

FINEVO: Financial Language Agents Shape the Markets That Evaluate Them

Mingxi Zou^{1,2,3,*}, Jiaxiang Chen^{2,*,†}, Langzhang Liang^{1,2}, Zhuo Wang²,
Aotian Luo^{1,2}, Chi Zhang³, Dongning Sun^{4,†}, and Zenglin Xu^{1,2,†}

¹Shanghai Academy of AI for Science (SAIS) ²Fudan University

³Tensorpacific ⁴Peng Cheng Laboratory

*Equal contribution. †Corresponding authors. ‡Project Lead.

{mxzou24, jiaxiangchen23}@m.fudan.edu.cn

Abstract

Financial language agents are often evaluated on fixed historical price paths, leaving unspecified how text-conditioned decisions interact with competing policies and the market in which they are assessed. We introduce FINEVO, a *text-to-market evaluation* environment in which time-stamped financial text is translated into executable orders, cleared in a shared simulated market, and evaluated under population feedback. FINEVO defines a *closed-loop evaluation signature* that jointly records a policy’s local outcome, market concentration, strategic reorganization, and regime-conditioned standing. This makes the evaluation target the policy in protocol, rather than its return against a frozen tape. Across controlled disturbances and historical-information settings, FINEVO makes changes in market organization observable alongside local returns. Ablations show that population-feedback mechanisms shape whether the evaluator remains contestable. The contribution is an evaluation protocol, not a new language model or evidence that language policies outperform forecasting baselines. FINEVO studies how financial-language interfaces behave when their actions help generate the trajectories on which they are evaluated.

1 Introduction

Financial text has long been used to measure risk, media sentiment, and the information content of news (Kogan et al., 2009; Tetlock, 2007; Chang et al., 2016). Financial language models now read earnings calls, news, filings, and market commentary to forecast prices or produce trading decisions (Zhou et al., 2021; Ding et al., 2015; Hu et al., 2018; Huang et al., 2023; Wu et al., 2023; Yu et al., 2024; Ma et al., 2025). Standard evaluation usually presents such a model with a historical record, asks for a decision, and scores that decision against a price path fixed outside the model. This is useful

for measuring a decision interface under replay, but it does not define the shared market in which the decision is evaluated once other policies can react.

That missing protocol matters for financial NLP. A text-conditioned policy receives a scheduled information stream, maps it into an order, and enters a market populated by policies with numerical, rule-based, and reinforcement-learning interfaces. The joint orders determine executions, prices, and the local outcomes recorded at the next step. A meaningful evaluation must consequently state both the financial-language interface and the market feedback through which the interface is assessed. It should not reduce the question to either text prediction or one scalar return on a frozen tape.

We call this object *text-to-market evaluation*. Its unit is a text-conditioned policy *in protocol*: a declared text-access schedule, an executable decision interface, a shared market transition, competing policies, and an update rule that changes the competitive population over time. This framing draws on market microstructure, where heterogeneous order submission jointly shapes liquidity, order flow, and price formation (Farmer and Joshi, 2002; Chiarella et al., 2009; Lillo et al., 2003). It makes a specifically NLP-facing question testable: how should a language-mediated financial action be evaluated when the trajectory on which it is judged is generated inside the protocol?

We introduce FINEVO, a closed-loop evaluation market for this setting. Time-stamped financial information reaches heterogeneous policies through their declared interfaces; orders meet in a common continuous double auction; local outcomes are recorded; and an archetype-level selection signal reallocates evaluation capital before the next epoch. Figure 1 shows the loop. The update combines performance-sensitive selection, innovation toward a broad prior, and stochastic perturbation, making concentration, reorganization, and historical-regime standing observable outputs rather than hid-

den simulator behavior.

The paper asks three tightly connected questions.

RQ1 (Joint disturbance response): How does a declared joint perturbation to text, the fundamental anchor, and liquidity reorganize an endogenous market? **RQ2 (Historical-regime standing):** How do documented text-conditioned interfaces occupy competitive positions across historical information regimes when their actions contribute to the price path on which they are evaluated? **RQ3 (Contestability):** Which feedback mechanisms prevent a closed-loop evaluator from collapsing into persistent dominance or unstructured churn?

The evidence follows this chain at the level of the declared protocol. Controlled disturbances make the response to the declared joint condition visible; historical information regimes expose changes in competitive standing; and mechanism ablations show how population feedback keeps the evaluator contestable. Together, these conditions make it possible to interpret episode outcomes in the shared market that records them.

The methodological contribution is the policy-in-protocol evaluation contract and its multidimensional report; the GPT-4o-mini policies are evaluated instantiations of that contract rather than new language-model methods. Our contributions are threefold. First, we formalize *text-to-market evaluation* for financial language agents, explicitly connecting a time-stamped text interface to executable actions, a shared market, and population feedback. Second, we introduce a *closed-loop evaluation signature* that reports local outcome together with market structure, reorganization, and historical-regime standing. Third, across controlled and historical information conditions plus feedback-mechanism ablations, we show why competitive context belongs in evaluating financial-language interfaces; we do not claim that these experiments isolate language-model quality or establish language-policy profitability.

2 Related Work

Financial text, events, and trading. Financial NLP connects textual events to market-relevant information. Studies of corporate disclosures, media tone, and financial-news semantics link text to risk, sentiment, and market outcomes (Kogan et al., 2009; Tetlock, 2007; Chang et al., 2016). Event-driven models convert identified news into trading signals (Zhou et al., 2021; Ding et al., 2015;

Hu et al., 2018), while datasets such as Astock pair stock-specific news with market factors for NLP-aided trading evaluation (Zou et al., 2022). Work on financial question answering and domain-adapted models further develops reasoning over reports, news, and numerical context (Chen et al., 2021; Huang et al., 2023; Wu et al., 2023; Liu et al., 2023; Xiong et al., 2025). These settings characterize a textual input or a local decision. FINEVO instead specifies the market protocol that evaluates an executable text-conditioned action jointly with competing actions.

Language agents for financial decision making.

LLM-based trading systems enrich conventional pipelines with reasoning, reflection, memory, and tool use (Yao et al., 2023; Wei et al., 2022; Yu et al., 2024; Xiao et al., 2024; Yu et al., 2025b). Recent benchmarks span financial competence, multi-asset sequential decision making, and financial-agent evaluation (Xie et al., 2024; Li et al., 2025; Guo et al., 2025); related FinNLP shared tasks evaluate agent-based decision making across assets and methods (Yu et al., 2025a). Language-agent trading arenas also demonstrate the value of allowing actions to enter a market rather than only replaying history (Ma et al., 2025). Our distinction is to make the *evaluation object* explicit: a language policy is assessed with a declared text interface, a competitive population, and a vector of local and market-mediated outputs, rather than as a generic predictor or a single leaderboard entry.

Agent-based market evaluation. Financial reinforcement-learning platforms provide data and market-environment infrastructure (Liu et al., 2022; Sun et al., 2023), while simulators such as ABIDES provide detailed multi-agent exchange infrastructure (Byrd et al., 2020; Amrouni et al., 2021); other systems support reinforcement-learning agents and high-fidelity trading environments (Mascioli et al., 2024; Frey et al., 2023). Double-auction and order-book studies establish that aggregate prices, liquidity, and order flow are jointly generated by exchange rules and heterogeneous trading behavior (Gode and Sunder, 1993; Chiarella and Iori, 2002; Chiarella et al., 2009; Cont et al., 2010). Ecological and evolutionary perspectives further motivate studying how strategies coexist and adapt through relative success (Hommes, 2006; Sandholm, 2010; Brock and Hommes, 1998; Lux and Marchesi, 1999). FINEVO combines these ingredients around a

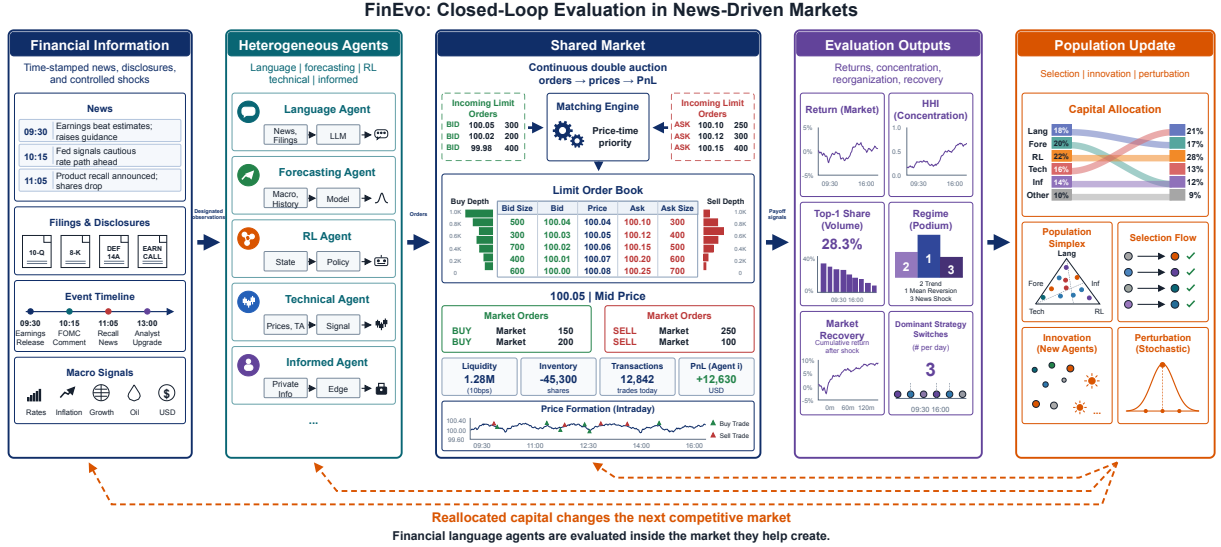


Figure 1: **FINOVO evaluates financial language policies inside a feedback market.** Time-stamped financial information reaches heterogeneous policies through their declared interfaces. Orders meet in a common continuous double auction; an archetype-level selection signal reallocates evaluation capital before the next epoch, changing the competition faced by each interface. Evaluation therefore records both local outcomes and market-level properties such as concentration, reorganization, and historical-regime standing.

FinNLP evaluation formulation: a *text-to-market contract* induces an endogenous evaluation trajectory, and a closed-loop signature records both policy-level outcomes and the shared market state in which they arise.

