Integrating Reinforcement Learning with Large Language Model Signals for High-Frequency Cryptocurrency Trading

Arpan Mukherjee
Jadavpur University
arpanm.civil.ug@jadavpuruniversity.in

Sohom Ghosh
Jadavpur University
sohom1ghosh@gmail.com

Sudip Kumar Naskar
Jadavpur University
sudip.naskar@gmail.com

***Abstract*—Cryptocurrency markets exhibit high volatility and strong sensitivity to news. We propose a hybrid high-frequency trading framework that integrates deep reinforcement learning (RL) agents with sentiment and risk signals generated by large language models (LLMs). Our approach combines expert policy rollouts using a modified DAgger algorithm, selectively incorporating LLM inputs through a Rational Inattention-based gating mechanism. This enables the agent to attend to informative signals while ignoring noise, improving robustness. Experiments show that our method achieves higher ROI and reduced volatility compared to baseline ensembles. This paper presents Team *LIPI's* submission to FinAI 2025 Task-1 "FinRL-DeepSeek for Crypto Trading", showcasing a synergy between deep RL, imitation learning, and LLM-driven market understanding for real-time trading decisions.**

I. Introduction

Large language models (LLMs) such as Gemini [1] and DeepSeek [2] are capable of extracting structured information, such as sentiment or risk signals, from unstructured news text. These soft signals can enrich decision-making, but naively feeding them into RL models risks overfitting or introducing noise without clear integration mechanisms.

To bridge this gap, we propose a hybrid framework that combines deep RL ensembles (Proximal Policy Optimization, PPO [3]; Deep Q-Networks, DQN [4]; Advantage Actor–Critic, A2C [5]) with LLM-generated sentiment and risk scores, orchestrated through a modified Dataset Aggregation (DAgger) imitation learning process [6]. In the original DAgger formulation, the agent iteratively collects states encountered by its own policy, solicits expert actions for those states, and retrains on the aggregated dataset, thus correcting for distribution mismatch and minimizing compounding errors. We introduce an attention-gating mechanism into the framework inspired by Rational Inattention (RI) theory in economics. Prior research [7] states that decision-makers face costly information processing and selectively attend to signals that justify their cognitive cost, optimally trading off expected utility gains against information acquisition costs. Our RI-gate computes a deviation metric from neutral sentiment/risk baselines and only allows LLM-based signals to influence the policy when that deviation exceeds a threshold—mimicking capacity-limited attention to salient inputs.

Thus, our approach blends imitation learning from an ensemble of RL "experts" with selectively gated LLM inputs, yielding a policy that is both robust to numeric fluctuations and responsive to contextual sentiment without succumbing to noise. This integration of DAgger and RI theory allows us to systematically control when LLM signals are incorporated, improving stability and interpretability in high-frequency trading environments.

II. Problem Statement and Dataset

In this shared task, we participate as Team **LIPI** in the *FinRL-DeepSeek for Crypto Trading* task of the FinAI Contest 2025 (Task I), which focuses on high-frequency cryptocurrency trading. As described in the official task overview, participants are provided with Limit Order Book (LOB) data and financial news aligned at one-second resolution, and are tasked with building robust trading agents using ensemble methods and factor mining techniques [8, 9].

Our objective is to build an automated trading agent acting on Bitcoin in a simulated high-frequency (1-second) trading environment. The agent observes state vectors composed of LOB-derived price and volume features, technical indicators, and optional LLM-generated sentiment/risk signals. The action space is discrete (*Buy*, *Sell*, *Hold*), and the goal is to maximize risk-adjusted return over unseen test periods. The system is evaluated under the same GPU-parallel trading infrastructure provided by the contest, ensuring fairness and consistency [9].
Dataset, split, and metrics: We use the official 1-second LOB+news dataset from the contest and an additional open-source dataset from Kaggle [1]; data is split chronologically into train and test sets in an 80:20 ratio. Evaluation follows the contest rules using cumulative return, Sharpe ratio, and maximum drawdown (with optional win/loss or Rachev ratios), with rankings based on a weighted combination of these metrics [9].

This problem statement sets the stage for our framework integrating RL ensembles, LLM-based signals, and Rational Inattention gating within the DAgger imitation learning pipeline.

[1] https://www.kaggle.com/datasets/martinsn/high-frequency-crypto-limit-order-book-data

III. Methodology

Dataset and preprocessing: We use two offline datasets curated for the FinAI Contest 2025 shared task: (i) high-frequency (1-second) BTC/USDT limit order book (top-of-book bid/ask prices and volumes), and (ii) timestamp-aligned crypto news headlines from financial sources (both contest-provided and an open Kaggle source). Headlines are aligned to trading timestamps via a time-aware left join (merge-asof style), associating each market second with the most recent prior headline. Numerical features (except the unscaled midpoint price used for portfolio valuation) are standardized using a scaler fit on the training split. LOB gaps are forward-filled and any residual missing values set to zero; missing LLM sentiment/risk signals default to neutral values (3 and 3) following bounded-rationality motivation [7].

