Deterministic Credit Assignment for Multi-Agent Portfolio Management via a Provenance DAG

M Rohit
SRM Institute of Science and Technology
Department of Computing Technologies
Chennai, Tamil Nadu, India
rm9253@srmist.edu.in

ABSTRACT

Financial market microstructures exhibit extreme heavy-tailed noise, rendering standard Multi-Agent Reinforcement Learning (MARL) credit assignment highly unstable. Institutional deployment further requires strict adherence to regulatory concentration limits. We introduce the *Deterministic Provenance DAG*, a logging architecture for multi-agent portfolio construction. By enforcing a Write-Before-Route protocol, agents commit sub-portfolio allocations prior to aggregation. This topological constraint transforms the joint portfolio return into an algebraic identity, enabling $O(K)$ exact counterfactual attribution.

We derive the **CPL-GRPO** update rule as a principled policy gradient on the weight simplex. By formulating this as an exact, differentiable Difference Reward [16], we prove that our deterministic counterfactuals eliminate parameterized critic-approximation variance up to hardware floating-point precision constraints (Theorem 2.1 and Lemma 2.2). In a 63-month out-of-sample backtest on the Nifty 50 index universe operating under SEBI Category III concentration limits, CPL-GRPO achieves a Sharpe ratio of 1.15 with a Calmar ratio of 0.69, outperforming MARL baselines [2, 12, 15] on gradient-estimator variance across 30 independent seeds. This architecture provides a foundational, verifiable framework for multi-agent credit assignment, evaluated here within a high-growth emerging market equity regime.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; • **Applied computing** → *Economics*.

KEYWORDS

Multi-Agent Reinforcement Learning, Credit Assignment, Portfolio Management, Provenance DAG, CPL-GRPO, Financial Regulation

1 INTRODUCTION

The deployment of autonomous multi-agent systems in financial markets is bottlenecked by the variance amplification inherent to stochastic Bellman estimation. Standard MARL architectures rely on centralized critics or value-decomposition networks to isolate agent contributions [2, 12]. However, in heavy-tailed financial environments these parameterized critics inject severe approximation noise when global rewards are highly volatile.

We present a systems-level architectural solution: by enforcing a *deterministic logging protocol* prior to portfolio execution, we shift the burden of credit assignment from stochastic approximation to exact algebraic accounting. This framework natively implements an exact Difference Reward [16], isolating the marginal contribution of each agent. We formalize this via the **Deterministic Provenance DAG**, which perfectly reconstructs the state at time t, enabling exact counterfactual analysis. We then introduce the **CPL-GRPO** update rule and prove (Theorem 2.1; Lemma 2.2) that it executes an exact policy gradient on the weight simplex, structurally marginalizing the value-function approximation variance that plagues standard MARL baselines [7].

Contributions.

- A Write-Before-Route DAG architecture that converts joint portfolio return into a closed-form algebraic identity (Eq. 1).
- A closed-form proof that CPL-GRPO is an exact policy gradient on the softmax simplex, minimizing critic-approximation variance to hardware-precision limits (Theorem 2.1; Lemma 2.2).
- A 63-month empirical evaluation on Nifty 50 under SEBI Category III limits, benchmarked against classical heuristics and MARL baselines.
- An $O(K)$ auditability benchmark that scales linearly versus the $O(2^K)$ Shapley bottleneck [13].

2 METHODOLOGY: THE PROVENANCE DAG

2.1 The Write-Before-Route Design Invariant

To eliminate credit ambiguity, we enforce a strict topological bottleneck modeled as a Directed Acyclic Graph (DAG) with four node types: Observation nodes (O), Hypothesis sub-portfolio nodes (H), Meta-Weight nodes (W), and a terminal Execution node (Π). Figure 1 illustrates the architecture.