3 FINEVO: Text-to-Market Evaluation

3.1 Text-Conditioned Decision Interface

We distinguish the three levels used throughout the paper. An *agent instance* j is a stateful simulated trader with its own cash, inventory, and internal state. A *policy* π_i maps the observations available at a decision time to an executable instruction. A *policy archetype* i is the named protocol configuration of that policy, including its observation interface, wake-up schedule, and execution constraints; multiple instances may instantiate one archetype. A policy family groups related archetypes. The unit reported by FINEVO is the policy archetype in protocol, rather than a language model in isolation. Let $g(j)$ map instance j to its archetype and let $\mathcal{L}$ index the text-conditioned archetypes. For an instance j with $i = g(j) \in \mathcal{L}$, let $\mathcal{N}_{\leq t}^{(i)}$ be the time-stamped financial text made available through interface i by its declared schedule up to time t . Given a shared market state z_t and policy state $\tilde{h}_{j,t}$, the interface

produces an executable trading instruction

$$o_{j,t} \sim \pi_i(\phi_i(\mathcal{N}_{\leq t}^{(i)}), z_t, \tilde{h}_{j,t}), \quad (1)$$

$$\text{for } g(j) = i, \quad i \in \mathcal{L}.$$

Here ϕ_i is the documented text-and-context construction for interface i , and $o_{j,t}$ is a market instruction or a null/hold decision. Holdings are supplied to FinCon; they may be empty for the other interfaces. The contract maps scheduled financial text and state to an action that the exchange can clear.

Table 1 instantiates this contract for the four language policies. The interfaces share the same exchange and information schedule but differ in their documented decision construction and executable output. Non-language policies retain their native numerical or rule-based interfaces; detailed wake-up and action specifications appear in Appendix A.

3.2 Shared Market Transition

Active language and non-language agent instances submit their instructions to the same continuous double auction (Kyle, 1985; Gode and Sunder, 1993; Chiarella and Iori, 2002; Cont et al., 2010). An instance that is not scheduled to act at time t emits $o_{j,t} = \emptyset$; scheduled news and the documented wake-up rules determine the active set. We write the resulting market transition as

$$(z_{t+1}, p_{t+1}, \ell_{1:N,t}) \sim \mathcal{M}(z_t, o_{1:N,t}; d_t, \omega_t), \quad (2)$$

Table 1: **Text-to-trade contract for the language policies.** The evaluated unit is a policy archetype under its declared interface; all executable outputs enter the same market transition.

Policy	Textual input	Decision construction	Executable output
ReAct	Scheduled financial news + market context	Reason-act procedure	Fixed-size BUY / SELL / HOLD instruction
CoT	Scheduled financial news + market context	Reasoning trace + final action	Fixed-size BUY / SELL / HOLD instruction
Trading	Scheduled financial news + market context	Structured trading plan	Action + quantity instruction
FinCon	Scheduled financial news + market context + holdings	Constrained trading plan	Constraint-checked action + quantity instruction

where p_{t+1} is the transaction-price state and $\ell_{1:N,t}$ records cleared fills and the resulting local cash, inventory, and payoff records. The declared condition d_t contains scheduled information and, in the controlled suite, the joint text, fundamental, and liquidity disturbance; ω_t denotes simulation randomness. Equation 2 therefore specifies the shared trajectory through which each policy is evaluated, rather than attributing market structure to one policy.

For a language interface π_i , the resulting closed-loop evaluation is a protocol-induced trajectory distribution:

$$\mathcal{E}_i^{\text{CL}} = \mathcal{E}(\pi_i; \mathcal{N}, \Pi_{-i}, \mathcal{M}, \mathcal{U}). \quad (3)$$

It depends on the declared information stream $\mathcal{N}$, competing-policy set Π_{-i} , market transition $\mathcal{M}$, and population update $\mathcal{U}$. It defines the conditions under which a text-conditioned policy’s local and market-mediated outcomes are observed.

3.3 Closed-Loop Evaluation Signature

For a declared information-and-market condition η , FINEVO reports the following closed-loop evaluation signature:

$$\mathcal{S}_i^{\text{CL}}(\eta) = (L_i(\eta), C(\eta), R(\eta), G_i(\eta)). \quad (4)$$

L_i is the policy’s local outcome record (terminal simple return in the reported instantiation). C is shared market concentration, summarized by HHI and Top-1 share; R is shared market reorganization, summarized by dominant-archetype switches; and G_i is the interface’s historical-regime standing. For each reported regime, G_i records the first-, second-, and third-place finish profile over the retained ranking checkpoints. It is reported alongside, rather than folded into, the local outcome record.

Concentration is measured by the Herfindahl–

Hirschman index

$$\text{HHI}(x^q) = \sum_{k=1}^K (x_k^q)^2, \quad (5)$$

where x^q is the allocation of evaluation capital over K policy archetypes at macro epoch q . C and R are shared-market coordinates. The signature therefore evaluates a policy together with the market structure under its declared protocol.

3.4 Population Feedback and Contestability

The market transition and population feedback are distinct. Let $x^q \in \Delta^{K-1}$ denote the archetype-level allocation at macro epoch q . Abstractly, the evaluator updates it through

$$x^{q+1} = \mathcal{U}(x^q, V^q; \theta_{\mathcal{U}}), \quad (6)$$

where V^q is the forward-looking, archetype-level value signal used for population reallocation and $\theta_{\mathcal{U}}$ specifies the update controls. At each five-minute epoch, its target is conditioned on recent realized payoff and exogenous information available in the evaluation state. The implementation updates capital shares as

$$\begin{aligned} r_k^q &= x_k^q \exp(\beta V_k^q + \gamma \sigma_k \epsilon_k^q), & \epsilon_k^q &\sim \mathcal{N}(0, 1), \\ \hat{x}_k^q &= \frac{r_k^q}{\sum_j r_j^q}, & x_k^{q+1} &= (1 - \mu) \hat{x}_k^q + \mu m_k^q. \end{aligned} \quad (7)$$

β controls performance-sensitive selection, μ innovation, γ stochastic perturbation, and σ_k the archetype-specific perturbation scale. The normalized exponential term and convex innovation mixture keep x^{q+1} on the simplex. The update reallocates capital before the next epoch, while reported local return remains a descriptive outcome in L_i .

Selection lets successful archetypes gain influence; a full-support innovation distribution with $\mu > 0$ allocates mass to every archetype at the

Table 2: **Information conditions in the evaluation contract.** Inputs are scheduled outside the market; the listed market record is generated by joint policy interaction.

Condition	External schedule	Endogenous market record
Controlled disturbances	Signed news, fundamental anchor, and liquidity condition	Orders, transaction prices, PnL, and allocation reorganization
Historical-news episode	Time-stamped financial news and disclosures	Orders, price path, order flow, and relative policy-archetype influence

next epoch; and perturbation lets the population respond to the configured stochastic condition. The selection–innovation–perturbation operator is the evaluator’s contestability mechanism. Its empirical role is tested by the ablation in Section 5.

4 Text-to-Market Evaluation Protocol

4.1 Policy Population and Execution Protocol

Default evaluations use $N = 120$ concurrent agent instances drawn from K policy archetypes and operating through the continuous double auction. The population spans technical and informed policies, time-series forecasters, reinforcement-learning traders, and the four text-conditioned interfaces in Table 1. All policies share exchange rules, equal initial capital, and five-minute population-update epochs; each configuration is summarized over 128 independent Monte Carlo runs. The detailed policy roster and operational schedules appear in Appendix A.

Because families retain their native observation and execution interfaces, cross-family results compare complete policy-in-protocol configurations. A ranking therefore reflects the joint effect of the decision model, information available at decision time, state construction, wake-up schedule, and order mapping; it does not isolate model quality under matched information and action interfaces.

4.2 Declared Information-and-Market Conditions

Evaluation separates an externally declared information schedule from the market trajectory generated in response. Table 2 makes this text-to-market boundary explicit.

The controlled suite jointly varies positive and

negative text, fundamental, and liquidity conditions to trace end-to-end information-to-market propagation. The historical record combines a July 7, 2025 intraday episode with bull (October 2024–July 2025) and bear (January–October 2022) regime summaries.

4.3 Estimating the Evaluation Signature

Each run yields an agent-level trading record and a market-level allocation trajectory. We summarize $\mathcal{S}_i^{\text{CL}}(\eta)$ for each declared condition η and report means with 95% confidence intervals over the 128-run suite. The historical podium figures report the aggregate first-, second-, and third-place finish profiles for the two regimes.

4.4 Specifying a Policy Comparison

The protocol also provides a way to specify subsequent comparisons. To compare decision constructions, an evaluator can hold the text schedule, observation fields, competing population, and execution constraints fixed while replacing the focal policy. To compare information access, the focal policy can instead be held fixed while its available text changes. These are different evaluation questions, even when both are summarized by return. In either case, the market must evolve again after the replacement: identical external inputs do not imply identical transaction-price paths once orders change. Reporting the changed component together with the resulting signature makes the comparison explicit. The present suite uses heterogeneous native interfaces to examine competition among complete policy configurations.

4.5 Research Questions

We operationalize the three questions introduced in Section 1 as an evidence map: RQ1 uses controlled disturbances; RQ2 combines the July 7 intraday episode with the bull/bear comparison; and RQ3 examines selection, innovation, and perturbation in turn.

5 Results: Text-to-Market Evaluation

5.1 A Joint Disturbance Reorganizes the Shared Market

Figure 2 addresses RQ1 by showing the shared-market trajectory after a controlled information-and-market disturbance. The positive and negative joint conditions are not mirror images. The positive condition distributes influence across several

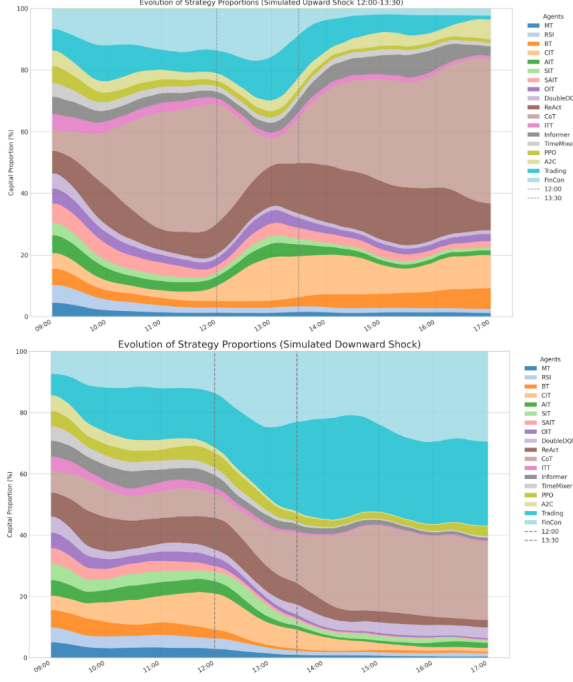


Figure 2: **Closed-loop response to controlled information-and-market disturbances.** Signed disturbances jointly perturb the text, fundamental, and liquidity channels. The upper and lower panels show the positive and negative conditions, respectively. Population allocations reorganize immediately and recover through the selection–innovation update. The suite uses 128 independent runs.

contenders, whereas the negative condition yields a more concentrated response. In both cases, the allocation first reorganizes and then recovers through the selection–innovation update rather than returning mechanically to its pre-disturbance state.

