How Well Do LLM-Based Simulations Predict? Lessons from Structural Models

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in large language models (LLMs) have driven rapid progress in social simulation, enabling heterogeneous agents to sustain complex interactions and reproduce a range of emergent phenomena. Yet it remains unclear whether such simulations can reliably predict how real populations respond to interventions. We examine the real-world predictive validity of LLM-based social simulations by comparing with structural models, which derive behavioral mechanisms from economic theory and estimate key parameters from observed data. Across the evaluated settings, LLM simulators consume substantial inference resources yet consistently underperform structural models in directional prediction. Scaling the simulated population does not close this gap. Analyzing agent response distributions, we find that larger deviations from structural response distributions are consistently associated with lower directional accuracy. Motivated by this finding, we propose Structure-Guided Behavioral Alignment (SGBA), which uses structural moments estimated from observable market data as behavioral supervision to calibrate distributions over executable agent policies while keeping LLM parameters fixed. SGBA improves directional accuracy by an average of 26.6 percentage points across the two evaluation settings and moves simulated response distributions closer to their structural references. Ablations show successive gains from explicit policy representation and structural distribution optimization. These findings highlight behavioral distribution alignment as promising principles for improving the empirical reliability of LLM-based social simulation. Code is available at `https://anonymous.4open.science/r/SGBA-2027`.

1 Introduction

LLM-based social simulation has rapidly developed from small interactive demonstrations into a general framework for studying human and collective behavior. Stanford's *Generative Agents* populated the well-known Smallville sandbox showed that LLM agents can sustain memory, planning, and social interaction over time (Park et al., 2023). *Turing Experiments* and synthetic human samples explored their ability to reproduce heterogeneous responses observed in human-subject settings (Aher et al., 2023; Argyle et al., 2023). More recent systems have substantially expanded the scale and scope of this paradigm, with *OASIS* supporting social interaction simulations with up to one million agents (Yang et al., 2024), *AgentSociety* enabling large-scale simulations of social interaction and interventions (Piao et al., 2025), and *SocioVerse* studying population dynamics using a large pool of real-world user profiles (Zhang et al., 2025). Beyond these advances, however, concerns about empirical grounding and validation raise doubts about the broader value of LLM-based social simulation (Larooij & Törnberg, 2026; Hu et al., 2026). **Do LLM-based simulations actually improve our ability to understand and predict the world?**

For decades, structural models have been widely used in economics and policy analysis to evaluate interventions and conduct counterfactual prediction (Low & Meghir, 2017; Blundell, 2017). As illustrated in Figure 1, they specify behavioral and equilibrium relationships from economic theory, estimate or calibrate unknown parameters using observed data, and use the fitted model to predict responses under interventions. With LLM-based social simulation advancing rapidly, we consider **how do LLM-based social simulation compare with structural models in predicting real responses to interventions?**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

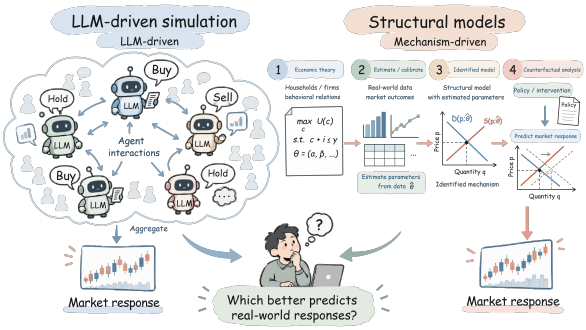


Figure 1: Motivation of our study.

To answer this question, we compare two structural models with three financial market simulators, each evaluated with three LLM backbones. The evaluation covers four Index Constituent Adjustment events in China, predicting returns of affected stocks, and four Fed Rate Change events in the United States, predicting short-window asset returns around policy announcements. Surprisingly, despite consuming up to 20.293 million inference tokens per event, the LLM simulations underperform the structural models across all configurations, with mean directional accuracy lower by 31.5 percentage points in Index Constituent Adjustment and 58.1 points in Fed Rate Change. We take a closer look at this gap and find that larger deviations of simulated behavioral response distributions from their structural references are consistently associated with lower directional accuracy. Increasing the simulated population by up to $8\times$ does not systematically close the predictive gap.

Inspired by our findings, we propose **Structure-Guided Behavioral Alignment (SGBA)**, which introduces distributional supervision over agent behavior. SGBA first uses a frozen LLM to generate explicit, executable policies for each agent, conditioned on its persona, initial state and factual constraints. It then calibrates the population distribution over these policies so that their induced responses align with structural moments. The aligned policies are executed through the simulator. Across the two evaluation settings, SGBA improves directional accuracy by an average of 26.6 percentage points and moves simulated response distributions closer to structural references. Ablations show successive gains from explicit policy representation and structural distribution optimization. SGBA further transfers across structurally distinct intervention settings.

Our work shows that population-level behavioral alignment provides a practical way to ground LLM-based social simulation in domain mechanisms and empirical evidence. We make contributions on: **(1) We evaluate the real-world predictive validity of LLM-based social simulation.** Despite rich agent behavior and emergent phenomena, current simulators exhibit a substantial gap in predicting realized collective responses. **(2) We identify behavioral distribution mismatch as a key source of simulation error.** Larger deviations from structural response distributions are associated with lower predictive accuracy, while population scaling alone cannot eliminate systematic response bias. **(3) We introduce SGBA to align agent response distributions.** SGBA improves prediction and behavioral alignment across simulators and LLM backbones. We hope this work advances social simulation toward more systematic integration of domain mechanisms and real-world data alongside advances in agent capability and simulation scale, helping develop LLM-based simulations into reliable and empirically testable tools for studying social processes and intervention effects.

2 Related Work

LLM-Based Social Simulation. Recent research has increasingly used LLM agents to study collective behavior in interactive environments. S^3 (Gao et al., 2023) models emotion, attitudes, and interpersonal interactions in social networks, illustrating how individual behaviors can generate population dynamics. EconAgent (Li et al., 2024) embeds heterogeneous LLM agents in a macroeconomic environment, where work and consumption decisions jointly produce aggregate economic patterns. AgentSociety (Piao et al., 2025) further supports large-scale simulations of social interaction and interventions across diverse societal settings. As these systems become more expressive, empirical fidelity has received growing attention. SimBench (Hu et al., 2026) systematically evaluates how closely LLM-generated responses reproduce observed human behavior across heterogeneous tasks and populations. This line of work reflects a broader shift from demonstrating simulation capability toward assessing correspondence with real populations.