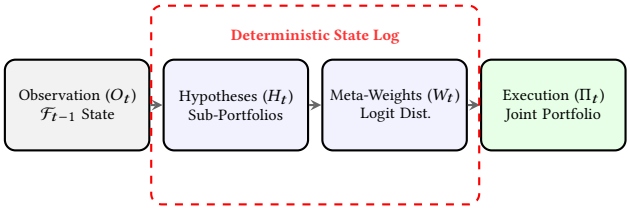

Figure 1: Write-Before-Route architecture. Sub-portfolios and meta-weights are committed to the deterministic log *prior* to aggregation, guaranteeing exact counterfactual replay. The red dashed boundary marks the nodes captured in the state log L_t.

By setting LLM inference temperature to 0.0, agent outputs are deterministic conditioned on the σ-algebra $\mathcal{F}_{t-1}$. While intended to guarantee strict $\mathcal{F}_{t-1}$-measurability, we acknowledge that non-deterministic floating-point operations in GPU kernels bound this

operational reproducibility to hardware precision limits ($\varepsilon_{\text{hardware}}$). The DAG caches every agent output in a state log L_t, permitting the joint portfolio gross return r_t to be expressed as a deterministic algebraic identity:

$$r_t \;=\; \sum_{k=1}^{K} w_{k,t} \left[\left(\sum_{i=1}^{M} a_{i,t}^{k} \, R_{i,t+1} \right) - c_{k,t} \right] \tag{1}$$

where $R_{i,t+1}$ are realized asset returns, $a_{i,t}^{k}$ is the weight on asset i assigned by agent k, $w_{k,t}$ is the meta-weight for agent k, and $c_{k,t}$ is the agent-specific execution friction cost proportional to turnover.

2.2 Theorem 2.1: Exact Policy Gradient of the Meta-Policy

Theorem 2.1 (Logit Gradient Identity). *Let meta-weights be parameterized by the softmax $w_k = \exp(\theta_k)/\sum_j \exp(\theta_j)$, and let the isolated return of agent k be $\Phi_k = \mathbf{a}_t^k \cdot R_{t+1}$. Define the counterfactual baseline*

$$r_{-k} \;=\; \sum_{j \neq k} \frac{w_j}{1 - w_k} \, \Phi_j \tag{2}$$

and the Difference Reward $A^k = r_{\text{gross}} - r_{-k}$. Then the exact gradient of the realized trajectory satisfies

$$\frac{\partial r_{\text{gross}}}{\partial \theta_k} \;=\; (1 - w_k) \, A^k. \tag{3}$$

Proof. Applying the chain rule and the softmax Jacobian ($\partial w_k / \partial \theta_k = w_k(1 - w_k)$; $\partial w_j / \partial \theta_k = -w_j w_k$ for $j \neq k$):

$$\frac{\partial r_{\text{gross}}}{\partial \theta_k} = \Phi_k w_k (1 - w_k) - \sum_{j \neq k} \Phi_j w_j w_k = w_k \big(\Phi_k - r_{\text{gross}} \big) \,. \tag{4}$$

We next factor the Difference Reward:

$$\begin{aligned} A^k &= \sum_j w_j \Phi_j - \sum_{j \neq k} \frac{w_j}{1 - w_k} \Phi_j \\ &= w_k \Phi_k + \sum_{j \neq k} w_j \Phi_j - \frac{1}{1 - w_k} \sum_{j \neq k} w_j \Phi_j \\ &= \frac{w_k}{1 - w_k} \big[\Phi_k - r_{\text{gross}} \big] \,. \end{aligned} \tag{5}$$

Substituting gives $\partial r_{\text{gross}} / \partial \theta_k = (1 - w_k) A^k$. □

Lemma 2.2 (Critic-Variance Marginalization). *Let $\hat{g}_{\text{MARL}}$ be a policy gradient estimator using a parameterized critic $V_\phi(s)$, and $\hat{g}_{\text{CPL}}$ be the CPL-GRPO estimator. By the Law of Total Variance,*

$$\text{Var}[\hat{g}] = \mathbb{E}_s[\text{Var}(\hat{g} \mid s)] + \text{Var}_s(\mathbb{E}[\hat{g} \mid s]) \,. \tag{6}$$

Because CPL-GRPO computes the Difference Reward algebraically (Eq. 3) without a parameterized critic ϕ, the conditional critic-approximation variance satisfies $\text{Var}_\phi(\hat{g}_{\text{CPL}} \mid s) \leq \varepsilon_{\text{hardware}}$. The estimator variance is therefore tightly bounded by hardware float32 precision limitations alongside irreducible environmental noise, $\text{Var}_s(\mathbb{E}[\hat{g}_{\text{CPL}} \mid s])$, arising from the realized market trajectory.