The disturbance is intentionally joint: its text, fundamental, and liquidity channels enter the protocol together. The result therefore traces information-to-market propagation through the full competitive population, showing the structural context in which a text-conditioned policy’s local outcome is recorded. It does not identify the marginal effect of text; that requires factorial channel ablations or text-placebo controls.

The allocation path shows how a declared disturbance redistributes evaluation capital among policies that interact through a common order book. A local gain therefore has a structural context: it can arise alongside a diffuse competitive response or during a concentration of competitive influence. Figure 3 records the corresponding transaction-price path. Together, the two views separate price movement from the redistribution of competitive

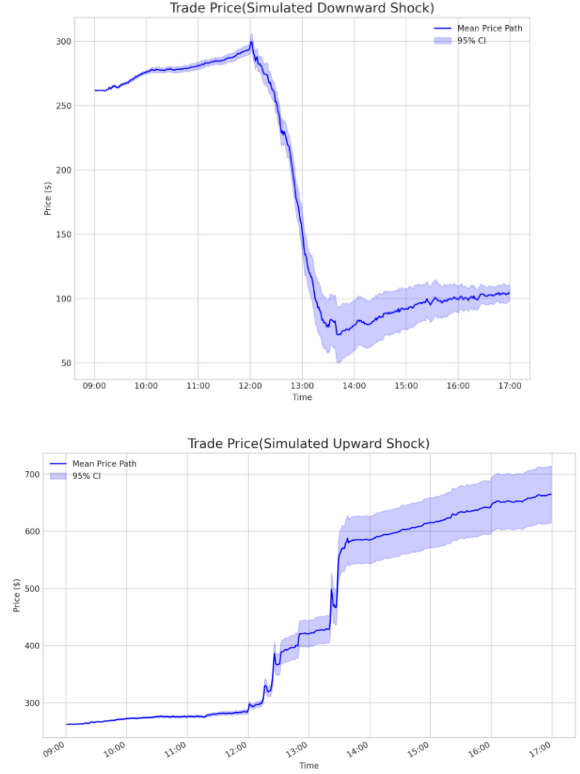


Figure 3: **Simulated transaction-price response under the controlled disturbance protocols.** The traces summarize cross-run mean price paths; the upper and lower panels show the negative and positive conditions, respectively. Shaded bands are the 95% confidence intervals rendered in the source result. Price formation is endogenous to the simulated market.

roles that produced it.

5.2 Observed Standings Differ Across Two Historical Schedules

RQ2 completes the historical signature at two time scales. The July 7, 2025 episode supplies the execution record from which local outcome, concentration, and reorganization are observed; Figure 4 shows its text-to-market allocation path. Figures 5 and 6 then report regime-conditioned standing for October 2024–July 2025 and January–October 2022, respectively.

The two schedules yield different descriptive top-three profiles; their aggregate rates are tabulated in Table 3. In the bull schedule, CoT and ReAct appear above Informer, while in the bear schedule ReAct and FinCon lead and DQN appears above Trading and CoT. The retained aggregate result does not provide run-clustered intervals, so these orderings are descriptive rather than claims of statistically significant differences. Writing $n_{i,c}^{(j)}$ for

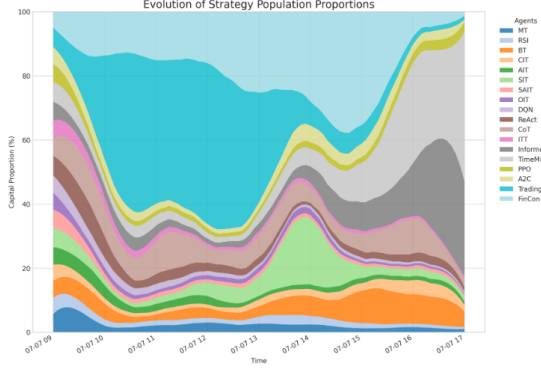


Figure 4: **July 7, 2025 historical-news-conditioned allocation path.** Each band is a policy’s share of evaluation capital over the intraday episode.

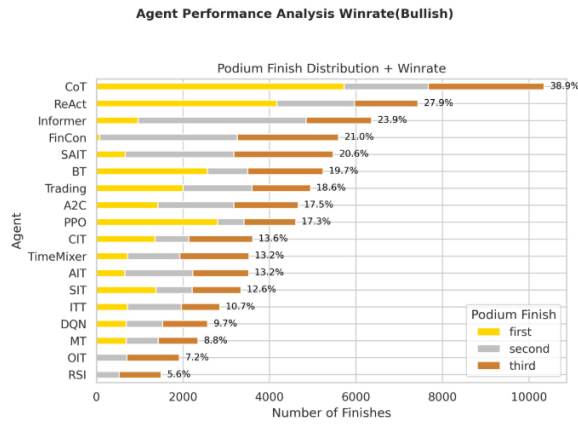


Figure 5: **Bull-period podium profile (October 2024–July 2025).** Stacked segments show the reported first-, second-, and third-place finish frequencies; the legacy vertical-axis label “Agent” denotes a policy archetype.

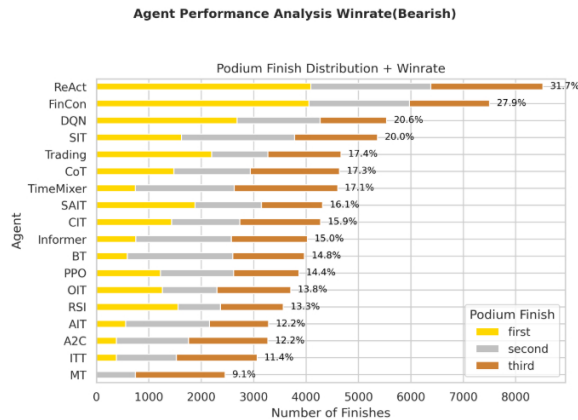


Figure 6: **Bear-period podium profile (January–October 2022).** Stacked segments show the reported first-, second-, and third-place finish frequencies; the legacy vertical-axis label “Agent” denotes a policy archetype.

policy i ’s rank- j count over M_c retained checkpoints in condition c , the displayed aggregate is $100 \sum_{j=1}^3 n_{i,c}^{(j)} / M_c$.

The July 7 local result also does not show a profitability advantage for the language policies. Table 13 reports negative mean terminal return for all four language policies, whereas Informer and TimeMixer are positive. Podium frequency is a separate regime-level diagnostic; it does not establish general language-agent superiority or real-world performance.

5.3 Innovation Keeps the Evaluator Contestable

Table 4 addresses RQ3 by comparing population-update components. Removing innovation produces a concentrated, low-switch condition, while removing perturbation yields an intermediate response. Appendix Figures 17 and 18 show the corresponding allocation paths and one-factor sensitivity. These contrasts make the separate roles of the update components visible: innovation supports contestability, and perturbation maintains responsiveness to changing conditions.

The magnitudes in Table 4 distinguish concentration from turnover in leadership. Without innovation, mean HHI rises from 0.189 to 0.477 and Top-1 share from 0.344 to 0.531, while the mean switch count falls from four to one. Without perturbation, the corresponding values are 0.245, 0.376, and three. Thus, in these configurations, a more concentrated allocation can also have fewer changes in its dominant set. A stable leader alone would miss the difference in competitive structure. The retained contrasts support reading concentration and reorganization together when assessing how the population operator changes the evaluation environment; they do not prescribe an optimal concentration level for every application.

Reading the signature. The four coordinates are reported separately. Local outcome records a policy’s trading result; concentration and reorganization characterize the shared evaluation market; and regime-conditioned standing records performance under a declared information regime. This separation preserves the market-mediated provenance of every local score. Supplementary policy-level outcomes and configuration checks appear in Appendix E and Appendix D.

Table 3: **Aggregate Top-3 finish rates (%) accompanying Figures 5 and 6.** Values transcribe the one-decimal totals retained in the rendered result. Each total aggregates first-, second-, and third-place finishes; per-place run-clustered intervals are not available in the retained artifact.

Policy	Bull	Bear	Policy	Bull	Bear
CoT	38.9	17.3	CIT	13.6	15.9
ReAct	27.9	31.7	TimeMixer	13.2	17.1
Informer	23.9	15.0	AIT	13.2	12.2
FinCon	21.0	27.9	SIT	12.6	20.0
SAIT	20.6	16.1	ITT	10.7	11.4
BT	19.7	14.8	DQN	9.7	20.6
Trading	18.6	17.4	MT	8.8	9.1
A2C	17.5	12.2	OIT	7.2	13.8
PPO	17.3	14.4	RSI	5.6	13.3

Table 4: Population-update ablations over 128 runs (mean $\pm$ 95% CI). Switches count changes in the dominant-archetype set.

Setting	HHI	Top-1	Switches
Full FINEVO	0.189 $\pm$ 0.012	0.344 $\pm$ 0.015	4 $\pm$ 0.663
No innovation	0.477 $\pm$ 0.014	0.531 $\pm$ 0.033	1 $\pm$ 0.483
No perturbation	0.245 $\pm$ 0.011	0.376 $\pm$ 0.017	3 $\pm$ 0.591

6 Discussion: Evaluation in Protocol

FINEVO makes the evaluation condition part of the reported object. The information schedule fixes when text and other declared signals become available, while order flow, allocation, and price formation arise inside the shared transition. The resulting question is therefore which financial-language interface is competitive under a named information–market condition, rather than whose action is scored against an unchanging tape.

The three result streams supply complementary evidence for that question. Controlled disturbances expose how a joint condition reorganizes the market; historical regimes reveal changes in competitive standing at a different time scale; and the update ablation shows that the evaluator’s population rule shapes the market in which every interface is assessed. These views make the provenance of an evaluation outcome inspectable: the outcome is attached to its information condition, market transition, and competitive population.

Interpreting disagreement between outcomes.