Structural Models. Structural models represent economic behavior through explicit primitives such as preferences, beliefs, constraints, and equilibrium conditions, which together determine how

agents respond to changes in their environment. Unknown parameters are disciplined by observed data, often by matching model-implied quantities to empirical moments through the generalized method of moments (Hansen, 1982) or simulated moments (McFadden, 1989). Once estimated, the same behavioral mechanisms can be used to evaluate responses to interventions and counterfactual states by changing the relevant policy or state inputs. Structural models are widely used for policy evaluation, counterfactual analysis, and market response prediction.

Behavioral Alignment. Behavioral alignment has been studied at both the model and population levels. Reinforcement learning from human feedback (Ouyang et al., 2022) adapts model parameters using human preference comparisons, while population simulation increasingly emphasizes matching heterogeneity across simulated individuals. Mixture of Personas (Bui et al., 2025) learns context-dependent weights over persona–exemplar components to reproduce population responses. Distributional Alignment for Social Simulation (Xie et al., 2026) calibrates mixtures of prompts toward observed response distributions. Related statistical methods address distribution matching through aggregate constraints. Entropy balancing (Hainmueller, 2012) reweights samples to satisfy empirical moments, while posterior regularization (Ganchev et al., 2010) constrains latent distributions through expectations of selected features. These approaches move behavioral alignment beyond individual outputs toward controlling the composition and aggregate properties of population behavior.

3 How Do LLM-Based Simulations Compare with Structural Models?

3.1 Experimental Setup

To systematically evaluate the predictive performance of LLM-based simulations, we compare three simulators against two structural models across eight historical events in financial markets. Figure 2 summarizes the evaluation framework.

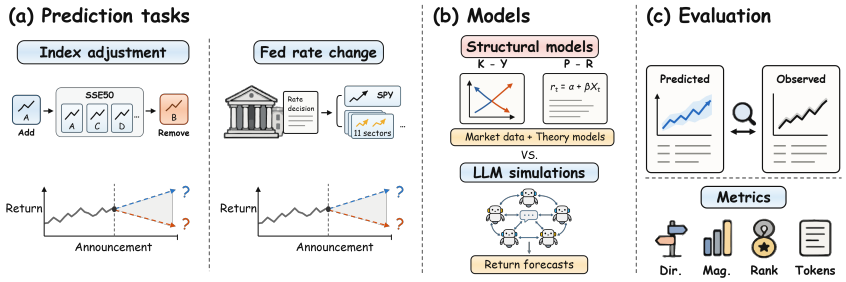


Figure 2: Evaluation framework.

Prediction Tasks. We evaluate how well LLM-based simulations and structural models predict responses to economic announcements in two financial market scenarios. *Index Constituent Adjustment* involves adding stocks to or removing stocks from the Shanghai Stock Exchange 50 Index. Models are asked to predict the returns of affected stocks from the announcement date to the effective date. *Fed Rate Change* involves announcements by the U.S. Federal Reserve about its policy interest rate. Models are asked to predict asset returns from immediately before the announcement to twenty minutes afterward. We select four historical events per scenario for a total of eight events, as detailed in Appendix A.

Structural Models. We use one structural model for each scenario to capture the economic mechanism underlying the announcement. Detailed descriptions and implementation are in Appendix B.1.

- Koijen–Yogo Demand System (K–Y) (Koijen & Yogo, 2019). This model describes how different investors allocate their portfolios and determines asset prices by equating aggregate demand with supply. For *Index Constituent Adjustment*, changes in index membership

shift investor demand for the affected stocks. The model predicts the resulting price adjustments and corresponding stock returns.

- Pflueger–Rinaldi Monetary Model (P–R) (Pflueger & Rinaldi, 2022). This model describes how monetary policy surprises affect economic activity and investors' risk aversion. These effects change expected cash flows and the returns investors require to hold risky assets, jointly determining stock valuations. For *Fed Rate Change*, we use this mechanism to predict asset returns following the announced policy decision.

LLM-Based Financial Market Simulators. We evaluate three LLM-based financial market simulators that differ in how agents make decisions and how trades determine market prices. 1) **StockAgent** (Zhang et al., 2024) models heterogeneous LLM investors who exchange information through a shared bulletin board and submit orders to a sequential matching market. 2) **TwinMarket** (Yang et al., 2025) generates trading decisions through updates to investors' beliefs, desires, and intentions alongside social interactions. Prices are determined through call auctions. 3) **MarketSim** (Piao et al., 2026) uses LLM managers to make portfolio decisions and rule-based traders to execute them through continuous double auctions. More implementation details are provided in Appendix B.2.

Evaluation. Within each scenario, all methods share the information cutoff, announced intervention, evaluated assets, and prediction horizon. Historical inputs and structural estimation use only information available by the cutoff, while subsequent realized returns are reserved for evaluation. Detailed input controls and information-cutoff rules are provided in Appendix B.3. Each configuration is evaluated with three random seeds per event. Our primary metrics are *Direction* and *Directional Value (DV)*, which assess directional correctness and account for realized return magnitudes, respectively. We also report MAE and RMSE for magnitude errors, Spearman correlation for cross-asset ranking, and LLM token usage for resource consumption. Complete metric definitions and reporting conventions are provided in Appendix A.2.

3.2 Comparative Results

As shown in Table 1, we compare the structural models with three LLM simulators. Each simulator is evaluated with DeepSeek-V4-Flash, GLM-5.3-Flash, and GPT-4o. We report the average predictive performance and LLM token usage across the four events in each scenario.

Table 1: Predictive performance in the two market scenarios. † Reasoning enabled.