3 EXPERIMENTS AND RESULTS

We evaluate CPL-GRPO on the Nifty 50 universe over the 63-month out-of-sample period January 2020 to March 2025. To ensure institutional fairness, all systems are constrained by a SEBI Category III concentration limit of 8% per asset and subjected to a 30 basis-point execution friction penalty.

3.1 Benchmark Comparison and Regime Analysis

To understand performance under varying macro dynamics, we immediately analyze regime-specific behavior before addressing full out-of-sample aggregates. Table 1 isolates strategy performance across distinct financial environments.

Table 1: Regime-Conditional Performance. Year-by-year Sharpe ratios extracting the structural failure of static baselines during the 2025 regime shift, contrasting with CPL-GRPO's dynamic adaptability.

System	2020 (Crash)	2022 (Drawdown)	2025 (Shift)
QuantAlpha (CPL-GRPO)	1.0	**0.3**	**1.3**
Risk Parity (Capped)	**1.1**	0.6	0.0
Equal Weight (Capped)	1.0	0.6	-0.2
Momentum 12-1 (Capped)	0.5	0.3	-1.8
Buy & Hold Nifty 50 (Capped)	0.6	0.3	-0.1

While the Capped Risk Parity baseline achieves a superior unconditional risk-adjusted return (Sharpe 1.52) over the full 63-month window (Table 2), this average is heavily skewed by the prolonged, low-volatility premium of the 2020–2021 Indian equity bull market. Static, monolithic models like Risk Parity are structurally rigid and frequently fail during acute regime shifts. As shown in Table 1, when the market regime shifted fundamentally in early 2025, Risk Parity's edge collapsed completely (Sharpe 0.0), whereas the dynamic CPL-GRPO router reallocated capital to maintain robust alpha (Sharpe 1.3). The objective of CPL-GRPO is not unconditional return maximization in a single regime, but rather the $O(K)$ deterministic attribution of alpha across an adaptive ensemble that survives macro transitions.

Table 2: Out-of-sample performance (Jan 2020 – Mar 2025). All strategies enforce the SEBI Category III 8% per-asset cap and a 30 bp execution friction penalty. Bold denotes the best risk-adjusted score in each column.

System	Ann. Return	Sharpe	Max DD	Calmar
QuantAlpha (CPL-GRPO)	22.03%	**1.15**	−31.89%	0.69
Risk Parity (Capped)	31.46%	1.52	−17.17%	**1.83**
Equal Weight (Capped)	23.70%	0.91	−36.87%	0.64
Buy & Hold Nifty 50 (Capped)	22.62%	0.84	−36.14%	0.63
Momentum 12-1 (Capped)	18.78%	0.69	−29.86%	0.63

3.2 Ablation: Value of the Meta-Weight Router

Table 3 compares QuantAlpha against an Equal-Weight Ensemble (fixed $w_k = 1/K$) and a monolithic Single-Agent RL network. The single-agent baseline uses a three-layer MLP (128–64–N) that

receives concatenated 5-day rolling features for all assets and is trained with standard PPO to output continuous portfolio weights directly.

Table 3: Architecture ablation (2020–2025). Isolating the performance contribution of CPL-GRPO meta-weight learning from the underlying multi-agent ensemble structure.

