

7459518-finrl-deepseek-revisited-a-reproduction-and-multi-seed-robustness-analysis-of-ll

AVTORJI: Arnold Leumale

POVZETEK: Raziskava ponovno izvede in razširi FinRL-DeepSeek [Benhenda, 2025], razvojno okolje za trgovino, ki kombinira Proximal Policy Optimization (PPO), Conditional Value-at-Risk (CVaR) omejitve in signalje z velikimi jezikovnimi modeli (LLM) iz finančnih novic. Izvedba četrth agentov iz začetka v PyTorchu in njihova povratna testiranje na oborožju Nasdaq-100 (2019–2023) proti enakomerno vloženemu vključenemu vključku (buy-and-hold) pokazuje, da LLM vključevanje poveča vrednost in Sharpe razmerje za omejeni in CVaR-omejeni agenciji. Ta trend pa ne je stabilen pri dveh neodvisnih semih, posebej za PPO-DeepSeek, ki se obrne v drugačno rangiranje vrednosti Sharpe. Vrstni red celotne vrednosti za CPPO-DeepSeek pa je najmanj splošen med dvema semima. Raziskava predstavlja te napačke kot prenovljeno dokazljivost in podpira, da je večsaga scale multi-seeda ocene potrebna v finančnem učenju s podobnimi raziskavami.

METODA: Raziskava ponovno izvede FinRL-DeepSeek, implementirane četrth agenti (PPO, CPPO, PPO-DeepSeek, CPPO-DeepSeek) v PyTorchu in povratno testira na oborožju Nasdaq-100 (2019–2023) proti enakomerno vloženemu vključku. Izvedba je preizkusljiva in dokumentirana, in vključuje vključek trgovinskega okolja, implementacije PPO/CPPO in evaluacijski potok. Raziskava uporablja dve neodvisni semi, da preveri stabilnost rezultatov.

KLJUČNE UGOTOVITVE:

- LLM vključevanje poveča vrednost in Sharpe razmerje za omejeni in CVaR-omejeni agenciji.
- Trenutni trend ni stabilen pri dveh neodvisnih semih, posebej za PPO-DeepSeek.
- CPPO-DeepSeek prikazuje najmanj splošen rangiranje celotne vrednosti med dvema semima.

POMEN ZA TRGOVANJE: Raziskava podpira, da je multi-seeda ocena neprimerno v finančnem učenju, kjer se pogosto izvedejo povratna testiranja na enem treniranem izpitu. To podpira potrebo po večji skalirani multi-seeda oceni, ki bi omogočila boljše razumevanje stabilnosti in prednosti različnih agencij.

FinRL-DeepSeek Revisited: A Reproduction and Multi-Seed Robustness Analysis of LLM-Infused Risk-Sensitive Trading Agents

Arnold Leumale

MSc AI & Data Science

aivancity School for Technology, Business & Society

leumalearnold411@gmail.com

Abstract

We reproduce and extend FinRL-DeepSeek [Benhenda, 2025], a trading framework combining Proximal Policy Optimization (PPO) [Schulman et al., 2017], Conditional Value-at-Risk (CVaR) constraints [Rockafellar and Uryasev, 2000], and large language model (LLM) signals extracted from financial news. Using the Nasdaq-100 universe (2013–2023) and the agent-ready datasets released by the original authors, we implement four agents from scratch in PyTorch — PPO, CPPO, PPO-DeepSeek, and CPPO-DeepSeek — and backtest them on the 2019–2023 period against an equal-weight buy-and-hold baseline. On a single seed, our results replicate the qualitative pattern reported by Benhenda (2025): LLM infusion improves both return and Sharpe ratio for both the unconstrained and CVaR-constrained agents. However, extending the evaluation to two independent seeds — a small but motivating step in light of well-documented seed sensitivity in deep RL [Henderson et al., 2018] — suggests that this single-seed pattern may not be stable, particularly for PPO-DeepSeek, whose Sharpe ratio ranking relative to plain PPO reverses between the two seeds we test. In contrast, CPPO-DeepSeek shows the smallest spread in total return across our two seeds of any agent. We present these observations as preliminary evidence, not a settled finding, and argue that they motivate wider-scale multi-seed evaluation in financial reinforcement learning research, an area where single-seed backtests remain the norm. We release our full codebase, including the trading environment, PPO/CPPO implementations, and evaluation pipeline, to support reproducibility.

