate sentiment and risk signals from financial news. This task encourages innovation in combining quantitative modeling with natural language understanding to improve trading performance.

We made the following three contributions:

- Adopting **Group Relative Policy Optimization (GRPO)**, a critic-free RL approach that reduces memory usage by sampling multiple actions from each state and normalizing the rewards within that group.
- Leveraging **Decoupled Clipping and Dynamic sAmpling Policy Optimization (DAPO)** [6], originally proposed for LLM preference tuning, to avoid issues like entropy collapse and to enable dynamic sampling of meaningful state-action pairs.
- Introducing an **adjustable reward formula** that allows more nuanced control. This enables us to emphasize sentiment or risk as desired.

Our tests on the NASDAQ-100 index show that our new DAPO algorithm surpasses the best performance of CPPO-DeepSeek [4] while significantly reduces computational requirements.

II. Related Work

A. Reinforcement Learning for Trading

PPO [1] is widely applied to stock trading scenarios [2], but it can be vulnerable to large drawdowns. CPPO [3] addresses tail risk using Conditional Value at Risk (CVaR) constraints, and FinRL-DeepSeek [5] extends CPPO by incorporating LLM-generated sentiment and risk signals from financial news data. While effective, these methods are typically resource-intensive due to high-dimensional state spaces, value functions, and extensive hyperparameter tuning.

B. GRPO and DAPO

Group Relative Policy Optimization (GRPO) eliminates the need for a value function by calculating group-level advantages, promising lower memory consumption and stable updates. **DAPO** [6] refines GRPO for large-scale language models, introducing:

- *Decoupled clipping*: Replaces symmetric clipping with an asymmetric range defined by parameters, ϵ_{low} and ϵ_{high}. This allows flexible policy updates, enhancing exploration in high-reward scenarios while constraining risk.
- *Dynamic sampling*: Samples with uniform risk-adjusted rewards are filtered out, ensuring that the model focuses on variations where learning signals are richer, thus promoting faster and more stable convergence.

We adapt these insights to a FinRL setting, focusing on daily aggregated sentiment and risk signals from LLMs.

III. Methodology

A. Stock Trading Environment

We follow the standard FinRL Contest setting for the definition of states, actions, and rewards as outlined in [9]. The state represents the current status of the trading environment, including the available cash, stock prices, number of shares held, and various technical indicators, including LLM

†These authors contributed equally to this work.

979-8-3315-9661-3/25/$31.00 ©2025 IEEE
DOI 10.1109/IDS66066.2025.00013

sentiment and risk values. The action refers to the decisions made by the trading agent, which includes buying, selling or holding stocks. The reward is the feedback signal received after taking an action, calculated as the change in total asset value.

B. GRPO with Exponentiated Sentiment-Risk Reward

Group Relative Policy Optimization (GRPO). For each state s_t, the policy π_θ samples a group of n potential actions $\{a_{t,1}, a_{t,2}, \ldots, a_{t,n}\}$. The group advantage for action $a_{t,i}$ is then defined as:

$$A^G(s_t, a_{t,i}) = \frac{r_{t,i} - \mu_t}{\sigma_t + \epsilon} \tag{1}$$

where $r_{t,i}$ denotes the reward for each candidate action, ϵ is a small constant to avoid division by zero, and

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

By relying on these group-based comparisons, GRPO obviates the need for a separate critic network, reducing computational overhead and potentially stabilizing updates.

Sentiment-Risk Adjusted Rewards. Our baseline reward $r_{t,i}$ for action $a_{t,i}$ is the portfolio return over one trading period. Let $S_{t,i}$ and $R_{t,i}$ be the aggregated sentiment and risk scores for each stock i at time t. We then incorporate sentiment and risk via:

$$r'_{t,i} = r_{t,i} \cdot \frac{\left(S_{t,i}\right)^{\alpha}}{\left(R_{t,i}\right)^{\beta} + 1e-8} \tag{3}$$

where

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

and weight $w_{t,i,j}$ reflects the proportion of stock j held by action $a_{t,i}$. Meanwhile, f map discrete LLM-generated sentiment and risk scores $(1 \ldots 5)$ to continuous factors ($0.99, 0.995, 1.0, 1.005, 1.01$), amplifying rewards under favorable sentiment and low risk. The exponents $\alpha, \beta \geq 0$ control the relative influence of sentiment vs. risk; for instance, $\alpha > \beta$ gives heavier weight to sentiment, while $\alpha = \beta = 0$ effectively disables sentiment-risk adjustments.