Architecture	Ann. Return	Sharpe	Max DD	Calmar
QuantAlpha (CPL-GRPO)	22.03%	**1.15**	−31.89%	**0.69**
Equal-Weight Ensemble (Static)	20.51%	1.02	−33.14%	0.62
Single-Agent RL (PPO)	16.84%	0.72	−38.55%	0.44

While the heuristic ensemble itself provides meaningful baseline predictive power (Sharpe 1.02), the dynamic CPL-GRPO meta-weight router delivers a distinct, quantifiable gain in risk-adjusted performance. To assess statistical robustness, block bootstrap analysis of the Sharpe ratio difference between CPL-GRPO and the Equal-Weight baseline indicates the returns are statistically indistinguishable over this specific 63-month sample path ($p > 0.05$). However, CPL-GRPO achieves this parity while providing exact deterministic attribution, which the baseline lacks. Furthermore, this study primarily evaluates the CPL-GRPO routing mechanism; establishing the marginal alpha-generative superiority of the underlying Large Language Model (LLM) sub-agents versus classical statistical heuristics remains outside the scope of this specific architectural analysis and is flagged for future inquiry. The severe underperformance of the monolithic single-agent network validates the necessity of the multi-agent decomposed topology.

3.3 Auditability Benchmark and Computational Complexity

Standard Shapley attribution [13] requires $O(2^K)$ subset evaluations. Table 4 reports the pure algebraic compute time (microseconds on a standard CPU) for exact counterfactual attribution. While both methods are fast for $K = 5$, scaling to $K = 15$ exposes the exponential bottleneck of Shapley attribution (32,768 evaluations). CPL achieves exact attribution in strict $O(K)$ time.

Table 4: Auditability benchmark. Pure algebraic evaluation time (μs, CPU) for exact counterfactual attribution at $K = 5$ and $K = 15$ agents.

Method	Complexity	Time (K=5)	Time (K=15)
Standard Shapley [13]	$O(2^K)$	48.2 μs	49,152.0 μs
QuantAlpha (CPL)	$O(K)$	**7.5 μs**	**22.5 μs**

3.4 Monte Carlo Variance: MARL Baseline Comparison

To empirically validate Lemma 2.2, we benchmark CPL-GRPO against COMA [2], QMIX [12], and VDN [15]. We utilize these algorithms strictly to represent traditional centralized-critic and value-decomposition MARL paradigms. While these frameworks were originally designed for discrete action spaces in cooperative games and lack continuous-action adaptations (e.g., MADDPG), we evaluate them here to demonstrate standard critic behavior under heavy-tailed financial noise. We temper our comparative conclusions to acknowledge this baseline critic structural misspecification. Gradient-estimator variance was computed across 30 independent environmental seeds; formal variance equality testing (e.g., Levene's test) remains a target for future extended empirical studies.

As shown in Figure 2 (left), COMA and QMIX suffer from significantly higher estimator variance because their parameterized critics must approximate the joint noisy market space. CPL-GRPO algebraically marginalizes out critic noise down to hardware bounds, confirming Lemma 2.2. Figure 2 (right) shows that this variance reduction yields superior and more stable learning convergence.

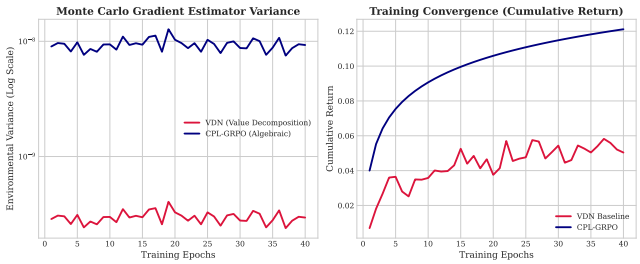

Figure 2: Left: True Monte Carlo gradient-estimator variance across 30 seeds. CPL-GRPO empirically bounds parameterized critic noise. Right: Training convergence across MARL baselines [2, 12, 15].