1 Introduction

Reinforcement learning (RL) has emerged as a promising paradigm for automated trading, allowing an agent to learn a trading policy directly from interaction with market data rather than relying on hand-crafted rules. However, two limitations restrict the practical use of standard RL objectives in finance. First, most RL trading agents rely solely on price and volume data, ignoring the rich contextual information available in financial news, earnings reports, and analyst commentary. Second, standard RL objectives maximize expected return, which is indifferent to rare but catastrophic losses in the tail of the return distribution.

FinRL-DeepSeek [Benhenda, 2025] addresses both limitations by combining (i) a CVaR-constrained variant of PPO (CPPO), inspired by risk-sensitive RL formulations [Rockafellar and Uryasev, 2000], with (ii) trading and risk signals extracted from financial news using a large language model (DeepSeek V3). The original paper reports that this combination, evaluated on the Nasdaq-100 index over 2019–2023, improves risk-adjusted performance relative to a plain PPO baseline.

Separately, a well-known line of work in the broader deep RL literature has shown that reported performance can be highly sensitive to random seed, and that single-seed or few-seed evaluations can produce misleading rankings between algorithms [Henderson et al., 2018]. This concern is rarely addressed in financial RL papers, where backtests are frequently reported on a single training run. Our work sits at the intersection of these two threads: we reproduce the FinRL-DeepSeek pipeline end to end — trading environment, PPO, CPPO, and LLM signal infusion — as a from-scratch PyTorch implementation, and use it to ask whether the original paper’s single-seed conclusions are robust to seed choice. Our contributions are:

- A clean, documented, from-scratch reproduction of the FinRL-DeepSeek agents (PPO, CPPO, PPO-DeepSeek, CPPO-DeepSeek) and trading environment, released publicly.
- A backtest on Nasdaq-100 (2019–2023) confirming the qualitative pattern of the original paper on a single seed.
- A small-scale ($n=2$ seeds) robustness check, explicitly framed as preliminary given the sample size, suggesting that this pattern may not generalize across seeds for the unconstrained LLM variant, while the CVaR-constrained LLM variant shows a notably smaller spread in total return between the two seeds tested.

2 Related Work

Deep RL for trading. Proximal Policy Optimization [Schulman et al., 2017] is a widely used policy-gradient algorithm valued for its training stability, achieved via a clipped surrogate objective that limits the size of policy updates. The FinRL framework [Liu et al., 2021] popularized standardized RL trading environments built on top of PPO and related algorithms.

Risk-sensitive RL. Standard RL objectives maximize expected return and are indifferent to tail risk. CVaR-constrained formulations address this by explicitly bounding the expected loss in the worst-case α -quantile of outcomes, following the Rockafellar-Uryasev formulation of CVaR as a convex optimization problem [Rockafellar and Uryasev, 2000]. FinRL-DeepSeek instantiates this idea for trading via a Lagrangian-relaxed PPO variant (CPPO).

LLMs for financial signal extraction. Recent work explores using LLMs to convert unstructured financial text (news, filings, social media) into structured trading signals. FinRL-DeepSeek [Benhenda, 2025] uses DeepSeek V3 to generate per-stock, per-day sentiment and risk scores from the FNSPID news dataset [Dong et al., 2024], which are then used to scale the RL agent’s actions and risk penalties. Our work reproduces this pipeline and evaluates its robustness under multiple random seeds.

Reproducibility and seed sensitivity in deep RL. Henderson et al. [Henderson et al., 2018] showed that policy-gradient methods, including PPO’s predecessors, can produce substantially different performance rankings depending on random seed, codebase, and even hyperparameter implementation details, and argued for reporting results across multiple seeds with variance estimates rather than single runs. This concern has motivated more careful evaluation protocols in general deep RL benchmarks, but is, to our knowledge, rarely applied in financial RL papers, where a single backtest run is the typical unit of evidence. Our multi-seed comparison should be read as an application of this well-established concern to the FinRL-DeepSeek setting, rather than as a novel methodological claim in itself.

3 Methodology

3.1 Data

We use daily OHLCV data and eight standard technical indicators (MACD, RSI-30, CCI-30, DX-30, Bollinger bands, and 30/60-day SMA) for 84 Nasdaq-100 constituents, spanning 2013–2023. We use the agent-ready datasets released by Benhenda [2025] (Hugging Face: `benstaf/nasdaq_2013_2023`), which additionally include DeepSeek V3 sentiment and risk scores derived from the FNSPID financial news corpus [Dong et al., 2024]. We follow the original train/test split: 2013–2018 for training, 2019–2023 for out-of-sample backtesting. We note that, as in the original paper, the constituent list reflects the *current* Nasdaq-100 composition, which introduces survivorship bias (discussed in Section 6).