The July 7 returns and the regime podium profiles illustrate why the signature preserves separate coordinates. A terminal return describes the change in portfolio value across one episode; a podium frequency describes relative placement across ranking checkpoints under a historical schedule. A policy can therefore appear frequently near the top of a relative ranking without the reported evidence estab-

lishing positive terminal return in another episode. These summaries also cover different horizons and conditions, so they should be read at their stated scope. For evaluation design, the useful question is which kind of consistency is required: profitable local execution, frequent competitive placement, or persistence under changing market conditions. Each calls for its corresponding evidence.

Using the market record to diagnose a policy. The protocol identifies a sequence of records through which an evaluator can investigate an outcome: the text available at a decision time, the instruction derived from it, the exchange execution, the subsequent portfolio state, and the population allocation. This sequence separates an interpretation problem from an execution problem. For example, a favorable interpretation of news does not by itself determine the execution price or the inventory exposure created by an order. Competitors also affect those quantities through the shared book. A detailed diagnosis would therefore inspect the action and fill records alongside the price and allocation paths. The aggregate evidence presented here motivates that diagnostic use; attribution of a particular gain or loss requires the corresponding event records. This gives future evaluations a concrete connection between language decisions and the market outcomes they seek to explain.

7 Conclusion

Financial-language evaluation should ask not only what action a policy derives from text, but also what market is allowed to evaluate that action. FINEVO makes this object explicit by evaluating text-conditioned policies within a shared, feedback-driven market protocol rather than against an externally fixed trajectory.

The controlled, historical-regime, and ablation evidence shows why this distinction matters: local outcomes coexist with different patterns of concentration, competitive reorganization, and regime-conditioned standing, while population feedback shapes whether the evaluation remains contestable. These quantities describe complementary aspects of one evaluation trajectory.

FINEVO therefore offers a protocol-level lens for comparative financial NLP. It can extend to new text sources, policy interfaces, and market settings while keeping the central question intact: how should an executable financial-language policy be assessed when its actions help shape the trajectory on which it is scored? The evidence supports the utility of making that protocol explicit; it does not establish that the evaluated language policies outperform numerical baselines or that simulated outcomes predict deployable real-market returns.

8 Scope and Limitations

FINEVO compares executable policy archetypes inside a declared closed-loop market protocol. Its quantitative outcomes are conditional on the specified information schedule, continuous double auction, transaction cost, competing population, and population-update rule. The sensitivity and stress tests in Appendix D provide within-simulator sensitivity evidence; they do not establish invariance across alternative clearing mechanisms or calibrated real-market microstructure.

Historical timestamps condition the information stream, but prices, order flow, and allocations are generated endogenously rather than replayed from historical markets. The bull and bear labels therefore identify two historical input schedules for the simulated evaluation, not estimates of real-world or out-of-sample trading performance. The reported standings cover one July 7 intraday episode and two prespecified periods, so broader regime generalization requires replication across additional episodes.

The controlled intervention jointly changes text,

the fundamental anchor, and liquidity. It identifies the response of the full protocol to that joint condition, not the marginal effect of text. Factorial channel ablations and shuffled-text or rule-based-parser controls are needed to isolate text-specific effects. Finally, because policy families retain native observations and execution interfaces, cross-family rankings describe complete policy-in-protocol configurations rather than model quality under matched information and action interfaces.

9 Ethical Considerations

FINEVO is a research evaluator; any use must preserve privacy, market integrity, and human oversight. Its outputs are protocol-conditioned research evidence, not automated investment advice; any deployment must satisfy applicable data, operational, and market-conduct requirements.

References

- Selim Amrouni, Aymeric Moulin, Jared Vann, Svitlana Vyetenko, Tucker Balch, and Manuela Veloso. 2021. ABIDES-Gym: gym environments for multi-agent discrete event simulation and application to financial markets. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9.
- William A. Brock and Cars H. Hommes. 1998. [Heterogeneous beliefs and routes to chaos in a simple asset pricing model](#). *Journal of Economic Dynamics and Control*, 22(8–9):1235–1274.
- David Byrd, Maria Hybinette, and Tucker Hybinette Balch. 2020. ABIDES: Towards high-fidelity multi-agent market simulation. In *Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 11–22.
- Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. 2016. [Measuring the information content of financial news](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225, Osaka, Japan. The COLING 2016 Organizing Committee.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Carl Chiarella and Giulia Iori. 2002. [A simulation analysis of the microstructure of double auction markets](#). *Quantitative Finance*, 2(5):346–353.
- Carl Chiarella, Giulia Iori, and Josep Perelló. 2009. [The impact of heterogeneous trading rules on the limit order book and order flows](#). *Journal of Economic Dynamics and Control*, 33(3):525–537.
- Rama Cont, Sasha Stoikov, and Rishi Talreja. 2010. [A stochastic model for order book dynamics](#). *Operations Research*, 58(3):549–563.
- Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Ijcai*, volume 15, pages 2327–2333.
- J. Doyne Farmer and Shareen Joshi. 2002. [The price dynamics of common trading strategies](#). *Journal of Economic Behavior & Organization*, 49(2):149–171.
- Sascha Yves Frey, Kang Li, Peer Nagy, Silvia Saporita, Christopher Lu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. 2023. JAX-LOB: A gpu-accelerated limit order book simulator to unlock large scale reinforcement learning for trading. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 583–591.
- Dhananjay K. Gode and Shyam Sunder. 1993. [Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality](#). *Journal of Political Economy*, 101(1):119–137.
- Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2025. [FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6258–6292, Albuquerque, New Mexico. Association for Computational Linguistics.
- Cars H Hommes. 2006. Heterogeneous agent models in economics and finance. *Handbook of computational economics*, 2:1109–1186.
- Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 261–269.
- Allen H Huang, Hui Wang, and Yi Yang. 2023. [FinBERT: A large language model for extracting information from financial text](#). *Contemporary Accounting Research*, 40(2):806–841.
- Shimon Kogan, Dmitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. [Predicting risk from financial reports with regression](#). In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.
- Albert S Kyle. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pages 1315–1335.
- Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. 2025. [INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525, Vienna, Austria. Association for Computational Linguistics.
- Fabrizio Lillo, J. Doyne Farmer, and Rosario N. Mantegna. 2003. [Master curve for price-impact function](#). *Nature*, 421(6919):129–130.
- Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. [FinGPT: Democratizing internet-scale data for financial large language models](#). *arXiv preprint arXiv:2307.10485*.

- Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhao-ran Wang, and Jian Guo. 2022. [FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 1835–1849. Curran Associates, Inc.
- Thomas Lux and Michele Marchesi. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719):498–500.
- Tianmi Ma, Jiawei Du, Wenxin Huang, Wenjie Wang, Liang Xie, Xian Zhong, and Joey Tianyi Zhou. 2025. [Agent trading arena: A study on numerical understanding in LLM-based agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5496–5514, Suzhou, China. Association for Computational Linguistics.
- Chris Mascioli, Anri Gu, Yongzhao Wang, Mithun Chakraborty, and Michael Wellman. 2024. A financial market simulation environment for trading agents using deep reinforcement learning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 117–125.
- William H Sandholm. 2010. *Population games and evolutionary dynamics*. MIT press.
- Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. 2023. [TradeMaster: A holistic quantitative trading platform empowered by reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 59047–59061. Curran Associates, Inc.
- Paul C. Tetlock. 2007. [Giving content to investor sentiment: The role of media in the stock market](#). *The Journal of Finance*, 62(3):1139–1168.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kam-badur, David Rosenberg, and Gideon Mann. 2023. [BloombergGPT: A large language model for finance](#). *Preprint*, arXiv:2303.17564.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. [FinBen: A holistic financial benchmark for large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95716–95743. Curran Associates, Inc.
- Guojun Xiong, Zhiyang Deng, Keyi Wang, Yupeng Cao, Haohang Li, Yangyang Yu, Xueqing Peng, Mingquan Lin, Kaleb E. Smith, Xiao-Yang Liu, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. 2025. [FLAG-TRADER: Fusion LLM-agent with gradient-based reinforcement learning for financial trading](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13921–13934, Vienna, Austria. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Yangyang Yu, Haohang Li, Yupeng Cao, Keyi Wang, Zhiyang Deng, Zhiyuan Yao, Yuechen Jiang, Dong Li, Ruey-Ling Weng, and Jordan W. Suchow. 2025a. [Finnlp-fnp-llmfinlegal @ COLING 2025 shared task: Agent-based single cryptocurrency trading challenge](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 401–406, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W Suchow, Denghui Zhang, and Khaldoun Khashanah. 2025b. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *IEEE Transactions on Big Data*.
- Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. 2024. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045.
- Zhihan Zhou, Liqian Ma, and Han Liu. 2021. [Trade the event: Corporate events detection for news-based event-driven trading](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.
- Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022. [Astock: A new dataset and automated stock trading based on stock-specific news analyzing model](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 178–186, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Appendix

A Operational Details of the Evaluation Market

A.1 Population Allocation and Simplex-Preserving Updates

The population variable x^q is the population-allocation vector, operationalized as the fraction of evaluation capital assigned to each policy archetype at epoch q , rather than an estimate of the number of real market participants. Let

$$z_k^q = x_k^q \exp(\beta V_k^q + \gamma \sigma_k \epsilon_k^q), \quad \hat{x}_k^q = \frac{z_k^q}{\sum_j z_j^q}.$$

The update in Eq. 7 is $x^{q+1} = (1 - \mu)\hat{x}^q + \mu m^q$. Since both $\hat{x}^q$ and m^q are probability vectors, the update preserves non-negativity and unit mass for $0 \leq \mu \leq 1$. With $\mu > 0$ and a full-support m^q , every archetype receives a positive allocation at the next epoch. This is the operational property needed by FINEVO: it reallocates finite evaluation capital without creating or destroying mass.

Each epoch aggregates market outcomes over five minutes of simulated trading. Selection is driven by the forward-looking, archetype-level value signal V^q . Its target is conditioned on recent realized payoff and exogenous information available in the evaluation state; it is then used to form the next archetype-level capital allocation. Innovation mixes the configured distribution m^q , and stochastic perturbation enters through ϵ^q .

Table 5: Protocol configuration: documented defaults for the evaluation suite.