LLM scoring: We used instruction-tuned LLMs offline — *GEMMA-7B-IT* [10], *DeepSeek-R1* [11], and *LLaMA-3 Instruct* [12, 13], to extract sentiment and risk scores on a 1–5 scale (1 = worst, 5 = best). Scores are parsed from JSON-style outputs; if a score is missing or arrives beyond a fixed 60-minute horizon, we apply neutral defaults. Delays arise because LLM inference and news ingestion are asynchronous relative to the exchange feed (GPU/CPU queueing, batching, and I/O), therefore alignment uses timestamps with a strict "most-recent prior" rule and a small tolerance window.

Trading environment: We implement a custom OpenAI Gym–compatible environment (i.e., the standard *reset/step/render* interface for RL experimentation [14]) that simulates high-frequency crypto trading with realistic frictions (e.g., a fixed transaction-cost rate such as 0.1%). Observations are flattened feature vectors over a fixed *lookback window* of recent timesteps (normalized LOB features plus optional LLM signals). The discrete action space is {Hold, Buy, Sell}. The per-step reward is defined as the portfolio return relative to the previous timestep, as given below, with portfolio value defined as cash plus BTC holdings marked to the unscaled midpoint price.

$$\text{reward} = \frac{\text{portfolio}_t - \text{portfolio}_{t-1}}{\text{initial capital}}$$

We track net-worth trajectories and terminate episodes at the end of the data.

Expert models and LLM signals: We train three deep RL agents - DQN, PPO, and A2C, using Stable-Baselines3[2] (same training environment and split), and form an ensemble expert by majority vote [4, 3, 5]. In parallel, LLM-derived sentiment/risk scores are converted into a deterministic signal expert: if sentiment ≥ 4 and risk $\leq 2 \rightarrow$ Buy; if sentiment ≤ 2 or risk $\geq 4 \rightarrow$ Sell; otherwise $\rightarrow$ Hold. We combine these via a Rational Inattention gate and define the deviation as given below, where s and r are the LLM sentiment and risk scores, respectively.

$$D(s, r) = |s - 3| + |r - 3|,$$

[2]https://github.com/DLR-RM/stable-baselines3

If $D(s, r) > \theta$ and the ensemble confidence is below 0.66, the LLM action overrides the ensemble; otherwise, the ensemble decision stands. This yields the final expert action label for imitation learning.

Dataset Aggregation (DAgger) with RI gating: DAgger [6] iteratively aggregates expert-labeled data. In the i^{th} iteration, a policy π_i is trained on $D_i = D_{i-1} \cup \{(s_t, \pi^*(s_t))\}_{t=1}^{T}$, minimizing supervised loss on the aggregated set. We run multiple DAgger iterations, gradually shifting from expert-dominant actions ($\beta \approx 1.0$) to model predictions ($\beta \approx 0.1$), with the RI gate enforcing selective attention to salient LLM signals.

IV. Results and Discussion

We evaluate the performance of our proposed framework using multiple metrics and visual comparisons across models.

A. Experimental Setup

We train on high-frequency BTC/USDT LOB data at 1-second resolution, using 80% data for training and the rest for testing. DQN/PPO/A2C agents (Stable-Baselines3) are trained individually in a shared Gym-compatible environment [15, 14]. Their majority vote forms the ensemble expert for DAgger [6]; a Rational Inattention gate [7] overrides the ensemble with LLM-derived actions when sentiment/risk deviations from neutral are salient. LLM scoring runs fully offline via `transformers` with 4-bit quantization - Gemma-7B-IT, LLaMA-3-8B-Instruct, and DeepSeek [10, 12, 11], using a single JSON prompt to produce integer sentiment/risk scores (1–5). On parse failure, we default to neutral $(3, 3)$. Scores are causally aligned to the LOB with a 60-minute influence window; outside of which we apply neutral values to avoid look-ahead. DAgger runs 10 iterations with β linearly decaying $1.0 \rightarrow 0.1$, 1000 steps/iteration, with fixed seeds and a consistent preprocessing pipeline.

B. Cumulative Return (ROI)

Figure 1 shows one of the best DAgger-only runs over a given 1-sec BTC-LOB dataset achieving over 5000% ROI.

Fig. 1: DAgger with Deepseek Generated Scores

C. Imitation Learning With Data Generation

A synthetic dataset was generated using a given BTC-LOB offline dataset containing different LLM scores.

D. DAgger With RI-Gate

Figure 3 shows the output of the DAgger run of the same setup mentioned in the previous subsection, but along with RI-Gate.