3.2 Trading Environment

We implement a discrete-time trading environment following the standard FinRL formulation. The state s_t is an 841-dimensional vector consisting of: available cash (1), closing prices for all 84 tickers (84), current share holdings per ticker (84), and the eight technical indicators per ticker (672). The action $a_t \in [-1, 1]^{84}$ specifies, for each ticker, a normalized trade intensity that is scaled by $h_{max} = 100$ shares. Sell orders are executed before buy orders at each step to free up cash. A transaction cost of 0.1% is applied to both buys and sells. The reward is the change in total portfolio value (cash plus holdings) between consecutive days, scaled by a factor of 10^{-4} for numerical stability. Each episode starts with \$1,000,000 in cash and runs for the full length of the training or test period.

3.3 PPO Agent

Our actor and critic are both two-layer MLPs (64 units, tanh activations). The actor outputs the mean μ of a diagonal Gaussian policy over the 84-dimensional action space; the log standard deviation is a learned, state-independent parameter. We use Generalized Advantage Estimation (GAE, $\lambda = 0.95$, $\gamma = 0.99$) and the standard PPO clipped surrogate objective with clip ratio $\epsilon = 0.2$, combined with a value-function loss (MSE) and an entropy bonus (coef = 0.01). Each training iteration consists of one full episode rollout followed by 10 epochs of minibatch (size 64) gradient updates. Agents are trained for 300 iterations using Adam (lr = 3×10^{-4}) on a single CPU (Apple Intel), each run taking approximately 10 minutes. Reward scaling and the CVaR threshold (Section 3.4) are fixed a priori rather than tuned; we did not perform a hyperparameter search, and we flag this as a limitation (Section 6).

3.4 CPPO: CVaR-Constrained PPO

Following a Lagrangian relaxation of the CVaR-constrained RL objective, we maintain two dual variables updated at every iteration: a VaR threshold η and a Lagrange multiplier $\lambda \geq 0$. Given the vector of relative daily portfolio losses L_t observed during the most recent rollout, we compute the CVaR estimate using the Rockafellar-Uryasev formula:

$$\text{CVaR}_\alpha = \eta + \frac{1}{\alpha} \mathbb{E}[(L_t - \eta)^+]$$

with $\alpha = 0.05$. At each iteration, η is updated by gradient descent to track the empirical $(1 - \alpha)$ -quantile of losses, and λ is updated by a projected sub-gradient step $\lambda \leftarrow \max(0, \lambda + \text{lr}_\lambda (\text{CVaR}_\alpha - d))$ where $d = 0.02$ is the tolerated tail-loss threshold. The advantage of each transition in the tail ($L_t > \eta$) is then penalized by $\frac{\lambda}{\alpha} (L_t - \eta)^+$ before the standard PPO update, discouraging the policy from trajectories that produce large tail losses. Figure 1 illustrates the resulting evolution of η and λ during training.

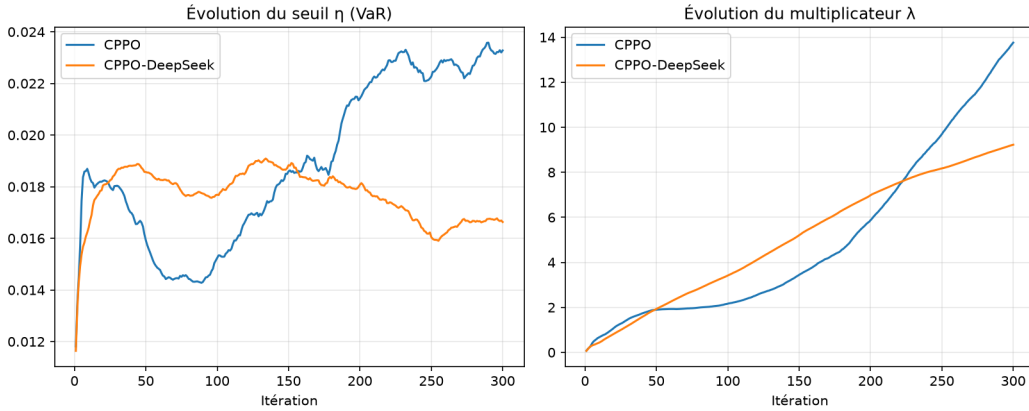


Figure 1: Evolution of the VaR threshold η and Lagrange multiplier λ during training, for CPPO and CPPO-DeepSeek (seed 42).