Scenario	Model	Backbone	Directional (%)		Error		Spearman ↑	Tokens (M) ↓
			Direction ↑	DV ↑	MAE ↓	RMSE ↓		
Panel A Index Constituent Adjustment	K–Y	–	**73.8**	**45.6**	8.43	13.39	**0.399**	**0.0**
	MarketSim	DeepSeek-V4-Flash†	39.4	−12.6	14.33	18.38	−0.274	92.1
		GLM-5.3-Flash	46.9	−12.5	13.86	18.08	−0.240	67.2
		GPT-4o	50.6	4.9	15.02	19.98	−0.179	39.2
	StockAgent	DeepSeek-V4-Flash†	35.6	−34.3	5.25	6.39	−0.207	203.7
		GLM-5.3-Flash	40.6	−28.0	5.35	6.50	−0.414	186.7
		GPT-4o	17.5	−41.4	**5.13**	**6.26**	0.087	164.8
	TwinMarket	DeepSeek-V4-Flash†	52.5	13.3	13.90	16.28	−0.309	243.5
		GLM-5.3-Flash	53.1	−24.3	17.51	19.33	−0.249	180.4
		GPT-4o	44.4	1.4	7.13	8.67	−0.027	174.6
Panel B Fed Rate Change	P–R	–	**83.3**	**66.3**	**19.47**	**26.39**	**0.395**	**0.0**
	MarketSim	DeepSeek-V4-Flash†	29.2	−39.3	44.45	56.19	−0.236	0.8
		GLM-5.3-Flash	33.3	−33.7	44.59	56.63	−0.252	0.5
		GPT-4o	26.9	−48.0	44.58	56.64	−0.274	0.4
	StockAgent	DeepSeek-V4-Flash†	28.6	−31.5	53.48	65.47	0.137	14.3
		GLM-5.3-Flash	10.4	−74.4	55.31	79.55	0.349	8.2
		GPT-4o	29.9	−37.7	42.17	52.20	0.253	7.1
	TwinMarket	DeepSeek-V4-Flash†	27.4	−42.4	41.18	50.54	0.222	18.5
		GLM-5.3-Flash	19.9	−58.1	43.72	55.37	0.337	9.5
		GPT-4o	20.8	−50.2	40.31	51.67	0.338	8.5

Structural Models vs. LLM Simulators. Structural models outperform all nine LLM configurations on both Direction and DV in each scenario. K–Y achieves 73.8% Direction in Index Constituent Adjustment and P–R achieves 83.3% in Fed Rate Change, exceeding the best LLM results

by 20.7 and 50.0 percentage points, respectively. Their corresponding DV advantages are 32.3 and 97.8 percentage points. The advantage is less uniform on the complementary metrics: some LLM configurations achieve lower magnitude errors in the index scenario. Directional prediction thus provides the clearest distinction between the two modeling approaches. Detailed event-level analysis is provided in Appendix C.1.

Cost Analysis. Across both scenarios, the LLM simulations consume 1.42 billion tokens in total. The structural models achieve stronger directional performance without requiring LLM calls. In addition, higher token usage does not consistently correspond to better directional performance across the evaluated configurations. Figure 10 in Appendix C.2 visualizes the relationship between token usage and Direction.

Finding 1 Our financial market experiments highlight ***structural models as effective and efficient alternatives to LLM-based social simulation***, with better performance and no LLM token usage.

4 Why Do LLM-Based Simulations Fail?

4.1 Diagnosing the Gap Between Structural Models and LLM Simulations

To investigate the predictive gap observed in Section 3.2, we compare simulated and structural response distributions and examine whether their mismatch is associated with directional prediction accuracy.

Comparing Response Distributions. We compare aggregate response distributions from the LLM simulators with those implied by the structural models. Figure 3 shows systematic differences between the simulated and structural response distributions in both settings. Detailed response construction and structural references are provided in Appendix D.1.

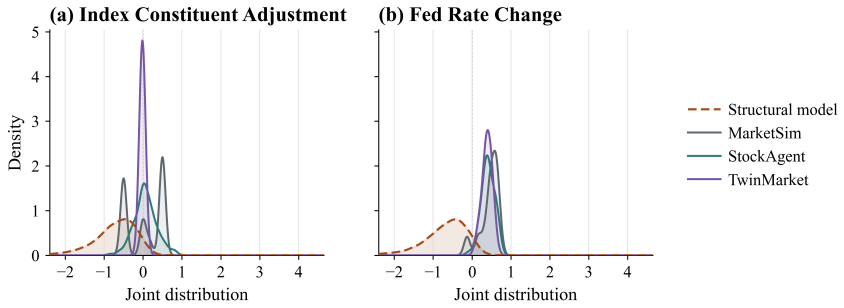


Figure 3: Response Distributions across models.

Relationship Between Response Deviations and Predictive Performance. We quantify structural mismatch at the run level. For each run, structural response deviation measures how far its simulated aggregate response departs from the matched structural reference. Appendix D.2 provides the detailed definitions. Figure 4 groups runs by setting-specific structural deviation. In both settings, directional accuracy decreases as deviation increases, linking greater distribution mismatch to poorer predictive performance.

Finding 2 LLM simulators exhibit systematic structural response mismatch, with larger deviations associated with lower directional accuracy. ***Response distribution mismatch is an informative diagnostic of predictive error.***

4.2 Understanding the Limits of Scaling LLM Simulations

We vary the number of agents in each of the three LLM-based simulators and compare their predictive performance with the corresponding structural benchmark. As Figure 5 a shows, scaling yields

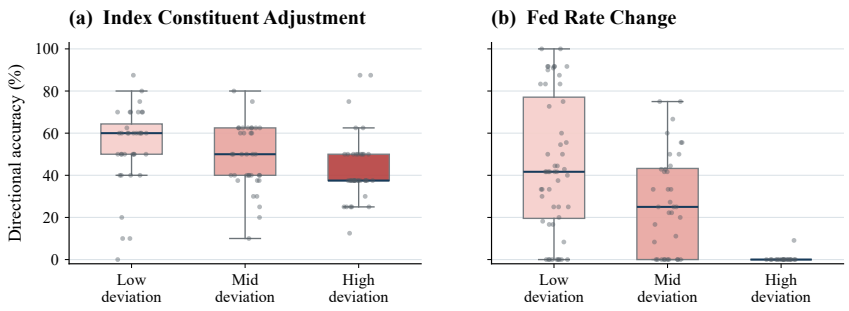


Figure 4: Directional accuracy by structural response deviation.

limited and non-monotonic gains. Relative to $1\times$, the best-scale Direction improves by 10–20 percentage points. MarketSim peaks at $2\times$, whereas StockAgent and TwinMarket perform best at $8\times$. Yet all three remain below the structural benchmark of 73.8%. Figure 5 b–d shows a similar pattern at the distributional level. Increasing the population changes the simulated response distributions, but does not induce systematic convergence toward the structural distribution. Across scales, the simulated distributions remain more concentrated and retain modes that differ from the structural target.

Finding 3 Increasing agent populations by up to 8× does not close the prediction gap or align response distributions systematically. ***Population scaling cannot eliminate systematic response bias.***

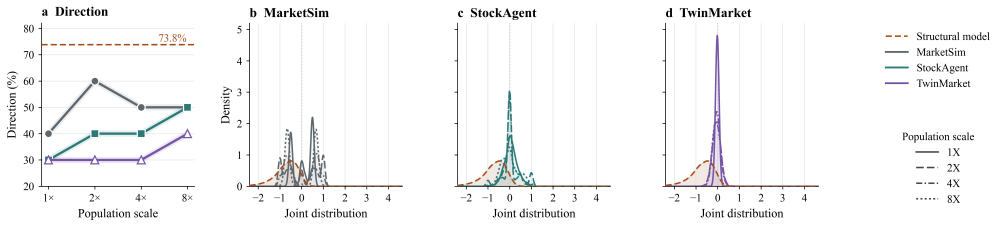


Figure 5: Prediction and response distribution under population scaling.