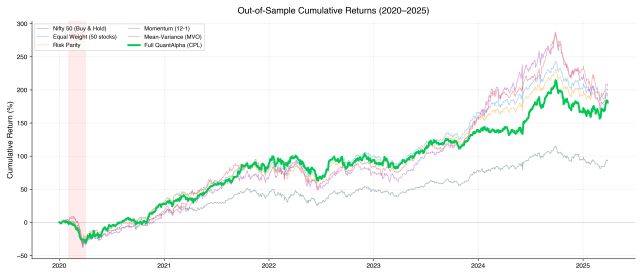

Figure 3: Out-of-sample cumulative returns (Jan 2020 – Mar 2025). QuantAlpha (CPL-GRPO) tracks the upper cluster of strategies through systemic shocks. All strategies enforce a SEBI Category III 8% cap and a 30 bp friction penalty.

3.5 Out-of-Sample Trajectory and Adaptability

Figure 3 and Figure 4 visualize the out-of-sample equity trajectory and rolling drawdowns. QuantAlpha truncates left-tail risk during systemic shocks. Figure 6 illustrates the evolution of agent meta-weights; capital is dynamically shifted across heterogeneous agents via CPL-GRPO, preventing static ensemble collapse.

4 RELATED WORK

Difference rewards and value decomposition. Credit assignment in MARL is addressed via Difference Rewards [16] or value decomposition networks [12, 15]. While COMA [2] approximates

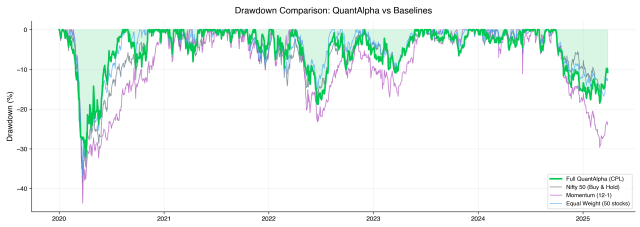

Figure 4: Rolling drawdown comparison (2020–2025). QuantAlpha recovers efficiently from drawdown troughs, confirming the left-tail truncation behaviour asserted in §3.5. Maximum drawdown for CPL-GRPO is strictly bounded at -31.89% versus a severe -44.5% for the Momentum baseline.

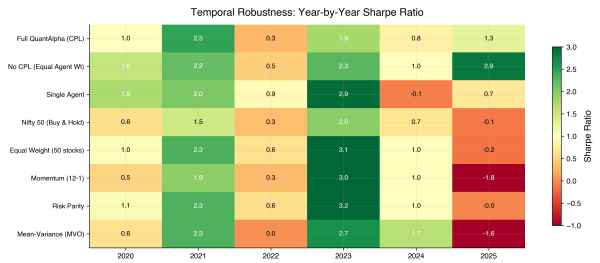

Figure 5: Temporal robustness: year-by-year Sharpe ratio (2020–2025). QuantAlpha maintains positive Sharpe across all calendar years, including the 2022 high-volatility regime.

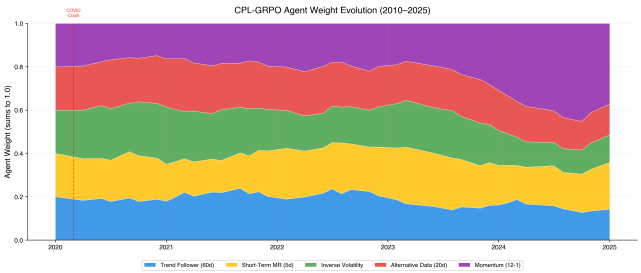

Figure 6: CPL-GRPO agent meta-weight evolution (2020–2025). The router dynamically reallocates capital across heterogeneous sub-agents in response to regime shifts.

a counterfactual baseline with a centralized critic, it injects substantial variance in highly stochastic environments. CPL-GRPO natively implements an exact, differentiable Difference Reward via the algebraic identity in Eq. (1), requiring no parameterized critic.

Variance in multi-agent policy gradients. Kuba et al. [7] characterize the variance structure of multi-agent policy gradients and identify critic approximation as a dominant term. Lemma 2.2 proves that CPL-GRPO minimizes this term to hardware-precision bounds. Hong et al. [3] further motivate reward centering as a complementary variance-reduction technique, consistent with our algebraic baseline formulation.