3.5 LLM Signal Infusion

DeepSeek V3 sentiment and risk scores (integers 1–5, with missing or invalid values treated as neutral) are converted to scaling factors in $[0.9, 1.1]$ via the affine map $f(\text{score}) = 0.9 + 0.05(\text{score} - 1)$, so that a neutral score of 3 yields a scaling factor of exactly 1.0 (no effect). In **PPO-DeepSeek**, the policy’s raw sampled action is multiplied element-wise by the per-ticker sentiment factor before being sent to the environment. The log-probability used by the PPO objective is computed on the pre-scaling action rather than the executed one: the policy being optimized is still the distribution over a_t , and the LLM scaling is applied only to the order sent to the environment, so this choice preserves an unbiased policy-gradient estimate while still letting the market-facing action reflect the LLM signal. In **CPPO-DeepSeek**, the same action scaling is applied, and additionally the daily loss series used in the CVaR update is scaled by an aggregate daily risk factor R_t , computed here as the mean risk factor across all 84 tickers. This is a simplification of the portfolio-weighted risk aggregation described in Benhenda [2025], which we adopt for implementation simplicity and document as a limitation (Section 6).

4 Experiments

We train all four agents (PPO, CPPO, PPO-DeepSeek, CPPO-DeepSeek) on the 2013–2018 period and evaluate them out-of-sample on 2019–2023 using a deterministic policy (the mean action μ , without sampling noise). We compare against an equal-weight buy-and-hold baseline that allocates \$1,000,000 evenly across all 84 tickers on the first trading day of 2019 and never trades again. We report four standard metrics: total return, annualized Sharpe ratio, maximum drawdown, and 5% CVaR of daily returns. To probe robustness, we repeat training and evaluation for two independent random seeds (42 and 123); we treat this as a first, resource-constrained step toward the wider multi-seed protocols called for by Henderson et al. [2018], not as a substitute for them.

5 Results

5.1 Single-Seed Results

Table 1 reports backtest performance for seed 42. LLM infusion improves both total return and Sharpe ratio relative to the corresponding non-LLM agent, for both the unconstrained (PPO vs. PPO-DeepSeek) and CVaR-constrained (CPPO vs. CPPO-DeepSeek) families. Notably, CPPO-DeepSeek achieves the best maximum drawdown (−34.5%) among our four agents, close to the buy-and-hold baseline (−33.9%), correcting the drawdown degradation observed for CPPO alone relative to plain PPO. Figure 2 shows the corresponding portfolio value trajectories.

We note that CPPO alone underperforms plain PPO on this seed across total return, Sharpe, and drawdown, despite achieving a better CVaR. A plausible explanation is a mismatch between the risk measure being penalized and the risk measure being reported: our CVaR penalty acts on the distribution of *isolated daily losses* observed during 2013–2018 training rollouts, whereas maximum drawdown is a property of *cumulative, path-dependent* loss sequences that can arise even when individual daily losses are unremarkable. An agent can therefore learn to avoid single bad days without learning to avoid prolonged adverse sequences, particularly when, as here, the training period (2013–2018) contains comparatively few sustained drawdown episodes for the CVaR mechanism to learn from, relative to the 2019–2023 test period, which includes both the COVID-19 crash and the 2022 bear market. This is consistent with CPPO-DeepSeek’s improved drawdown: the LLM risk signal provides contemporaneous, test-time information that the CVaR mechanism alone, trained only on 2013–2018 statistics, does not have access to.

5.2 Multi-Seed Observations

Table 2 reports the mean and standard deviation across our two seeds (42 and 123). We stress that with $n = 2$, these are descriptive statistics, not estimates with any formal statistical guarantee, and should be read as preliminary evidence of instability rather than as a precise characterization of variance. With that caveat, the single-seed pattern from Section 5.1 does not fully hold across the two seeds we test: PPO-DeepSeek has both the highest mean return and the largest spread (std = 0.286) of any agent, and its Sharpe ratio is in fact lower than plain PPO’s on seed 123 (0.696 vs. 0.725), reversing

Table 1: Backtest performance on Nasdaq-100, 2019–2023, seed 42.