Item	Setting
Concurrent population	$N = 120$ agent instances
Session / update cadence	09:00–17:00 simulated session; five-minute epochs
Initial endowment	Equal capital across agent instances
Transaction cost	0.5% per trade, including slippage
Monte Carlo protocol	128 independent seeds per configuration
Selection / innovation	$\beta = 2.0$; $\mu = 0.05$; Dirichlet prior with $\alpha = 10$
Base perturbation scale	0.1

A.2 Market Engine and Execution

Orders meet in a continuous double auction with price–time priority. A marketable order consumes available depth at the best price and successive levels; a non-marketable residual remains queued

as a limit order. If an order executes at levels $\ell = 1, \dots, L$ with quantities v_ℓ and prices p_ℓ , its volume-weighted execution price is

$$\bar{p}_t = \frac{\sum_{\ell=1}^L v_\ell p_\ell}{\sum_{\ell=1}^L v_\ell}.$$

Cash and inventory are marked to the prevailing mid-price after execution. The recorded 0.5% per-trade friction includes the specified slippage setting. Consequently, a policy’s local PnL and the population allocation updated at the next epoch are both downstream of the same shared order book.

A.3 Agent Interfaces and Decision Policies

Table 6 records the operational comparison used in the evaluation suite. Families share the exchange and capital clock, but retain their native observations and decision interfaces.

The implementation uses GPT-4o-mini as the backbone for four text-conditioned policies. At a decision point, ReAct, CoT, and Trading receive processed market indicators x_t and recent scheduled news n_t ; FinCon also receives holdings h_t . A scheduled news event triggers an immediate indicator request and decision, while no-news intervals use randomized 5–15-minute wake-ups. ReAct maps a reason–act procedure to a fixed-quantity market buy/sell/hold order; CoT maps its reasoning trace to the same directional market-order interface; Trading emits an action–quantity plan; and FinCon emits a position-constrained action–quantity plan. Quantities are discretized in multiples of 10 units. Holding exchange mechanics fixed, the suite evaluates these text-conditioned policies against forecasting, reinforcement-learning, and rule-based policies.

A.4 Information Conditions and Alignment

The two information conditions share the same market engine but differ in the externally scheduled inputs. Table 2 in the main paper makes that division explicit. The input clock fixes when information becomes available; order flow, prices, and capital allocations are generated by the agents’ interaction.

A.5 Policy Legend for Supplemental Figures

The relational diagnostics use compact policy labels to keep the individual policy traces readable. Table 7 provides the exact mapping. It also makes clear that a family label is not a single policy: the

Table 6: Operational policy interfaces. The table describes the reported interfaces used in the shared market, rather than treating all agents as variants of one prediction model.

Family	Observation interface	Decision and execution interface
Technical	Recent prices and technical indicators	Momentum, RSI, or breakout buy/sell/hold rules; market orders
Informed	Fundamental-value anchor, inventory, and order-book state	Mispricing rules; conservative, aggressive, scaled, or spread-aware limit/market execution
Forecasting	Numerical indicator history and learned features	Informer, TimeMixer, or iTransformer directional signal; market order or hold
Reinforcement learning	State, reward, and action history	Double DQN, PPO, or A2C policy over buy/sell/hold actions
Language	Processed indicators and scheduled text; FinCon also receives position state	GPT-4o-mini ReAct, CoT, Trading, or FinCon action interface

Table 7: Individual-policy legend used in the supplementary diagnostics.

Family	Policy	Label
Technical	MomentumTrader	MT
Technical	RSITrader	RSI
Technical	BreakoutTrader	BT
Informed	ConservativeInformedTrader	CIT
Informed	AggressiveInformedTrader	AIT
Informed	SpreadAwareInformedTrader	SAIT
Informed	ScaledInformedTrader	SIT
Informed	OpportunisticInformedTrader	OIT
Forecasting	iTransformerTrader	ITT
Forecasting	InformerTrader	IT
Forecasting	TimeMixerTrader	TM
Reinforcement learning	DoubleDQNTrader	DQN
Reinforcement learning	PPOTrader	PPO
Reinforcement learning	A2CTrader	A2C
Language	LLMReActTrader	ReAct
Language	LLMCoTTrader	CoT
Language	LLMTradingTrader	Trading
Language	LLMFinConTrader	FinCon

evaluation compares several execution styles within the informed, forecasting, reinforcement-learning, and language-conditioned families.

A.6 Metrics and Statistical Aggregation

Terminal simple return is the reported local trading outcome. The primary market-structure metrics are population-allocation concentration (Eq. 5), Top-1 share, and the count of dominant-archetype switches. A switch occurs when the identity of the dominant-archetype set changes between consecutive epochs. Together these measures distinguish high-return behavior from persistent dominance, crowding, or rapid reorganization. For the regime comparison, the podium profile stacks first-, second-, and third-place finishes by policy. We use the resulting display as a within-regime standing diagnostic; it is distinct from local terminal return and allocation concentration. The action-entropy diagnostic used in the sensitivity plot is a separate summary of the distribution of discrete actions; it is not used as a synonym for population-allocation concentration.

How to read the supplemental diagnostics. The appendix separates three complementary layers of evidence. Terminal simple return summarizes what a policy earns in the reported episode. Allocation metrics such as HHI, Top-1 share, and dominant-archetype switches summarize how the evaluation market redistributes influence. Relational diagnostics—pairwise correlations, thresholded co-occurrence, and mutual-information views—show which individual policies are prominent together at a given stage. The three layers deliberately answer different questions: a local winner need not remain the dominant population member, and a concentrated population need not exhibit the same pairwise organization under two information conditions.

B Supplementary Controlled-Shock Evidence

B.1 Representative Pairwise Correlation Paths

The allocation paths in the main paper summarize which policy archetypes hold influence. Figure 7 adds a time-resolved view of selected individual-policy pairs under the two shock signs. It reveals the relationship layer of the market: policies can rise and fall together, separate, or re-align while the aggregate allocation statistics evolve. The curves are pairwise correlations, so they describe co-movement rather than a causal action by one policy on another.

This distinction matters for an evaluator of financial language agents. A leaderboard reports an endpoint; a competitive market also has a path and an organization. The pairwise traces expose whether the same apparent winner emerges through a stable coalition of policies or through a rapidly changing set of relationships. In conjunction with

HHI and dominant-archetype switches, they make allocation dynamics visible at a resolution that a terminal return table cannot provide.

B.2 Full Shock-Sign Correlation Snapshots

Figure 8 retains the complete correlation-matrix diagnostic behind the controlled-shock result. The upper and lower rows correspond to the positive- and negative-shock conditions. These matrices are descriptive summaries of policy co-movement; they complement the allocation trajectories and do not identify causal influence between policies.

B.3 Thresholded Co-occurrence Diagnostics

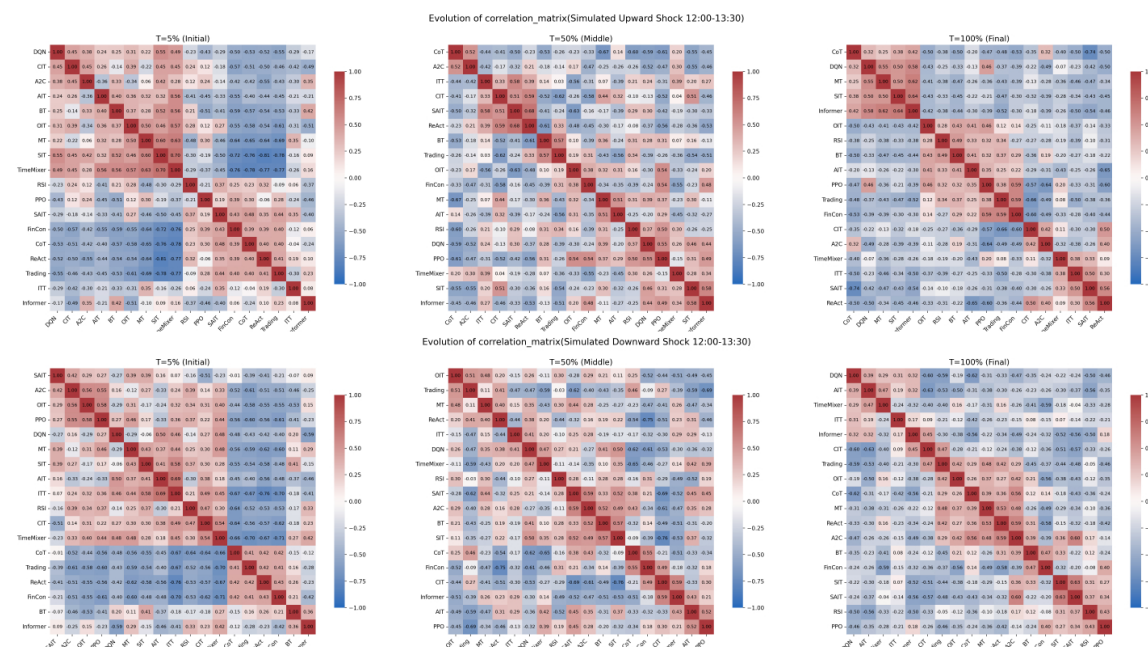
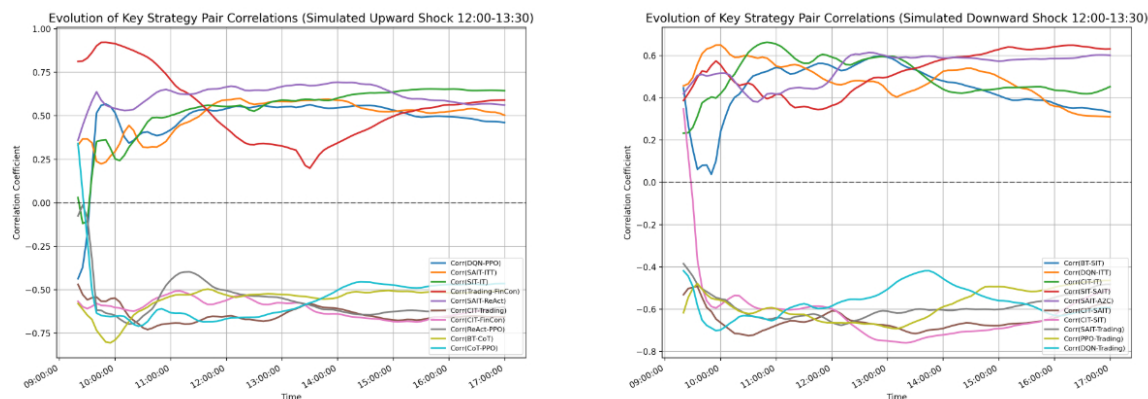
Figure 9 reports co-occurrence summaries for the positive and negative controlled-shock conditions. A cell records the frequency with which a pair of policies is simultaneously above the figure’s dominance threshold (12%) at the displayed stage. This diagnostic complements HHI and Top-1 by showing whether influence is concentrated in one policy archetype or distributed across pairs.