We formalize why population scaling cannot eliminate systematic response error using a bias–variance decomposition of the simulated population response. For agents with characteristics z, let $P^\star(\pi \mid z)$ be the target population distribution over policies π, and $Q^0(\pi \mid z)$ the corresponding distribution in the original simulator. Each policy $\pi(a \mid z, S)$ maps agent characteristics z and market state S to a distribution over actions a. For event e, the response feature $\phi_e(\pi)$ extracts from policy π the behavior associated with one of the structural moments in Section 4.1. Averaging this feature over the target and simulated policy distributions gives the corresponding population responses

$$\mu_e^\star = \mathbb{E}_{\pi\sim P^\star(\cdot|z)}\left[\phi_e(\pi)\right], \qquad \mu_e^0 = \mathbb{E}_{\pi\sim Q^0(\cdot|z)}\left[\phi_e(\pi)\right], \tag{1}$$

Here, $\mu_e^\star$ is the target population response for event e, and μ_e^0 is the response implied by the original simulator. The structural moments provide low-dimensional empirical targets for selected features of these population responses.

Under a fixed conditioning state, consider N policies sampled from Q^0, where $\widehat{\mu}_{e,N}$ is their average response. Its expected squared deviation from the target response decomposes as

$$\mathbb{E}\left[\left(\widehat{\mu}_{e,N} - \mu_e^\star\right)^2\right] = \underbrace{\left(\mu_e^0 - \mu_e^\star\right)^2}_{\text{squared response bias}} + \underbrace{\frac{\sigma_{e,0}^2}{N}}_{\text{sampling variance}}, \tag{2}$$

where $\sigma_{e,0}^2$ is the variance of response feature $\phi_e(\pi)$ under the simulator distribution Q^0.

The decomposition separates two mechanisms behind population-level error. Increasing the simulated population size N reduces sampling variance around the existing policy distribution, while policy alignment changes Q^0 itself and directly reduces systematic response bias. This distinction links the structural deviations observed in Section 4.1 to the scaling analysis and the policy-distribution alignment introduced in Section 5.

5 How Can Structural Models Improve LLM-Based Simulations?

Section 4 shows that structural response mismatch is associated with weaker directional prediction, motivating direct alignment of the underlying policy distribution. We therefore introduce **Structure-Guided Behavioral Alignment (SGBA)**, which uses structural response moments estimated from observable market data as indirect supervision for policies generated by frozen LLM agents.

5.1 Structure-Guided Behavioral Alignment

Figure 6 summarizes SGBA. The method (1) constructs executable policy support, derives population-level behavioral targets from structural models, and (2) aligns the policy distribution before executing the original simulator.

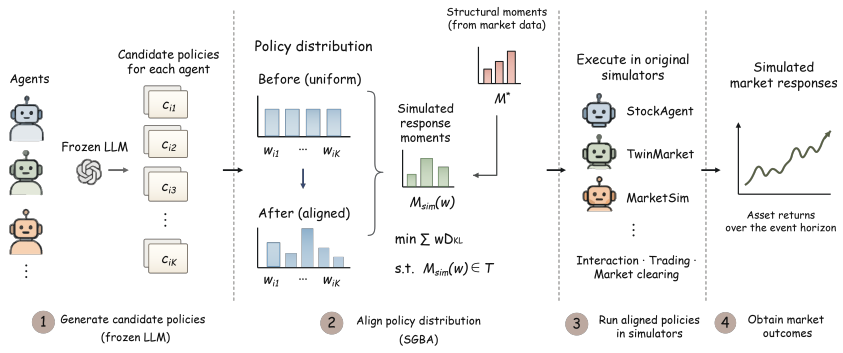


Figure 6: Overview of SGBA.

Policy generation. For each agent i, a frozen LLM generates a set of candidate policies conditioned on its persona and initial state,

$$\mathcal{F}_i = \{c_{i1}, \ldots, c_{iK_i}\}. \tag{3}$$

Each policy c_{ik} is an explicit executable decision rule that specifies how the agent responds at each decision point, given the intervention and current market state. Appendix E.1 provides an illustrative executable policy. We define the reference weights over the feasible candidates as $w^0_{ik} = 1/K_i$. This converts each agent from a single persona-conditioned decision process into an explicit distribution over executable response policies while keeping the LLM parameters fixed.

Structural supervision. Structural models combine observable market data with economic restrictions to estimate response patterns that summarize how market participants adjust to changing conditions. For event e, let $M_{\text{target},e}$ denote the resulting vector of target structural moments. Candidate policy c_{ik} is evaluated through the corresponding response features $\phi_e(c_{ik})$. Given fixed initial risk-capital shares $\omega_{i,e}$, the population response implied by policy weights w is

$$M_{\text{sim},e}(w) = \sum_i \omega_{i,e} \sum_{k=1}^{K_i} w_{ik}\phi_e(c_{ik}). \tag{4}$$

The K–Y structural model supplies demand and substitution responses for index constituent adjustments, whereas the P–R model supplies macroeconomic, risk-premium, and valuation responses for Fed rate changes. The policy features use matching conditioning states, units, and normalization so that candidate responses and structural targets occupy the same response space.

Policy distribution alignment. SGBA adjusts the relative frequency of the generated policies so that their aggregate responses are consistent with the structural targets. Let $\bar{\Sigma}_e$ denote the regularized

covariance of the structural moments and c_e the tolerance calibrated from structural draws. We solve

$$\begin{aligned} w_e^\star &= \arg\min_w \sum_i \omega_{i,e} \sum_k w_{ik} \log \frac{w_{ik}}{w_{ik}^0} \\ \text{s.t.} \quad & [M_{\text{sim},e}(w) - M_{\text{target},e}]^\top \bar{\Sigma}_e^{-1} [M_{\text{sim},e}(w) - M_{\text{target},e}] \le c_e, \\ & w_i \in \Delta(\mathcal{F}_i) \qquad \forall i. \end{aligned} \tag{5}$$

The KL objective selects the policy mixture closest to the original uniform reference among mixtures consistent with the structural response region. With fixed candidate features, the optimization is convex and requires no update to the underlying LLM.