Agent	Total Return	Sharpe	Max Drawdown	CVaR (5%)
PPO	1.4472	0.824	−0.394	−0.0388
CPPO	1.1301	0.746	−0.412	−0.0361
PPO-DeepSeek	1.5888	0.862	−0.412	−0.0384
CPPO-DeepSeek	1.3099	0.794	−0.345	−0.0379
Buy & Hold	2.3500	0.991	−0.339	−0.0422

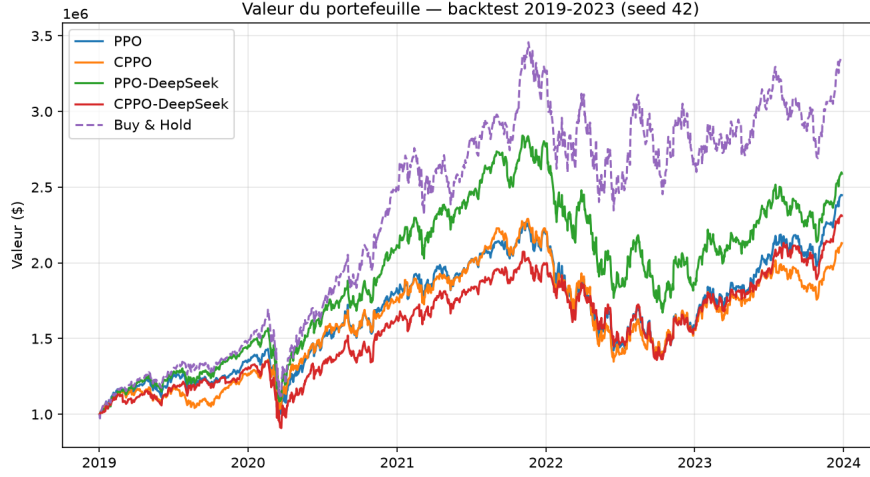


Figure 2: Portfolio value over the 2019–2023 backtest period (seed 42), for all four agents and the equal-weight buy-and-hold baseline.

the single-seed ranking. In contrast, CPPO-DeepSeek shows the smallest spread in total return of any agent (std = 0.003, an order of magnitude below the others) while still improving mean return and drawdown relative to CPPO alone. Figure 3 visualizes this spread across seeds. We view this asymmetry — the CVaR-constrained LLM variant appearing more consistent than the unconstrained one — as the paper’s most interesting preliminary observation, and one that a larger-scale multi-seed study should test directly.

Table 2: Backtest performance, mean $\pm$ std across seeds 42 and 123 ($n = 2$; descriptive only, see Section 5.2).

Agent	Total Return	Sharpe	Max Drawdown	CVaR (5%)
PPO	1.361 ± 0.123	0.775 ± 0.070	-0.385 ± 0.013	-0.0404 ± 0.0023
CPPO	1.246 ± 0.164	0.782 ± 0.052	-0.376 ± 0.051	-0.0369 ± 0.0011
PPO-DeepSeek	1.387 ± 0.286	0.779 ± 0.117	-0.396 ± 0.022	-0.0406 ± 0.0031
CPPO-DeepSeek	1.312 ± 0.003	0.763 ± 0.045	-0.363 ± 0.025	-0.0404 ± 0.0035

6 Limitations

Sample size. Our multi-seed analysis uses only $n = 2$ seeds due to compute and time constraints, which is insufficient for formal statistical significance testing and does not, on its own, establish the kind of multi-seed evaluation protocol advocated by Henderson et al. [2018]. We present our results as preliminary, motivating evidence rather than a settled finding, and recommend $n \geq 5$ seeds, ideally with a formal variance or significance test, before drawing firm conclusions about relative agent stability. We leave this to future work.