The comparison is deliberately sign-specific. The shock direction changes the information condition, while the market engine and evaluation clock remain fixed. The resulting matrices therefore provide a compact description of how the competitive field reorganizes under two controlled conditions. They should be read as a structural summary of the simulated population, alongside—not as a substitute for—the local trading metrics reported in the main paper.

Diagnostic reading protocol. Read the displays in sequence: HHI and Top-1 summarize concentration, switches summarize reorganization, and correlations, co-occurrence, and mutual information describe the population’s associative structure. Together they contextualize the shared market under a declared joint disturbance; none is a policy-specific causal attribution.

Table 8: How the controlled-shock diagnostics complement one another.

Diagnostic	Evaluation question
HHI / Top-1	How concentrated is the allocation of evaluation capital?
Dominant-archetype switches	How often does the leading allocation change?
Pairwise correlation	Which individual policy traces move together over time?
Co-occurrence	Which policies are simultaneously prominent at a specified threshold?
Mutual information	Which policy signals show statistical dependence over the episode?



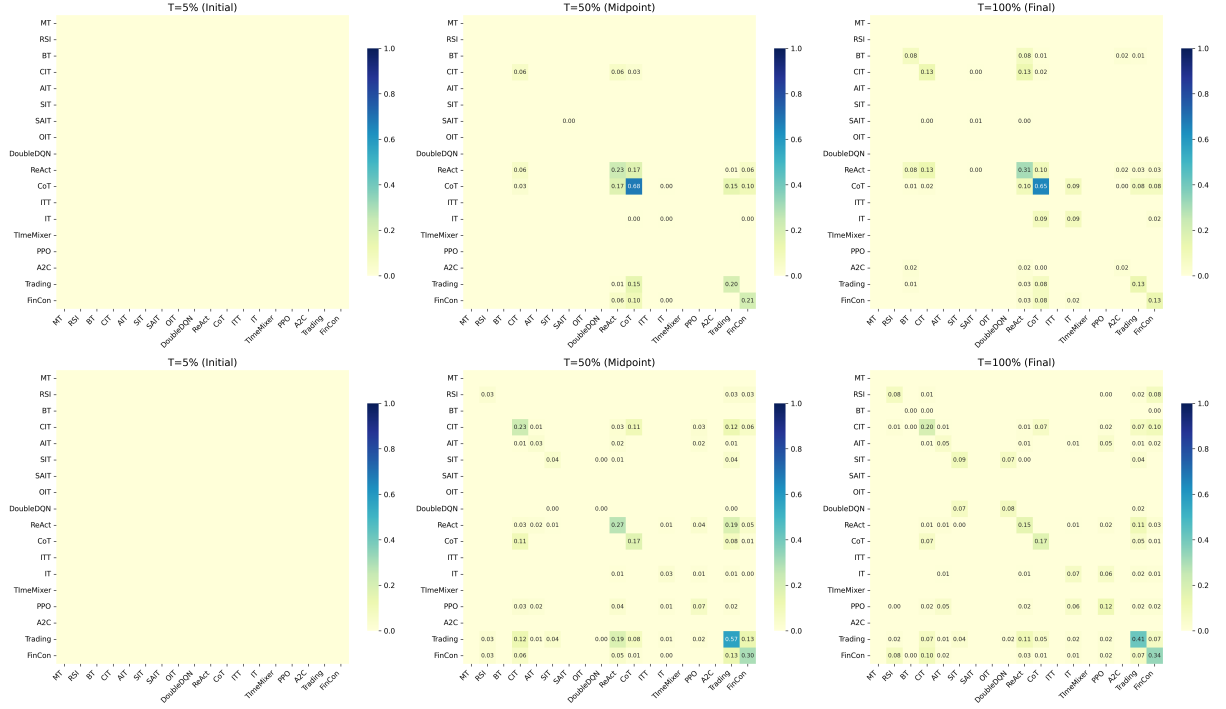


Figure 9: **Co-occurrence diagnostics across controlled shock signs.** The upper row shows positive-shock snapshots; the lower row shows the corresponding negative-shock snapshots. Entries quantify thresholded joint dominance among individual policies at the initial, midpoint, and final stages.

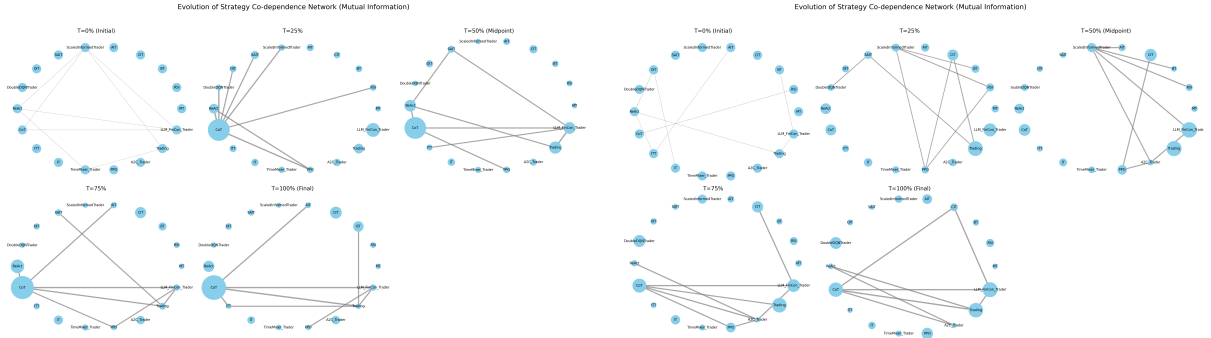


Figure 10: **Mutual-information diagnostics under controlled information shocks.** The left and right panels show the positive- and negative-shock conditions at successive checkpoints. Node size and edge thickness retain the prominence and statistical-dependence encodings of the source visualization. These networks are associative and do not establish causal influence between policies.

C Supplementary Historical-News Diagnostics

C.1 Episode Trajectories and Correlation Snapshots

The main paper presents population allocation and local cumulative-return paths from the July 7 episode as direct checks on the distinction between local outcomes and regime standing. This supplement retains the normalized volatility trajectory and three pairwise-correlation snapshots at a readable single-column scale. Together they provide a descriptive audit of the evolving market state, not a causal account of one policy’s effect on another.

C.2 Co-occurrence over the Historical-News Episode

The historical-news-conditioned evaluation complements the controlled shocks with a time-stamped information stream. The episode preserves the same closed-loop market: text is scheduled outside the exchange, whereas orders, prices, PnL, and allocations arise from the policies’ interaction. This separation lets the appendix ask a market-structure question that is distinct from an endpoint comparison: how does the joint prominence of individual policies evolve as the information clock advances?

Figure 15 gives the corresponding thresholded co-occurrence snapshots for the July 7 episode. The visual is useful because allocation concentration alone cannot distinguish a single dominant archetype from several policies repeatedly occupying the dominant set together. Reading it with the policy legend in Table 7 makes the individual-level organization behind the aggregate historical-news result directly inspectable.

C.3 Mutual-Information View of the Same Episode

Co-occurrence captures simultaneous prominence at a chosen threshold. Figure 16 adds a second associative view, visualizing the source mutual-information diagnostic at successive stages of the same episode. The two views are intentionally complementary: one describes which policies share a dominant state, and the other summarizes statistical dependence in their evolving signals. Neither view is a causal network, but together they reveal that a market evaluator exposes relationships that per-policy scores omit.

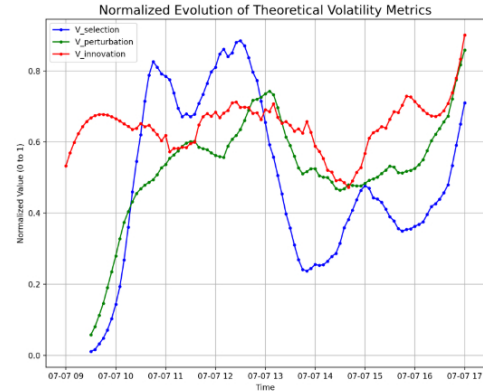


Figure 11: **Volatility components over the July 7 episode.** The three displayed components evolve under the shared information and trading clock.

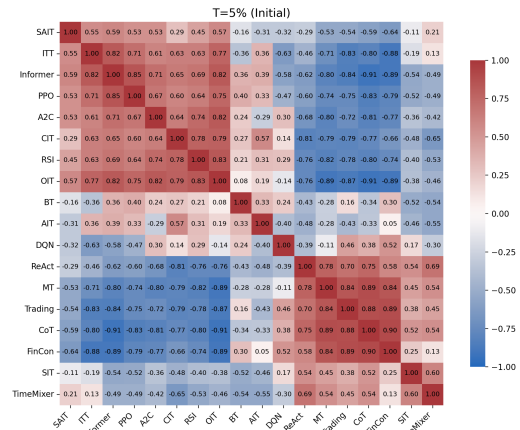


Figure 12: **Initial pairwise-correlation snapshot for the July 7 episode.** Colors encode signed correlations among individual policy traces.

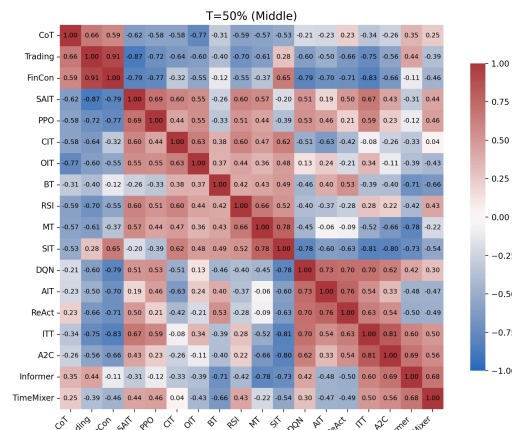


Figure 13: **Middle pairwise-correlation snapshot for the July 7 episode.** Table 7 maps the compact policy labels.

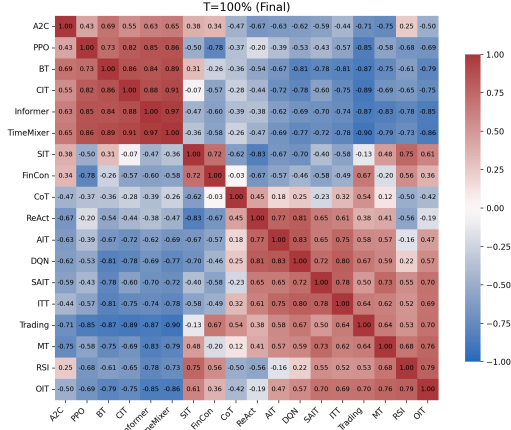


Figure 14: **Final pairwise-correlation snapshot for the July 7 episode.** The three snapshots are descriptive rather than causal.