Finally, a constrained assignment selects one aligned policy for each LLM decision unit. These policies are returned to the original multi-agent simulator, where agents interact, trade, and generate prices through its existing mechanisms. SGBA intervenes on the **distribution of agent policies**. Appendix E.1 provides the structural coordinates, feasibility screening, support expansion, and assignment checks.

5.2 Predictive Improvement

SGBA improvement. Table 2 shows consistent improvements in directional metrics for Fed Rate Change. SGBA improves Direction and DV in all configurations, with mean gains of 40.8 and 89.0 percentage points, respectively. Complete SGBA results across both settings are reported in Appendix E.2.1, where Direction and DV each improve eight of nine in Index setting.

Table 2: Within-setting and cross-setting SGBA performance. † denotes reasoning enabled.

Model	Backbone	Fed→Fed		Index→Fed	
		Direction (%)	DV (%)	Direction (%)	DV (%)
MarketSim	DeepSeek-V4-Flash†	91.7 (**+62.5**)	94.3 (**+133.6**)	41.6 (**+12.4**)	-11.2 (**+28.1**)
	GLM-5.3-Flash	91.7 (**+58.4**)	94.3 (**+128.0**)	45.8 (**+12.5**)	-5.2 (**+28.5**)
	GPT-4o	75.0 (**+48.1**)	66.4 (**+114.4**)	53.8 (**+26.9**)	9.5 (**+57.5**)
StockAgent	DeepSeek-V4-Flash†	41.7 (**+13.1**)	-0.6 (**+30.9**)	46.6 (**+18.0**)	5.6 (**+37.1**)
	GLM-5.3-Flash	16.7 (**+6.3**)	-44.4 (**+30.0**)	12.5 (**+2.1**)	-63.3 (**+11.1**)
	GPT-4o	58.3 (**+28.4**)	43.8 (**+81.5**)	40.0 (**+10.1**)	-6.0 (**+31.7**)
TwinMarket	DeepSeek-V4-Flash†	66.7 (**+39.3**)	26.9 (**+69.3**)	43.0 (**+15.6**)	-9.9 (**+32.5**)
	GLM-5.3-Flash	70.0 (**+50.1**)	30.6 (**+88.7**)	40.1 (**+20.2**)	-17.5 (**+40.6**)
	GPT-4o	81.8 (**+61.0**)	74.5 (**+124.7**)	55.0 (**+34.2**)	13.6 (**+63.8**)
	Mean gain (pp)	**+40.8**	**+89.0**	**+16.9**	**+36.8**

Cross-setting transfer. We further impose a setting transfer as stress test SGBA's generality by using structural moments from Index setting to align policies in Fed setting. Although the two differ substantially in both economic mechanism and underlying market data, Index→Fed still improves Direction and DV by average of 16.9 and 36.8 percentage points in all nine configurations, respectively. These show that SGBA's core mechanism of explicit policy representation and population policy distribution alignment remains effective even when the structural targets are not tailored to the target setting.

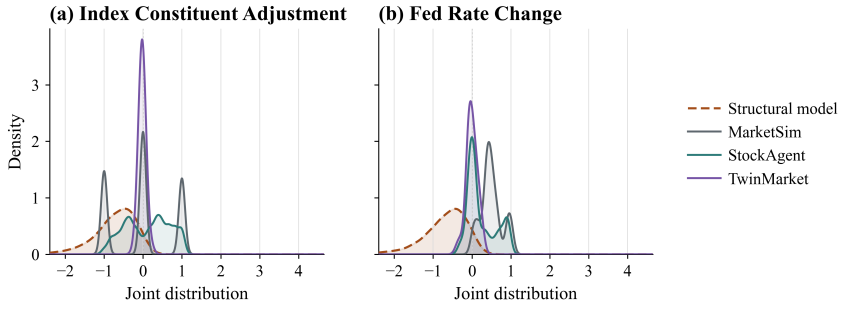


Figure 7: Response distributions after SGBA.

Distributional alignment. For within-setting alignment, Figure 7 shows that simulated response distributions move closer to their structural references after SGBA, reducing the mismatch identified in Section 4. This shift provides complementary evidence that SGBA's predictive gains coincide with the population level response alignment targeted by the method.

Finding 4 SGBA improves directional prediction both within and cross settings. ***Behavioral distribution alignment can improve predictive validity without updating the LLM.***

5.3 Policy Representation and Distribution Optimization

We isolate two main components of SGBA: explicit policy representation and structural distribution optimization. The ablation shows complementary gains from policy representation and distribution alignment. As Figure 8 shows, explicit executable policies raise mean Direction from 22.7% to 50.0% and DV from -50.2% to 8.4%, while policy distribution alignment further increases them to 72.8% and 44.0%.

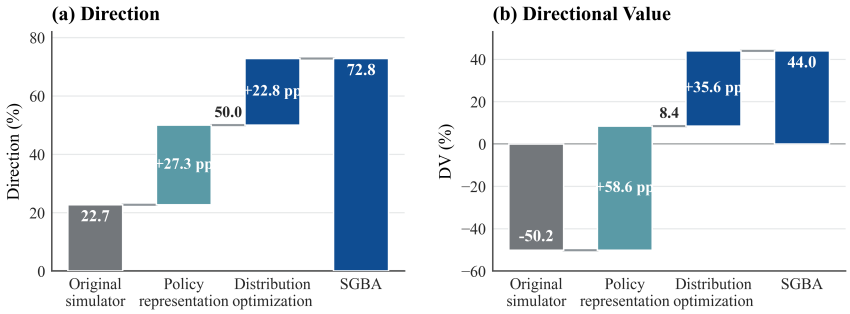


Figure 8: Ablation of policy representation and distribution optimization.

Finding 5 Explicit policies improve prediction, and policy distribution optimization yields further gains. ***Policy representation and policy distribution alignment provide successive gains.***

6 Conclusion

We evaluate the predictive validity of LLM based social simulation against structural models across two intervention settings. LLM simulators exhibit a substantial gap in directional prediction, and larger deviations from structural response distributions are consistently associated with lower directional accuracy. Motivated by this finding, we introduce Structure Guided Behavioral Alignment (SGBA), which represents agent decisions as explicit executable policies and aligns the population distribution over these policies using structural moments estimated from observable market data while keeping the LLM fixed.

Across three simulators and three LLM backbones, SGBA improves mean Direction by 12.4 percentage points in Index Constituent Adjustment and 40.8 points in Fed Rate Change, while moving simulated response distributions closer to their structural references. Ablations show complementary gains from explicit policy representation and distribution alignment. These results highlight explicit policy representation and population distribution alignment as promising principles for grounding LLM based social simulation in domain structure and empirical evidence.