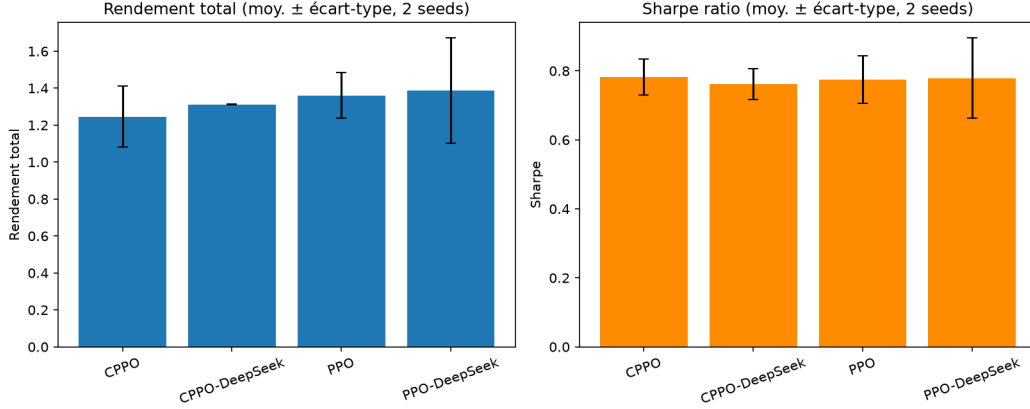


Figure 3: Total return and Sharpe ratio (mean $\pm$ standard deviation across our two seeds) for each agent. Error bars from $n = 2$ should be read as descriptive, not as confidence intervals.

Survivorship bias. As in the original paper, we use the current Nasdaq-100 constituent list applied retroactively to 2013–2023, which excludes companies that were delisted or removed from the index (e.g. via acquisition) during the period, potentially inflating reported returns.

Simplified risk aggregation. Our CPPO-DeepSeek implementation scales the daily CVaR loss by the unweighted mean risk factor across tickers, rather than the portfolio-weighted aggregation described in the original paper. This simplification may understate the effect of LLM risk signals on positions with large portfolio weight.

Reliance on pre-generated LLM signals. Due to API budget constraints, we reuse the DeepSeek V3 sentiment and risk scores released by Benhenda [2025] rather than regenerating them independently. Our reproduction therefore validates the RL and CVaR components end to end, but does not independently verify the LLM signal generation step.

Single hyperparameter configuration. We did not perform hyperparameter tuning for either the PPO/CPPO optimization parameters or the CVaR threshold d ; results may be sensitive to these choices, and we cannot rule out that a different threshold would change the CPPO vs. PPO comparison discussed in Section 5.1.

7 Broader Impact

This work studies automated trading agents on historical, publicly available market data, and we do not deploy any agent with real capital. We nonetheless note two risks relevant to this line of research more broadly. First, backtested performance on historical data, even when out-of-sample, does not guarantee future performance, and results here should not be interpreted as investment advice or as evidence that any of these agents would perform similarly if deployed live; overfitting to a particular historical backtest window is a well-known risk in quantitative finance, and our small-seed evaluation is, if anything, an illustration of how easily a single favorable run can overstate an agent’s reliability. Second, LLM-infused trading signals derived from public news could, at scale, contribute to correlated trading behavior across market participants using similar underlying models, a systemic-risk concern that is outside the scope of this paper’s backtest-based evaluation but that we believe merits attention as such systems see wider adoption. We did not identify concerns specific to this paper requiring an ethics review under the NeurIPS ethics guidelines.

8 Conclusion

We reproduced the FinRL-DeepSeek pipeline from scratch and confirmed, on a single seed, that LLM signal infusion improves risk-adjusted performance for both unconstrained and CVaR-constrained PPO trading agents on Nasdaq-100. A small-scale, two-seed robustness check complicates this picture: the unconstrained LLM variant (PPO-DeepSeek) shows the largest spread across our two seeds,

while the CVaR-constrained LLM variant (CPPO-DeepSeek) shows the smallest, alongside improved mean return and drawdown relative to CPPO alone. Consistent with long-standing concerns about seed sensitivity in deep RL [Henderson et al., 2018], we present these observations as preliminary rather than conclusive, and as motivation for larger-scale multi-seed evaluation as standard practice in financial reinforcement learning research — an area where, to our knowledge, this concern is rarely addressed explicitly. Future work should extend this analysis to a larger number of seeds with formal variance estimates, independently regenerated LLM signals, and a portfolio-weighted risk aggregation matching the original formulation.

References

- Mostapha Benhenda. FinRL-DeepSeek: LLM-Infused Risk-Sensitive Reinforcement Learning for Trading Agents. *arXiv preprint arXiv:2502.07393*, 2025.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- R. Tyler Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000.
- Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance. *arXiv preprint arXiv:2011.09607*, 2021.
- Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A Comprehensive Financial News Dataset in Time Series. *arXiv preprint arXiv:2402.06698*, 2024.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning that Matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.