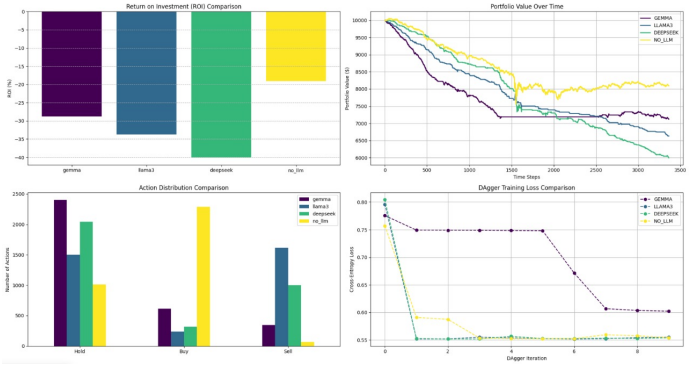

Fig. 2: DAgger without RI-Gate with Data Scraping

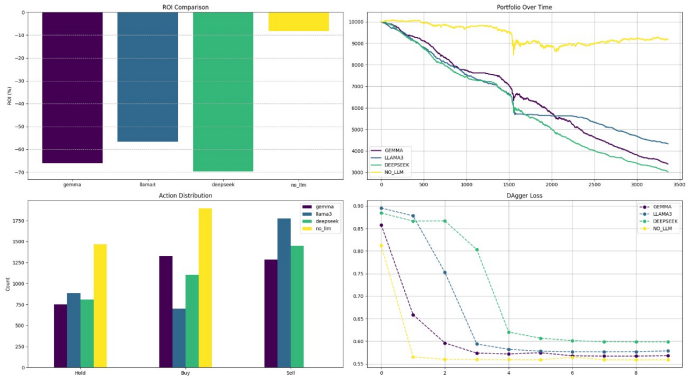

Fig. 3: DAgger with RI-Gate

E. Performance Insights

The use of LLM-derived signals in combination with a Rational Inattention (RI) gating strategy allows the agent to selectively leverage high-value information. The DAgger framework enables safe and efficient imitation learning from both learned experts and LLM heuristics. Again, over multiple runs, we consistently observe that when no LLM scores are used, the cumulative ROI is greater than when DAgger is being trained using LLM scores.

V. Conclusion and Future Work

Our results echo the observations of Benhenda, 2025, finrl-deepseek [16], specifically the FinRLDeepSeek study, which also highlighted that during periods of market growth, PPO without the added layer of an LLM tends to be more effective than agents that incorporate them. Delving into the mechanisms that drive this characteristic is a compelling area for further investigation.

Future directions include adaptive tuning of hyperparameters, RI threshold θ, using API calling for LLMs instead of using transformers in an offline setting, performing early stopping, attention-based neural networks for gating, and multi-asset portfolio extension at improved ROI with real-time LLM signals.

References

[1] Gemini Team, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," Google, Tech. Rep., 2025.

[2] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[3] J. e. a. Schulman, "Proximal policy optimization algorithms," in *NeurIPS*, 2017.

[4] V. Mnih, K. Kavukcuoglu, and D. e. a. Silver, "Human-level control through deep reinforcement learning," in *Nature*, vol. 518, 2015, pp. 529–533.

[5] V. e. a. Mnih, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016.

[6] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *AISTATS*, 2011, pp. 627–635.

[7] C. A. Sims, "Implications of rational inattention," *Journal of Monetary Economics*, vol. 50, no. 3, pp. 665–690, 2003.

[8] Y. Yu, H. Li, Y. Cao, and K. e. a. Wang, "Finnlp-fnp-llmfinlegal @ coling 2025 shared task: Agent-based single cryptocurrency trading challenge," in *FinNLP-FNP-LLMFinLegal @ COLING 2025*. Association for Computational Linguistics, 2025, pp. 401–406.

[9] K. Wang, K. Xiao, and X.-Y. e. a. Liu, "Parallel market environments for finrl contests 2023–2025," *arXiv preprint arXiv:2504.02281*, 2025.

[10] K. Mo, W. Liu, X. Xu, C. Yu, Y. Zou, and F. Xia, "Fine-tuning gemma-7b for enhanced sentiment analysis of financial news headlines," *arXiv preprint arXiv:2406.13626*, 2024.

[11] D. Huang and Z. Wang, "Explainable sentiment analysis with deepseek-r1: Performance, efficiency, and few-shot learning," *arXiv preprint arXiv:2503.11655*, 2025.

[12] L. Meng, W. Weng, and P. Mane, "Financial sentiment analysis meets llama 3: A comprehensive analysis," in *WASSA Workshop, ACM Digital Library*, 2025.

[13] T. Konstantinidis, G. Iacovides, M. Xu, A. G. Constantinides, and D. Mandic, "Finllama: Financial sentiment classification for algorithmic trading applications," *arXiv preprint arXiv:2403.12285*, 2024.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[15] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," 2021, gitHub repository.

[16] M. Benhenda, "Finrl-deepseek: Llm-infused risk-sensitive reinforcement learning for trading agents," *arXiv preprint arXiv:2502.07393*, 2025.