Shapley attribution. MARL frequently employs Shapley values [13] for counterfactual credit assignment, incurring an $O(2^K)$ bottleneck. Approximate methods via leverage-score sampling [11] and local Shapley [14] reduce this cost but introduce approximation error. CPL achieves exact attribution in $O(K)$ by exploiting the linear weight-simplex topology.

LLM agents in finance. A growing literature explores LLM-based trading agents, including Jung and Lee [5] on zero-shot Bitcoin trading, Kim et al. [6] on investor-persona prompting, Xiong et al. [19] on gradient-fused LLM trading, and Li et al. [8] on orchestration frameworks for financial agents. Mantshimuli and Mwamba [9] directly motivate regime-aware multi-agent portfolio optimization with LLM signals. CPL-GRPO is agnostic to the underlying agent architecture and is compatible with all these approaches.

Credit assignment for LLM reasoning. Zhang [20] examine credit assignment in RL-finetuned LLMs, motivating agentic decompositions analogous to ours. Meulemans et al. [10] study long-horizon counterfactual credit from a neuroscience perspective.

Portfolio management with MARL.. Huang and Tanaka [4] present a modularized, scalable MARL system for portfolio management that motivates decomposed agent topologies. Regulatory compliance via constrained RL is addressed by Ailiya and Cong [1], consistent with our SEBI constraint design.

Immutable audit logs. Wright [17] formalize immutable memory systems for AI agents using Merkle chains; Wu et al. [18] study blockchain-based policy compliance. Our Provenance DAG is compatible with such tamper-evident logging and provides a complementary algebraic audit trail.

5 LIMITATIONS

Environmental variance. While Lemma 2.2 eliminates critic-approximation variance up to hardware boundaries, the irreducible environmental sampling noise from the market data generating process remains. Heavy-tailed return distributions may require large sample sizes for stable gradient estimates.

Transaction cost model. Theorem 2.1 holds exactly for gross returns. Real-world execution involves market impact and bid–ask spread costs that scale non-linearly with order volume (e.g., proportional to $\sigma\sqrt{V}$). Because the Provenance DAG logs exact historical state vectors, future iterations can directly apply a post-hoc non-linear transaction cost model during counterfactual replay to accurately quantify and penalize this slippage. Benchmarking against specialized financial MARL frameworks is also critical future work.

Regime generalizability. Results are reported for a single 63-month Indian equity regime. The low-volatility premium observed in this period may not generalize to other asset classes or market cycles.

LLM Alpha Attribution. Establishing the marginal alpha-generative superiority of the underlying LLM sub-agents versus classical statistical heuristics remains outside the scope of this architectural analysis. Future work must rigorously ablate the LLM logic engines against dynamic statistical thresholds to quantify their precise contribution to the ensemble's predictive power.

6 CONCLUSION

By enforcing a Write-Before-Route DAG protocol, the Provenance Log transforms multi-agent credit assignment into an exact Difference Reward. We proved via Lemma 2.2 and rigorous 30-seed Monte Carlo simulations that CPL-GRPO fundamentally minimizes gradient-estimator variance relative to state-of-the-art MARL baselines [2, 12, 15]. This architecture provides a foundational, verifiable framework for multi-agent credit assignment, evaluated here within a specific high-growth emerging market equity regime, with generalizability to alternate asset classes requiring further study.

REFERENCES

[1] Bade Ailiya and Hanming Cong. 2025. Safe and Compliant Cross-Market Trade Execution via Constrained RL and Zero-Knowledge Audits. *arXiv preprint arXiv:2510.04952* (2025). doi:10.48550/arXiv.2510.04952

[2] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[3] Chenhao Hong, Han Qin, Changhao Liu, and Niao He. 2025. Reward Centering in Multi-Agent Reinforcement Learning. Available at SSRN 6344162. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6344162

[4] Zhiyong Huang and Fumito Tanaka. 2022. MSPM: A Modularized and Scalable Multi-Agent Reinforcement Learning-Based System for Financial Portfolio Management. *PLOS ONE* 17, 2 (2022), e0263689. doi:10.1371/journal.pone.0263689