C. GRPO-Inspired DAPO

Policy Optimization with Decoupled Clipping. While GRPO offers a critic-free and memory-efficient alternative to traditional RL methods, it still suffers from limited exploration and learning inefficiencies in flat reward landscapes. To address these issues, we replace the uniform symmetric clipping parameter ϵ in GRPO with two asymmetric thresholds: ϵ_{low} and ϵ_{high}. The GRPO loss is updated as:

$$\mathcal{L}^{GRPO}(\theta) = \mathbb{E}\Big[\min\left(r_t(\theta)\, A_t^G,\ \text{clip}\big(r_t(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big) A_t^G\right)\Big], \tag{4}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. By enabling asymmetric clipping, the policy has greater flexibility to reinforce advantageous actions, while large negative updates remain constrained for stability.

Dynamic Sampling. To improve sample efficiency, we filter out states with all the same rewards. The policy parameters θ are updated via stochastic gradient descent:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}^{GRPO}(\theta).$$

By filtering out uninformative samples and employing decoupled clipping, our GRPO-inspired DAPO algorithm balances exploration and stability in large-scale financial trading.

IV. Preliminary Results and Discussion

A. Dataset

We use the FNSPID dataset [7], which includes 15.7 million time-aligned financial news records from 1999–2023. Since our primary objective is not to derive sentiment and risk signals, we adopt the pre-extracted LLM-based metrics from FinRL-DeepSeek [5], avoiding the overhead of retraining or fine-tuning large language models.

B. Impacts of Risk and Sentiment Signals

We first assess the contribution of risk and sentiment signals by comparing four configurations: (1) using only risk signals ($\alpha = 0, \beta = 1$), (2) only sentiment signals ($\alpha = 1, \beta = 0$), (3) balanced risk and sentiment ($\alpha = 1, \beta = 1$), with (4) the NASDAQ-100 index as a baseline. Fig 1 shows that the balanced configuration achieves a higher cumulative returns and improved risk-adjusted metrics compared to using either signal alone. Incorporating both signals allows the agent to favor actions with high sentiment (indicating market optimism) while avoiding high-risk scenarios, leading to more robust trading decisions.

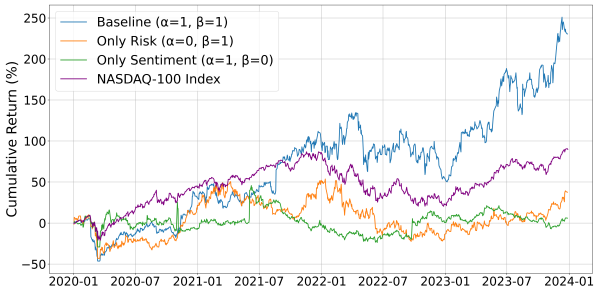

Fig. 1. Comparison of the Baseline, Risk-Only, and Sentiment-Only reward settings vs. the NASDAQ-100, showing 2020–2023 cumulative returns after 100 training epochs on six years of data.

C. Effects of Varying α and β

Next, we evaluate how different exponents for sentiment (α) and risk (β) influence the reward shaping. We compare five settings: (1) Risk-Heavy ($\alpha = 1, \beta = 3$), (2) Balanced ($\alpha = 2, \beta = 2$), (3) Sentiment Emphasis ($\alpha = 3, \beta = 1$), (4)

Strong Sentiment ($\alpha = 5, \beta = 1$), and (5) the NASDAQ-100 index. Fig 2 illustrates that a sentiment-emphasized configuration ($\alpha = 3, \beta = 1$) yields notably higher cumulative returns than either balanced or risk-focused settings, suggesting that strong sentiment signals can drive profitable trades.

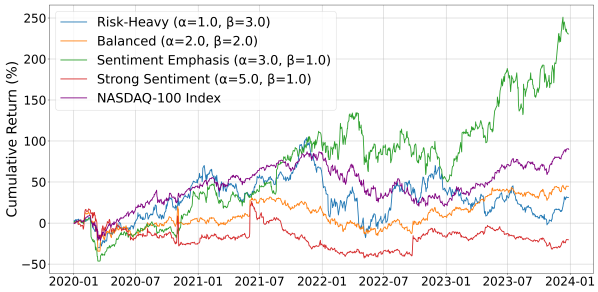

Fig. 2. Comparison of four sentiment–risk weightings vs. the NASDAQ-100, showing 2020–2023 cumulative returns after 100 training epochs on six years of data.