Limitations and future work. SGBA currently relies on the selection of appropriate structural models and the construction of corresponding structural targets and policy features. Future work could automate these steps and investigate more general forms of structural guidance that transfer across intervention settings and domains.

AI USE STATEMENT

Codex (GPT-5.6 sol) was used in this revision to assist with manuscript editing, LaTeX preparation, consistency checks. We reviewed all AI-assisted work and take full responsibility for the final manuscript and reported results. The LLMs evaluated as simulator backbones are described separately in the experimental setup.

ETHICS STATEMENT

This study evaluates agent policies through simulations of historical financial market events. The findings are intended for research and do not constitute investment advice. Interpretation and practical use require careful assessment of the selected structural models, simulation assumptions, and applicability to other market conditions.

REPRODUCIBILITY STATEMENT

The Appendix documents the evaluation protocol, data and information cutoffs, structural model implementation, response diagnostics, and SGBA construction. Additional implementation details, configurations, and code are provided in the anonymous repository linked in the abstract.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2. URL https://doi.org/10.1017/pan.2023.2.

Richard Blundell. What have we learned from structural models? *American Economic Review*, 107(5):287–292, 2017. doi: 10.1257/aer.p20171116. URL https://doi.org/10.1257/aer.p20171116.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24761–24778. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1271. URL https://aclanthology.org/2025.findings-acl.1271/.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010. URL https://www.jmlr.org/papers/v11/ganchev10a.html.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023. doi: 10.48550/arXiv.2307.14984. URL https://arxiv.org/abs/2307.14984.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. doi: 10.1093/pan/mpr025. URL https://doi.org/10.1093/pan/mpr025.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775. URL https://doi.org/10.2307/1912775.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2fc4d6a6400addda7493b9db258c34e0-Abstract-Conference.html.

Ralph S. J. Koijen and Motohiro Yogo. A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515, 2019. doi: 10.1086/701683. URL https://doi.org/10.1086/701683.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2026. doi: 10.1007/s10462-025-11412-6. URL https://doi.org/10.1007/s10462-025-11412-6.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15523–15536, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.829. URL https://aclanthology.org/2024.acl-long.829/.

Hamish Low and Costas Meghir. The use of structural models in econometrics. *Journal of Economic Perspectives*, 31(2):33–58, 2017. doi: 10.1257/jep.31.2.33. URL https://doi.org/10.1257/jep.31.2.33.

Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, 1989. doi: 10.2307/1913621. URL https://doi.org/10.2307/1913621.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Carolin Pflueger and Gianluca Rinaldi. Why does the Fed move markets so much? a model of monetary policy and time-varying risk aversion. *Journal of Financial Economics*, 146(1):71–89, 2022. doi: 10.1016/j.jfineco.2022.06.002. URL https://doi.org/10.1016/j.jfineco.2022.06.002.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. doi: 10.48550/arXiv.2502.08691. URL https://arxiv.org/abs/2502.08691.

Jinghua Piao, Zhentao Liu, Cheng Huang, Jiarui Huang, Songwei Li, Ranran Wang, and Yong Li. MarketSim: Simulating stock markets with large-scale generative agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=EzpJxPDqXB.

Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. Distributional alignment for social simulation with LLMs: A mixture modeling approach. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 12492–12503. ACM, 2026. doi: 10.1145/3770855.3818919. URL https://doi.org/10.1145/3770855.3818919.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70830–70880, 2025. doi: 10.52202/085713-2132.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024. URL https://arxiv.org/abs/2411.11581.

Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. When AI meets finance (StockAgent): Large language model-based stock trading in simulated real-world environments, 2024. URL https://arxiv.org/abs/2407.18957.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. SocioVerse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users, 2025. URL https://arxiv.org/abs/2504.10157.

A Evaluation Protocol and Metrics

A.1 Event Protocol

Table A1 records the event dates and evaluation timing for both settings.In Panel B, the first date is the effective close and the second is the first trading day used in this study. Index targets are stock log returns and cumulative abnormal returns from the announcement close to the effective close.

Table A1: Event protocol.

Panel A: FOMC Events			
Event	**Date**	**Timing**	**Scored assets**
E01	2024-09-18	Announcement: 14:00	SPY, XLB, XLC, XLE,
E02	2024-11-07	State cutoff: 13:59:59	XLF, XLI, XLK, XLP,
E03	2024-12-18	Scoring: 14:19:30–14:20:30	XLRE, XLU, XLV, XLY
E04	2025-01-29		
Panel B: SSE50 Constituent Adjustments			
Event date	**Eff. / Trade**	**Additions**	**Deletions**
2024-11-29	2024-12-13 2024-12-16	601127, 601766, 601816, 603993, 688256	600089, 600104, 600436, 601669, 603986
2025-05-30	2025-06-13 2025-06-16	600760, 601211, 601600, 688008	600438, 601390, 601633, 603288
2025-11-28	2025-12-12 2025-12-15	600104, 600111, 600930, 603019	600048, 600941, 601600, 601766
2026-05-29	2026-06-12 2026-06-15	600089, 600183, 601600, 601688, 603986	600104, 600690, 601225, 601816, 601985

A.2 Evaluation Metrics

Let $\mathcal{A}_e$ be the prespecified set of scored assets in event e, with $n_e = |\mathcal{A}_e|$. For asset i, $\widehat{y}_{ei}$ and y_{ei} denote the predicted and realized responses on the same outcome scale. We define the event-level

directional measures as

$$\text{Direction}_e = \frac{1}{n_e} \sum_{i \in \mathcal{A}_e} \mathbb{I}\{\text{sign}(\widehat{y}_{ei}) = \text{sign}(y_{ei})\}, \tag{6}$$

$$\text{DV}_e = \frac{\sum_{i \in \mathcal{A}_e} |y_{ei}| \left(2\mathbb{I}\{\widehat{y}_{ei} y_{ei} > 0\} - 1\right)}{\sum_{i \in \mathcal{A}_e} |y_{ei}|}. \tag{7}$$

Direction is the share of assets whose predicted sign matches the realized sign, with $\text{sign}(0) = 0$. Directional Value (DV) weights each directional judgment by the magnitude of the realized response. Both measures are reported as percentages.