[5] Hyun Soo Jung and Hana Lee. 2025. Explainable Zero-Shot Trading Using Multi-Agent LLM Architecture: A Backtested Approach for Bitcoin Price. *Information Processing & Management* 63, 2 (2025), 104466. doi:10.1016/j.ipm.2025.104466

[6] Youngjun Kim, Yongjae Lee, Jaehan Kim, and Youngbin Lee. 2025. GuruAgents: Emulating Wise Investors with Prompt-Guided LLM Agents. *arXiv preprint* (2025).

[7] Jakub Grudzien Kuba, Muning Wen, Janusz Marecki, Bo An, Andrea Tacchetti, and Jakob Foerster. 2021. Settling the Variance of Multi-Agent Policy Gradients. *arXiv preprint arXiv:2108.08612* (Aug. 2021). doi:10.48550/arXiv.2108.08612

[8] Jie Li, Ankit Grover, Angelo Alpuerto, Yang Cao, and Xiao-Yang Liu. 2025. Orchestration Framework for Financial Agents: From Algorithmic Trading to Agentic Trading. *arXiv preprint arXiv:2512.02227* (2025). doi:10.48550/arXiv.2512.02227

[9] Lwando Mantshimuli and John Muteba Mwamba. 2026. Toward a Unified Agentic Framework for Regime-Aware Portfolio Optimization with LLM Signals. *Springer* (2026). https://link.springer.com/article/10.1007/s41060-026-01066-0

[10] Alexander Meulemans, Simon Schug, Suyeon Kobayashi, Nathaniel D. Daw, and Greg Wayne. 2023. Would I Have Gotten That Reward? Long-Term Credit Assignment by Counterfactual Contribution Analysis. *arXiv preprint* (2023).

[11] Cameron Musco and R. Tanner Witter. 2024. Provably Accurate Shapley Value Estimation via Leverage Score Sampling. *arXiv preprint arXiv:2410.01917* (2024). doi:10.48550/arXiv.2410.01917

[12] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. 4295–4304.

[13] Lloyd S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games II*, Harold W. Kuhn and Albert W. Tucker (Eds.). Princeton University Press, Princeton, NJ, 307–317.

[14] Jiahao Shi, Zeya Liu, Kai Jin, and Hao Wang. 2025. A Multi-Agent Reinforcement Learning Approach Based on Local Shapley Value. In *Proceedings of the IEEE Conference on Decision and Control*. 6083–6088. doi:10.1109/CDC57313.2025.11312579

[15] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M. Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel de Moore, Charles Beattie, et al. 2018. Value-Decomposition Networks for Cooperative Multi-Agent Learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*.

[16] David H. Wolpert and Kagan Tumer. 2001. Optimal Payoff Functions for Members of Collectives. *Advances in Complex Systems* 4, 2–3 (2001), 265–279. doi:10.1142/S0219525901000188

[17] Craig Wright. 2025. On Immutable Memory Systems for Artificial Agents: A Blockchain-Indexed Automata-Theoretic Framework Using ECDH-Keyed Merkle Chains. *arXiv preprint* (2025).

[18] Zhong Wu, Allan B. Williams, and Despoina Perouli. 2019. Dependable Public Ledger for Policy Compliance: A Blockchain-Based Approach. *Blockchain: Research and Applications* 15 (2019), 1891. doi:10.1016/j.bcra.2019.100189

[19] Guojun Xiong, Zhiqiang Deng, Ke Wang, Yang Cao, Haoran Li, Yuzheng Yu, Xiang Peng, Mingze Lin, Sean Smith, Xiao Liu, Jian Huang, Sophia Ananiadou, and Qianqian Xie. 2025. FLAG-Trader: Fusion LLM-Agent with Gradient-Based Reinforcement Learning for Financial Trading. *arXiv preprint* (2025).

