The Losing Winner: An LLM Agent that Predicts the Market but Loses Money

Youwon Jang[1*] Joochan Kim[2*] Byoung-Tak Zhang[1†]
[1]Seoul National University [2]Korea Institute of Science and Technology
{sharifa, btzhang}@snu.ac.kr, joochan.k@kist.re.kr

Introduction

- Fine-tuned small-scale LLM (Qwen2.5-3B-Instruct) for BTC trading via **market state prediction**.
- Used daily price, volume, and technical indicators; trained with RLVR to classify next-day market: bullish, consolidation, bearish.
- Achieved **higher classification accuracy** than zero-shot baseline.
- **Paradox:** better predictions led to **lower cumulative trading returns**.
- Cause: **objective mismatch**—model optimized for classification, not profit, ignoring risk and price magnitude.
- Highlights challenges in reward design and the need to align proxy tasks with actual profit goals.

Methodology

- **Supervised Fine-Tuning (SFT)**
  - Initial task-specific adaptation on historical market data.
  - Model learns to classify next-day market state: *Bullish*, *Consolidation*, *Bearish*.
  - Trained with **cross-entropy loss** to maximize probability of correct labels.
  - Provides **baseline policy** for reinforcement learning.

$$\pi_{\text{SFT}} = \arg\max_{\pi} \mathbb{E}_{(s_t, a_t^*) \sim \mathcal{D}} \left[\log \pi(a_t^* \mid s_t) \right]$$

- **Reinforcement Fine-Tuning (RFT)**
  - Policy refined based on **outcome-based rewards** of model predictions. Trained with Low-Rank Adaptation **(LoRA)**.
  - Uses Reinforcement Learning with Verifiable Reward **(RLVR)** and Guided Reward Policy Optimization **(GRPO)**.
  - Model predicts an action (market state), evaluated against ground-truth to generate **verifiable reward** R_t.
  - Reward considers **prediction accuracy** and **format accuracy**; binary scoring (+1 / 0).
  - Optimization includes **KL regularization** to prevent deviation from initial SFT policy.

$$\pi_{\text{RFT}} = \arg\max_{\pi} \mathbb{E}_{s_t \sim \mathcal{D}} \left[\mathbb{E}_{a_t \sim \pi(\cdot|s_t)}[R_t(a_t)] - \beta\, \mathbb{D}_{\text{KL}}\big(\pi(\cdot|s_t) \,||\, \pi_{\text{SFT}}(\cdot|s_t)\big) \right] \qquad R_t = w_{\text{format}} \cdot R_{\text{format}} + (1 - w_{\text{format}}) \cdot R_{\text{acc}}$$

Results

Strategy	Validation Set			Bullish Set			Bearish Set		
	Acc	Cum. Return	Sharpe	Acc	Cum. Return	Sharpe	Acc	Cum. Return	Sharpe
Buy & Hold	-	+15.4%	2.03	-	+78.3%	8.91	-	-17.5%	-5.02
Qwen2.5-3B-Instruct	25.4%	+10.1%	2.30	8.3%	+6.2%	3.49	38.5%	-0.4%	-2.65
+ SFT	44.1%	+8.3%	2.03	27.1%	+50.9%	8.05	23.1%	-10.2%	-2.76
+ RFT	64.4%	+0.0%	0.0	66.7%	+0.0%	0.0	75.0%	+0.0%	0.0
+ SFT & RFT	45.8%	+15.4%	2.03	41.7%	+78.3%	8.91	44.2%	-14.8%	-3.09

Conclusion

- Fine-tuned LLM agent shows higher predictive accuracy but lower trading returns.
- Reinforcement Learning leads to '**reward hacking**,' ignoring return magnitude and risk management.
- '**Losing Winner**' highlights risks of treating complex financial tasks as simple classification problems.
- Success of generative AI in finance relies on designing reward functions aligned with risk-adjusted profit maximization, not just proxy metrics.