For magnitude errors, we use

$$\text{MAE}_e = \frac{1}{n_e} \sum_{i \in \mathcal{A}_e} |\widehat{y}_{ei} - y_{ei}|, \tag{8}$$

$$\text{RMSE}_e = \sqrt{\frac{1}{n_e} \sum_{i \in \mathcal{A}_e} (\widehat{y}_{ei} - y_{ei})^2}. \tag{9}$$

To assess cross-asset ordering, we compute Spearman correlation within each event:

$$\text{Spearman}_e = \text{Corr}_{i \in \mathcal{A}_e}\left(R_{ei}^{\text{pred}}, R_{ei}^{\text{real}}\right), \tag{10}$$

where R_{ei}^{pred} and R_{ei}^{real} are the within-event ranks of predicted and realized responses, with average ranks assigned to ties. MAE and RMSE are expressed in percentage points for Index Constituent Adjustment and in basis points for Fed Rate Change. Spearman correlation has no return unit and ranges from -1 to 1.

B Model Implementation

B.1 Structural Models

The base structural predictions use parameters and state measurements fixed without reference to realized returns from the four evaluation events. Inputs for each event are restricted to information available at the cutoff in Table A1, together with the intervention when it is announced.

B.1.1 K–Y Demand System

Model specification. The Koijen–Yogo system assigns each investor group a demand curve over stocks and an outside asset based on stock characteristics. Portfolio weights depend on prices and characteristics; market clearing equates the resulting share demand, including demand from index trackers and residual holdings, with point-in-time free-float share supply. We estimate separate demand parameters for active equity funds, active mixed funds, and enhanced index funds, each split into three AUM groups. Mechanical passive index funds are excluded from the estimation groups and enter the counterfactual through their holdings and flows induced by index changes.

For fund h in estimation group $g(h)$, the characteristic-demand equation for a positive holding is

$$\begin{aligned} r_{hi,t} \equiv \frac{w_{hi,t}}{w_{h0,t}} &= \exp\left(a_{h,t} + \beta_{g(h),t} \log M_{i,t} + \boldsymbol{\theta}_{g(h),t}^{\top} X_{i,t} + \xi_{hi,t}\right), \\ w_{hi,t} &= \frac{r_{hi,t}}{1 + \sum_j r_{hj,t}}. \end{aligned} \tag{11}$$

Here $w_{h0,t}$ is the weight in the outside asset, $M_{i,t} = p_{i,t} S_{i,t}^{\text{float}}$ is the free-float market equity used in this implementation, $X_{i,t}$ contains the observed characteristics, and $\xi_{hi,t}$ is latent demand. Zero holdings enter the nonlinear GMM moments without taking their logarithm. In the event counterfactual, the estimated price sensitivities and recovered latent demand terms are held fixed while prices change.

Let A_e be the AUM tracking the SSE50 before the announcement, and let ω_{ie}^- and ω_{ie}^+ be its weights before the change and those implied by the announcement. The resulting SSE50 cash flow shock and market clearing condition are

$$F_{ie} = A_e(\omega_{ie}^+ - \omega_{ie}^-), \tag{12}$$

$$\sum_{h \in \mathcal{H}_e} W_{he} w_{hi}(p) = p_i \left(S_{ie}^{\text{float}} - H_{ie}^{\text{pass}} - \frac{F_{ie}}{p_{ie}^0} \right). \tag{13}$$

Here $\mathcal{H}_e$ includes the active and residual demand units, W_{he} is their investable wealth, H_{ie}^{pass} is disclosed passive tracker ownership, and p_{ie}^0 converts the flow in currency into shares. Solving the clearing equation before and after the flow produces each draw's predicted change in log price, $\log(p_{ie}^+/p_{ie}^-)$.

Calibration data and parameters. CSMAR supplies fund holdings, fund classifications and AUM, stock prices and returns, and accounting variables. RESSET supplies historical free-float shares; each record is admitted only after its publication and effective dates. The estimation panel contains 19 semiannual snapshots from June 2015 through June 2024. The first six establish the rolling investment universe, so demand parameters are estimated on snapshots from June 2018 through June 2024 and are held fixed for all four events. The five standardized stock characteristics are log book equity, profitability, investment, dividends relative to book equity, and market beta. Characteristics are winsorized at the first and 99th percentiles; beta uses up to 60 monthly returns, with at least 36 required. A linear fit with instrumental variables initializes the nonlinear GMM estimator, which incorporates zero holdings and caps the price coefficient at 0.99.

B.1.2 P–R Monetary Model

Model specification. The Pflueger–Rinaldi model maps an innovation in monetary policy into changes in macroeconomic states, surplus consumption, and the present value of an equity claim. We use the locked model calibration to construct full and risk-neutral valuation grids in five state dimensions. The implementation preserves, among other upstream parameters, $\gamma = 2$, $\phi = 0.93$, $\delta = 0.6666$, $\gamma_\pi = 1.5$, and $\rho_i = 0.8$; the four evaluation events do not re-estimate these values. The habit state is the surplus consumption ratio $S_t = (C_t - H_t)/C_t$, represented in the numerical solution by $\hat{s}_t = \log S_t - \log \bar{S}$, its log deviation from steady state. The difference between full and risk-neutral responses identifies the model's risk premium component.

Let $\Delta i_e = i_e^{\text{ann}} - \mathbb{E}_{e^-}[i_e]$ be the surprise in the policy rate, measured in percentage points, and λ_{rate} the model's rate loading. For the macro state z_e^- before the announcement, the event update is

$$\begin{aligned} u_e &= \frac{\Delta i_e}{\lambda_{\text{rate}}}, & z_e^+ &= z_e^- + AQ_{\cdot 3} u_e, \\ \hat{s}_e^+ &= \hat{s}_e^- + \Lambda(\hat{s}_e^-) Q_{13} u_e. \end{aligned} \tag{14}$$

Within the admissible state range, $\Lambda(\hat{s}) = \sqrt{1 - 2\hat{s}}/\bar{S} - 1$. The fifth state, for lagged growth, remains fixed during the instantaneous update. The locked full and risk-neutral grids provide valuation functions $V_k(s)$, $k \in \{\text{full}, \text{rn}\}$. For each draw, the gross claim response implied by the model and the log return of asset i are

$$\begin{aligned} R_e^{\text{claim},k} &= \exp(Q_{13} u_e) \frac{V_k(s_e^+)}{V_k(s_e^-)}, \\ \widehat{y}_{ei}^{\,k} &= \log\left(\frac{R_e^{\text{claim},k} - (1 - \delta_i)}{\delta_i} \right). \end{aligned} \tag{15}$$

where δ_i is the leverage parameter of the claim matched to the ETF's beta measured before the event. The full return is the structural prediction, and $\widehat{y}_{ei}^{\text{full}} - \widehat{y}_{ei}^{\text{rn}}$ is its risk premium component.