Both balanced ($\alpha = 1, \beta = 1$) and Sentiment Emphasis ($\alpha = 3, \beta = 1$) achieve $\sim$ 230.49% return, reflecting the reward function's robustness within a reasonable α, β range. Moderate α values leverage sentiment effectively, while $\beta = 1$ ensures balanced risk consideration. In contrast, Risk-Heavy ($\alpha = 1, \beta = 3$) over-penalizes risk, and Strong Sentiment ($\alpha = 5, \beta = 1$) over-emphasizes sentiment, leading to suboptimal results. This indicates α and β should remain balanced to maintain consistent performance across market conditions.

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

Table 1. Comparison of key metrics between our method and CPPO-DeepSeek 10%, which incorporates DeepSeek data.

To comply with the contest requirement, we also evaluated our model on the 2019–2023 period. Table 2 summarizes the performance metrics over this timeframe:

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%

Table 2. Backtesting results of our algorithm for the 2019–2023 period, meeting the contest's guideline for evaluation.

D. Improved Computation Efficiency

Our DAPO algorithm significantly reduces computational requirements compared to the CPPO-DeepSeek [5]. Experiments were conducted on a system with an Intel(R) Xeon(R) CPU @ 2.20GHz, 12 logical cores, 6 physical cores.

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

Table 3. Comparison of memory and training time (100 epochs, 20k steps each).

V. Conclusion and Future Work

We presented an improved DAPO algorithm for algorithmic trading that integrates exponent-weighted sentiment-risk scores. This approach allows flexible tuning of the reward to emphasize sentiment or risk. Results on the NASDAQ-100 index demonstrate potential for superior returns vs. prior SOTA, with faster training and much less memory usage.

Future directions include:

1) **Intraday Timescales**: Extending the approach to shorter intervals for rapid event-driven trading.
2) **Enhanced Prompts**: Using more detailed LLM prompts for sector-specific or event-specific signals.
3) **Adaptive** α, β: Automatically adjusting exponents based on market regimes.

References

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[2] X.-Y. Liu, H. Yang, Q. Chen, *et al.*, "FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance," *arXiv preprint arXiv:2011.09607*, 2022.

[3] C. Y. Ying, X. Zhou, H. Su, *et al.*, "Towards safe reinforcement learning via constraining conditional value-at-risk," in *IJCAI*, 2022, pp. 3673–3680.

[4] DeepSeek-AI *et al.*, "DeepSeek-V3 Technical Report," *arXiv preprint arXiv:2412.19437*, 2024.

[5] M. Benhenda, "FinRL-DeepSeek: LLM-infused risk-sensitive reinforcement learning for trading agents," *arXiv preprint arXiv:2502.07393*, 2025.

[6] Q. Yu, Z. Zhang, R. Zhu, *et al.*, "DAPO: An open-source LLM reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[7] Z. Dong, X. Fan, Z. Peng, *et al.*, "Fnspid: A comprehensive financial news dataset in time series," *arXiv preprint arXiv:2402.06698*, 2024.

[8] C. Y. Ying, X. Zhou, H. Su, D. Yan, N. Chen, and J. Zhu, "Towards safe reinforcement learning via constraining conditional value-at-risk," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pp. 3673–3680, 2022, doi: 10.24963/ijcai.2022/510.

[9] K. Wang, K. Xiao, and X.Y. Liu, "Parallel Market Environments for FinRL Contests," *arXiv preprint arXiv:2504.02281*, 2025.

[10] X.-Y. Liu, Z. Xia, H. Yang, J. Gao, D. Zha, M. Zhu, C. D. Wang, Z. Wang, and J. Guo,"Dynamic datasets and market environments for financial reinforcement learning, " *Machine Learning - Nature*, 2024.

[11] H. H. Yang, X.-Y. Liu, H. Tong, Y. Zhang, J. Wang, and L. Deng, "Deep reinforcement learning for automated stock trading: an ensemble strategy," *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 2245–2252, 2020.

[1]The Rachev Ratio measures tail-risk-adjusted performance by comparing expected tail gains to expected tail losses.

[2]Information Ratio evaluates the consistency of excess returns relative to a benchmark.