The figure is therefore best read as a diagnostic of evaluation structure. A financial language policy is not only assessed against a fixed dataset; it joins a population whose relative influence and information relationships change as trading unfolds. That is precisely the closed-loop property FINEVO is designed to make measurable.

D Supplementary Mechanism and Robustness Results

D.1 Visible Update-Component Trajectories

The ablation table in the main paper summarizes selected population-update conditions. Figure 17 shows the full allocation trajectory together with the three individual component removals. Together, the figure and Table 4 connect visible market paths with the 128-run aggregate outcomes.

The comparison is useful because each update component plays a different market-design role. Selection translates recent value into reallocation; innovation maintains a route for low-share archetypes to re-enter; and perturbation makes the allocation process responsive to the configured stochastic component. Removing a component changes the trajectory itself, not merely a scalar score at the end of the session. The main-paper ablation therefore has a concrete dynamical counterpart in the paths shown here.

D.2 One-Factor Sensitivity

Figure 18 varies one update parameter at a time. It supplies a local sensitivity map for the selection, innovation, and perturbation controls rather than a new agent-comparison experiment. The panels

show which aspects of the evaluator respond when one control is moved while the rest of the protocol is held fixed. This makes the design knobs auditable: selection changes the strength with which recent outcomes reshape the population, innovation changes the renewal rate, and perturbation changes the scale of the stochastic allocation term.

D.3 Across-Seed Trajectories

The across-seed diagnostic fixes the market configuration and varies the seed across 128 runs. It records a simulated trade-price path and action entropy, allowing the uncertainty bands to be inspected directly rather than inferred from a single trajectory. The paired display is especially helpful for interpreting the main figures: price is an exchange-level outcome, whereas action entropy summarizes dispersion in the discrete decisions that generate order flow. Their joint evolution makes the simulation variability legible without conflating either quantity with population concentration.

D.4 Configuration Stress Tests

The following checks vary one aspect of the evaluation configuration while preserving the core closed-loop protocol. Together they probe whether the market signatures used in the paper remain interpretable when agent parameters, market size, noise structure, or asset identity changes.

Table 9 perturbs the documented agent parameters by $\pm 20\%$. Table 10 changes the concurrent population from 60 to 120 agents, testing whether the reported distributional summaries depend on a single market scale. The values provide a compact stability check for the evaluation environment rather than a separate agent-ranking claim.

The heavy-tailed setting in Table 11 increases tail exposure and introduces common-factor correlation, which tests the evaluator under a more demanding noise process. Finally, Table 12 records the same distributional summaries for four asset configurations. The four panels of evidence distinguish protocol-level behavior from a result that would depend on one narrowly chosen market setting.

E Supplementary Policy-Level Evidence

E.1 Text-to-Trade Contract

Financial language evaluation begins with an executable decision interface: time-stamped text must become an order that enters the same market as the

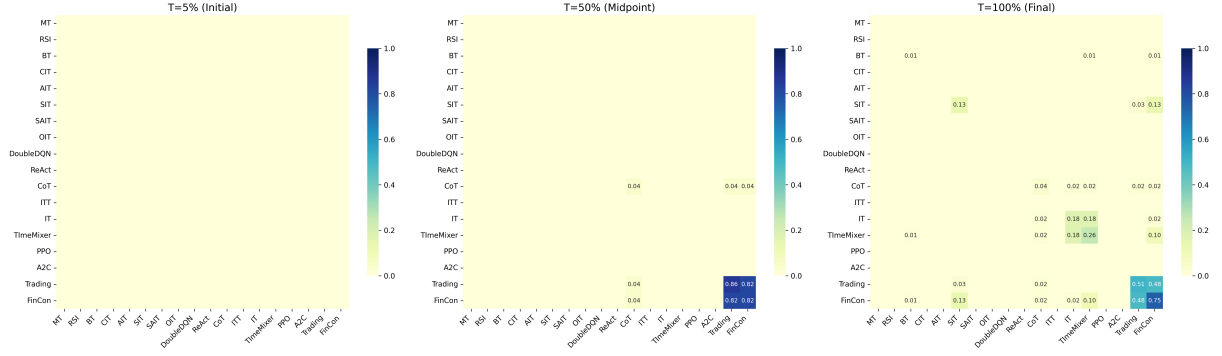


Figure 15: **Historical-news co-occurrence snapshots for the July 7 episode.** Entries show thresholded joint prominence among individual policies at the source figure’s initial, intermediate, and final checkpoints.

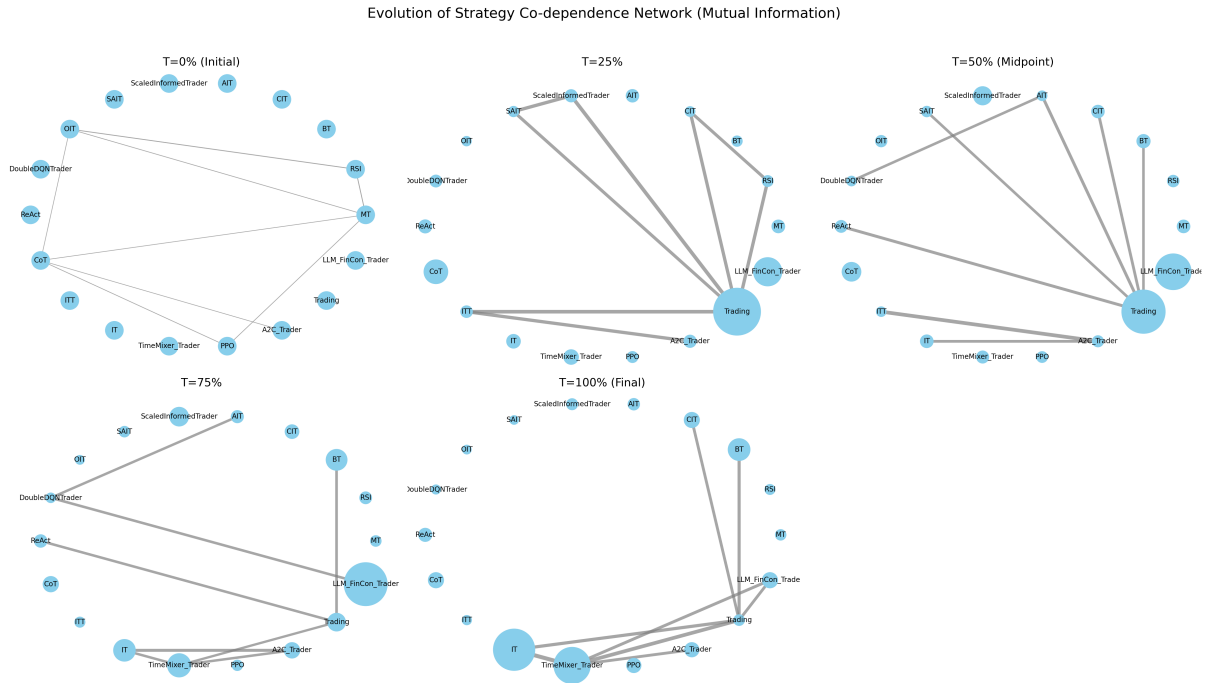


Figure 16: **Mutual-information diagnostic for the July 7 historical-news episode.** Node size and edge thickness retain the quantities encoded in the visualization at successive checkpoints. The display is associative rather than causal.

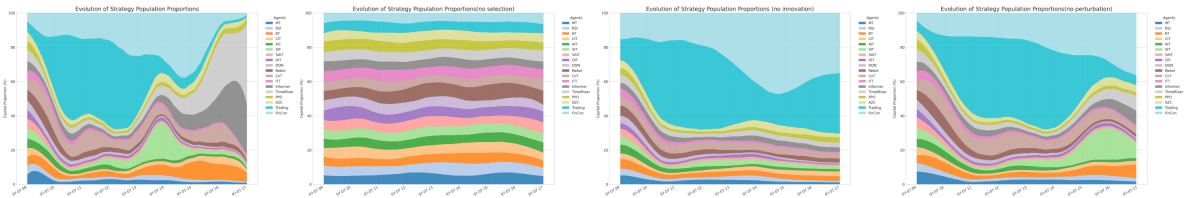


Figure 17: **Population-update ablation trajectories.** From left to right: full FINEVO, no selection, no innovation, and no perturbation. The main-paper ablation table reports the retained aggregate comparisons across 128 runs.

non-language policies. The text-conditioned policy has access to the scheduled message stream and market context at its decision time; its action is then cleared by the shared exchange, producing local outcomes and an updated competitive allocation.

The evaluation therefore measures a policy in the market it helps create, not only in a fixed record of labels or returns.

The four language-conditioned policies share clearing and equal-initial-capital conditions, but

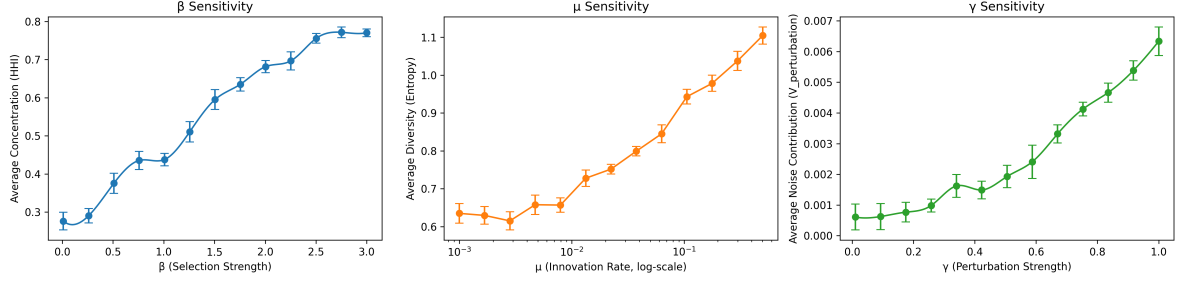


Figure 18: **One-factor sensitivity checks.** The three panels report HHI against selection strength β , action entropy against innovation rate μ , and the perturbation component against perturbation strength γ . Error bars represent 95% confidence intervals across 128 runs.