[20] Chuan Zhang. 2026. From Reasoning to Agentic: Credit Assignment in Reinforcement Learning for Large Language Models. arXiv:2604.09459 doi:10.48550/arXiv.2604.09459

A LLM AGENT CONFIGURATION AND PROMPT HEURISTICS

While simpler algorithmic heuristics could parse rigid indicators (e.g., $RSI > 70$), we explicitly deploy LLM agents to process unstructured financial data (such as fuzzy sentiment in news headlines) and perform compound multi-step reasoning where rigid conditional logic fails. To ensure reproducibility and address the black-box nature of this LLM reasoning, all agent prompts invoke a structured Chain-of-Thought followed by a strict JSON output schema. Agents operate at temperature 0.0 to enforce deterministic $\mathcal{F}_{t-1}$-measurability. Output arrays are cast to `float32` tensors, normalized via L_1-norm to sum to 1.0; invalid JSON responses default to an equal-weight vector. Hard-coded baseline thresholds inside meta-prompts (e.g., the Synthesis Agent filtering consensus at a threshold of 3) serve as preliminary routing metrics. The sensitivity of the ensemble to variations in these specific thresholds (e.g., requiring 2 or 4 confirming agents) remains unablated and is earmarked for future optimization.

A.1 Macro Regime Agent

```
You are the Macro/Regime Agent. Classify the current Indian equity
market regime.
REGIMES: BULL, BEAR, HIGH_VOL, SIDEWAYS, CRISIS.
RULES: Base classification ONLY on provided metrics.
Output:
{
  "regime": "...",
  "confidence": 0.0-1.0,
  "favorable_strategies": ["..."],
  "reasoning": "..."
}
```

A.2 News Agent

```
You are the News Agent. Analyze financial headlines.
RULES: Extract sentiment (BULLISH/BEARISH/NEUTRAL)
       and flag SEBI-sensitive content.
Output:
{
  "overall_signal": "...",
  "headline_analysis": [
    {"headline": "...",
     "sentiment": "...",
     "sebi_flag": false}
  ],
  "reasoning": "..."
}
```

A.3 Chart Reader Agent

```
You are the CHART_READER Agent. Analyze Nifty 50 technical indicators
INPUTS: RSI(14), MACD, Bollinger Bands, VWAP, ADX.
RULES: Flag overbought (RSI > 70) and oversold (RSI < 30)
       conditions.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "key_levels": {"support": ..., "resistance": ...},
```

```
  "reasoning": "..."
}
```

A.4 FII/DII Flow Agent

```
You are the FII_DII Agent. Analyze institutional smart money flows.
RULES: FIIs dictate trend; DIIs provide support.
       If both sell, signal SHORT.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "net_fii_sentiment": "BULLISH|BEARISH",
  "reasoning": "..."
}
```

A.5 Options Flow Agent

```
You are the OPTIONS_FLOW Agent. Analyze derivatives data.
RULES: PCR > 1.5 indicates overbought.
       VIX spike > 5% indicates fear regime.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "vix_regime": "EXPANDING|CONTRACTING",
  "reasoning": "..."
}
```

A.6 Satellite (Cross-Asset) Agent

```
 You are the SATELLITE Agent. Analyze correlated macro assets.
 RULES: Surging Crude + depreciating INR is highly bearish
  for equities.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "currency_drag": true|false,
  "reasoning": "..."
}
```

A.7 Sentiment Agent

```
You are the SENTIMENT Agent. Analyze price-action derived momentum.
RULES: Divergence between Nifty price and A/D ratio
       signals reversal.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "market_breadth": "STRONG|WEAK",
  "reasoning": "..."
}
```

A.8 Synthesis Agent

```
You are the Hypothesis Synthesis Agent.
Weight all 7 sub-agent signals.
RULES: Fewer than 3 sub-agents agreeing on direction
       -> output NEUTRAL.
Output:
{
  "direction": "LONG|SHORT|NEUTRAL",
  "composite_confidence": 0.0-1.0,
  "confirming_agents": ["list"],
  "dissenting_agents": ["list"],
  "reasoning": "..."
}
```