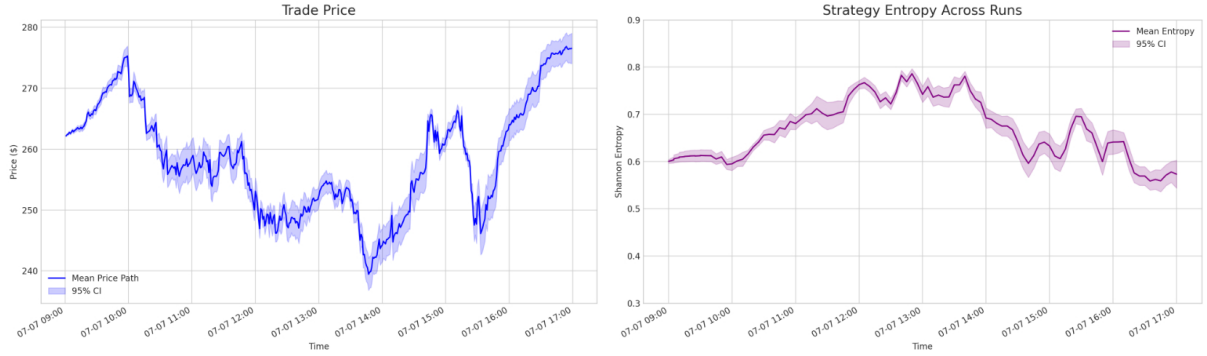


Figure 19: **Across-seed diagnostic for the July 7, 2025 episode.** The left panel gives the simulated trade-price trajectories and the right panel the corresponding action entropy. Curves are means and shaded bands are 95% confidence intervals over 128 runs.

Table 9: $\pm 20\%$ agent-parameter stress test (128 runs; mean $\pm$ 95% CI).

Regime	Setting	Excess kurtosis	Skewness	Vol-of-Vol
Bull	Baseline	3.163 ± 0.138	-1.532 ± 0.129	0.008 ± 0.0004
Bull	Parameters perturbed	3.141 ± 0.152	-1.596 ± 0.138	0.008 ± 0.0004
Bear	Baseline	4.259 ± 0.172	-1.793 ± 0.149	0.013 ± 0.0007
Bear	Parameters perturbed	4.196 ± 0.181	-1.692 ± 0.157	0.012 ± 0.0008

Table 10: Market-size diagnostic (128 runs per configuration; mean $\pm$ 95% CI).

Regime	N	Excess kurtosis	Skewness	Vol-of-Vol
Bull	120	3.143 ± 0.141	-1.539 ± 0.132	0.009 ± 0.0004
Bull	100	3.163 ± 0.138	-1.532 ± 0.129	0.008 ± 0.0004
Bull	80	3.178 ± 0.165	-1.541 ± 0.158	0.008 ± 0.0005
Bull	60	3.159 ± 0.178	-1.516 ± 0.162	0.007 ± 0.0005
Bear	120	4.326 ± 0.169	-1.805 ± 0.152	0.012 ± 0.0006
Bear	100	4.259 ± 0.172	-1.793 ± 0.149	0.013 ± 0.0007
Bear	80	4.115 ± 0.175	-1.765 ± 0.161	0.013 ± 0.0007
Bear	60	4.251 ± 0.197	-1.772 ± 0.183	0.011 ± 0.0007

Table 11: Correlated heavy-tailed shock stress test (128 runs; mean $\pm$ 95% CI). The alternate configuration uses a common-factor correlation of $\rho = 0.5$ and standardized Student- t innovations with $\nu = 5$.

Regime	Setting	Excess kurtosis	Skewness	Vol-of-Vol
Bull	Baseline	3.163 ± 0.138	-1.532 ± 0.129	0.008 ± 0.0004
Bull	Correlated heavy-tailed	5.278 ± 0.151	-1.632 ± 0.133	0.011 ± 0.0007
Bear	Baseline	4.259 ± 0.172	-1.793 ± 0.149	0.013 ± 0.0007
Bear	Correlated heavy-tailed	6.399 ± 0.214	-1.932 ± 0.175	0.015 ± 0.0009

differ in how they turn their documented inputs into an executable action. ReAct, CoT, and Trading

Table 12: Cross-asset diagnostic (128 runs per asset; mean $\pm$ 95% CI).

Asset	Excess kurtosis	Skewness	Vol-of-Vol
AAPL	4.354 ± 0.410	-1.656 ± 0.180	0.011 ± 0.0015
TSLA	5.578 ± 0.530	-2.344 ± 0.240	0.015 ± 0.0023
MSFT	3.176 ± 0.350	-0.967 ± 0.150	0.008 ± 0.0016
XOM	3.741 ± 0.380	-0.573 ± 0.120	0.008 ± 0.0011

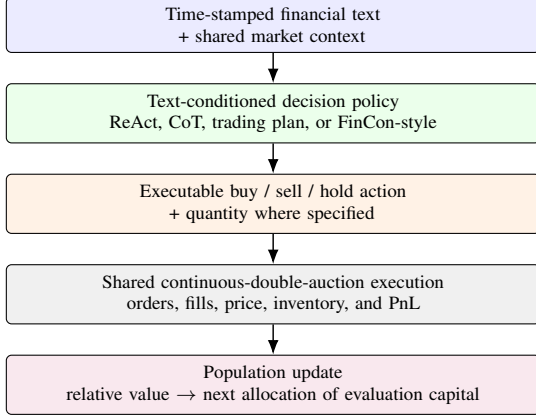


Figure 20: **Text-to-trade contract in FINEVO.** Text-conditioned policies enter the same execution and population-update loop as every other policy family.

receive processed indicators x_t and recent scheduled news n_t ; FinCon additionally receives holdings h_t . The event schedule fixes when textual information enters this interface. ReAct and CoT select a directional market order, Trading emits an action–quantity plan, and FinCon applies a position constraint before issuing its action–quantity plan. The comparison therefore targets a text-to-trade interface rather than a different exchange for each language policy.

E.2 Per-Policy Outcome Profile

Table 13 reports terminal simple return for the July 7 historical-news-conditioned episode. For policy k in run r , the metric is $\text{Ret}_{k,r} = PV_{k,r,T}/PV_{k,r,0} - 1$. Values are means with 95% confidence intervals across the 128 retained run-level returns; for example, 6.738 denotes a 673.8% terminal gain. This auditable episode metric complements, rather than replaces, the allocation and podium summaries.

Across this episode, all four language policies have negative mean terminal return, whereas Informer and TimeMixer are positive. The result supports the paper’s separation of local outcomes from market structure; it does not support a claim that language-conditioned policies are the most profitable family.

F Submission and Reproducibility Checklist

The reproducibility unit for a closed-loop evaluator is the protocol, not just the final score table. A release should preserve the configuration that determines what each policy observes, how it acts, how orders clear, and how outcomes feed back into population allocation. The following manifest specifies the artifacts needed to reproduce the evidence presented in this submission.

Text sources and model outputs should be released in accordance with their licences and privacy conditions. With these artifacts, future work can test whether a new financial language agent changes the same closed-loop properties: local outcomes, allocation concentration, reorganization, and information- conditioned standing.

F.1 Replay Integrity Checks

A complete release record supports a concrete audit of the evidence chain. A reviewer should be able to begin from the scheduled information inputs, recreate policy actions and market executions for a declared seed, and then recompute the allocation summaries and confidence intervals. This is more specific than verifying a final leaderboard: the same protocol must reproduce the sequence through which a policy’s local behavior becomes a market-level evaluation result.

These checks turn the appendix from a list of implementation details into a reviewable contract between a financial-language policy, the market protocol, and the reported result. They also make it straightforward to add a new model: hold the market and declared information condition fixed, record its text-to-trade interface, and compare its local and population-level outcomes under the same evaluation trace.

Table 13: **Terminal simple return by policy for the July 7, 2025 historical-news-conditioned episode.** Values are mean $\pm$ 95% CI across 128 run-level terminal returns; DQN labels DoubleDQN. Sharpe, drawdown, and turnover summaries from the reviewed version are omitted because their run-level uncertainty and sampling or annualization conventions are not retained in the available aggregate result.

Family	Policy	Terminal return	Family	Policy	Terminal return
Technical	MT	-0.832 ± 0.064	Language	ReAct	-0.685 ± 0.025
Technical	RSI	-0.849 ± 0.023	Language	CoT	-0.560 ± 0.046
Technical	BT	0.034 ± 0.010	Language	Trading	-0.746 ± 0.048
Informed	CIT	-0.544 ± 0.032	Language	FinCon	-0.685 ± 0.045
Informed	AIT	-0.807 ± 0.026	Forecasting	ITT	-0.896 ± 0.066
Informed	SIT	-0.673 ± 0.057	Forecasting	Informer	4.333 ± 0.022
Informed	SAIT	-0.866 ± 0.010	Forecasting	TimeMixer	6.738 ± 0.031
Informed	OIT	-0.915 ± 0.032	RL	PPO	-0.431 ± 0.033
RL	A2C	-0.733 ± 0.004	RL	DQN	-0.885 ± 0.034

Table 14: Documented text-to-trade interfaces for the four language-conditioned policies.

Policy	Inputs and decision construction	Executable output
ReAct	$x_t + n_t$; reason-act procedure selects BUY / SELL / HOLD	Fixed-quantity market order
CoT	$x_t + n_t$; reasoning trace produces a final action	Fixed-quantity market order
Trading	$x_t + n_t$; trading plan produces {action, quantity}	Market order specified by plan
FinCon	$x_t + n_t + h_t$; constrained plan respects position limit	Market order only when constraint is met

Table 15: Reproducibility manifest for a FINEVO evaluation release.

Artifact	Role in reproducing the evaluation
Market configuration	Session clock, endowments, transaction costs, order-book rules, and population-update parameters
Policy specifications	Policy observations, decision interfaces, execution constraints, and versioned language-agent prompts and inference settings
Information alignment	Time-stamped message/disclosure schedule and the procedure that makes inputs available to policies
Randomness record	Seed list, sampling configuration, and aggregation code for confidence intervals and podium-profile summaries
Evaluation outputs	Orders, transactions, allocations, trajectories, and the data used to create every figure and table

Table 16: Integrity checks for a closed-loop evaluation replay.

Check	Evidence to compare
Information timing	Recorded schedules show when each policy’s documented inputs become available for a condition.
Decision replay	Policy inputs, inference settings, and constraints reproduce the recorded action trace.
Market replay	Orders and book rules reproduce fills, prices, cash, and inventory transitions.
Allocation replay	Recorded values reproduce the next population allocation under the documented update.
Aggregate replay	The seed set reproduces reported intervals, concentration, switches, and podium-profile summaries.