Calibration data and parameters. Macro states before each event are filtered from the Philadelphia Fed Real-Time Data Set for Macroeconomists series for real output, real consumption, and core consumption prices, together with the New York Fed effective federal funds rate. A one-sided trend using up to 40 quarters measures the output gap, and inflation is measured relative to a 2% annual target. The most recent admissible data vintage is the previous calendar month. The prior

trading day's Cboe VIX close anchors the surplus consumption state. Calibration of the state filter ends on July 31, 2024. Asset exposures are estimated from the preceding 60 complete months of total returns for SPY and the sector ETFs, constructed from State Street's official NAV and distributions; resampling in six-month moving blocks yields joint beta draws. The policy surprise is the announced midpoint of the target range, verified against archived Federal Reserve statements, less the probability-weighted midpoint expected before the announcement as reconstructed from CME 30-Day Federal Funds futures settlements.

B.2 LLM SIMULATORS

We adapt StockAgent, TwinMarket, and MarketSim to two experimental settings: Index Constituent Adjustment and Fed Rate Change. Within each setting, the historical information cutoff, announced intervention, evaluated assets, and prediction horizon are held constant across the three simulators. In the baseline runs, each simulator retains its own agent decision process and market-clearing mechanism. The LLM backbones are DeepSeek-V4-Flash, GLM-5.3-Flash, and GPT-4o. For each event, every simulator–backbone configuration is evaluated using three random seeds. At each decision step, a simulator constructs inputs from information available at that time and its current market state. Investors or managers make decisions, orders are executed through the corresponding market mechanism, and the resulting account and market states enter subsequent decisions.

B.2.1 STOCKAGENT

StockAgent consists of heterogeneous LLM investors, a public forum, and a sequential order-matching market. For each index event, a baseline StockAgent run uses 100 LLM investors to trade the evaluated stocks. For each Fed event, a baseline run uses 24 LLM investors and one neutral liquidity provider.

Before making a trading decision, an investor receives its profile, current cash balance and holdings, simulated market prices and quotes, available public information, and visible forum messages.

In the Index Constituent Adjustment setting, an investor generates a structured decision to buy, sell, or abstain. The decision specifies an asset, order quantity, and intended price. The execution layer converts the decision into a limit order and verifies that the asset is tradable, the quantity and price are valid, and the investor has sufficient cash or sellable holdings. Orders that fail these checks are not submitted to the order book.

Valid orders are submitted sequentially according to the agents' action order. An incoming order executes when its price crosses a resting order on the opposite side of the book. Eligible unfilled quantities may remain in the book until they are filled, replaced, or expire. Each trade updates the cash and holdings of the buyer and seller, the latest transaction price, and the market state observed by subsequent agents. New forum messages enter investors' information sets once they become visible.

In the Fed Rate Change setting, investors specify target portfolio weights and limit prices for the evaluated assets. The execution layer converts these targets into whole-share buy or sell quantities using current portfolio value, asset prices, and existing holdings. The resulting orders are submitted subject to cash and inventory constraints.

B.2.2 TWINMARKET

TwinMarket incorporates investor profiles, account states, market observations, and social interactions into a belief–desire–intention (BDI) decision process. For each index event, a baseline TwinMarket run uses 100 LLM investors. For each Fed event, a baseline run uses 24 LLM investors and one neutral liquidity provider.

Investors form structured trading intentions based on their current states. The execution layer converts these intentions into orders specifying an asset, side, quantity, and limit price, subject to cash, inventory, and price constraints. Intentions that fail these checks are excluded from market clearing.

Valid orders are grouped by asset for a call auction. Buy and sell limit prices and executable quantities determine the clearing price and matched volume, subject to account constraints. Trades update

investors' cash balances, holdings, and asset prices. The resulting market observations and newly visible social information enter the next BDI update and trading decision.

B.2.3 MARKETSIM

MarketSim separates LLM-based strategy generation by managers from rule-based order execution by traders. For each evaluated stock in an index event, a baseline MarketSim run uses 10 LLM managers and 2,950 manager-guided rule-based traders. The market also contains stock-specific populations of rule-based noise, value, momentum, and market-making agents. For each Fed event, a baseline run uses four LLM managers directing 24 rule-based execution accounts and one neutral liquidity provider.

In the Index Constituent Adjustment setting, a manager observes multiple levels of the order book, recent trades, technical indicators, the aggregate performance of its supervised traders, and an available public-information signal. The manager produces strategy guidance and an indicative price. Rule-based traders use the manager's guidance, the best bid and ask prices, and their displayed quantities to determine order side and limit price under predefined trading rules. They submit limit orders with preset quantities.

The MarketSim market operates as a continuous double auction. When a new order arrives, the system examines the opposite side of the order book in price–time order. Executable quantities trade immediately, while any remainder rests in the book or is canceled according to the market's order rules. Trades update both counterparties' accounts, the latest transaction price, and the order book. Price indicators are updated as trading proceeds, and these outcomes enter subsequent manager and trader decisions.

In the Fed Rate Change setting, managers observe the announcement, market quotes, and the states of their managed portfolios. They specify target asset weights and limit prices. The execution layer allocates these institutional targets among rule-based traders, which then generate orders based on their current accounts and the prevailing order book.

Prediction extraction. In the Index Constituent Adjustment setting, simulated price paths are used to calculate predicted stock log returns and cumulative abnormal returns from the announcement close to the effective close. In the Fed Rate Change setting, let P_j^{pre} denote the frozen pre-announcement reference price for asset j, and let $m_{j,t}$ denote a valid bid–ask midquote at time t. Let $\widetilde{m}_j$ be the median of valid midquotes observed from 1170 to 1230 seconds after the announcement. The simulated return in basis points is

$$\widehat{r}_j = 10^4 \log\left(\frac{\widetilde{m}_j}{P_j^{\mathrm{pre}}}\right). \tag{16}$$

B.3 INPUT CONTROLS

All comparisons use the event definitions, scored assets, and timing in Table A1; a common cutoff does not imply identical model inputs. Structural models use holdings, macroeconomic states, and measures of market expectations specific to each model, while the LLM simulators use their admissible inputs for each event. The intervention is introduced at simulation start. Prices and realized returns after the event do not enter the generation of raw prediction draws. Estimation of structural parameters and asset betas, construction of event states, generation of candidate policies, and calibration of structural targets are restricted to information available by the relevant cutoff or released with the intervention. In the Fed setting, the surprise in the target rate compares the announced decision with the expectation reconstructed from futures data available before the announcement.

C ADDITIONAL BASELINE RESULTS

C.1 EVENT-LEVEL RESULTS

Figure 9 retains the DV distributions for each event across structural draws and simulator seeds. Structural boxes show variation across draws, while simulator markers and ranges show variation across seeds; these are distinct uncertainty sources.

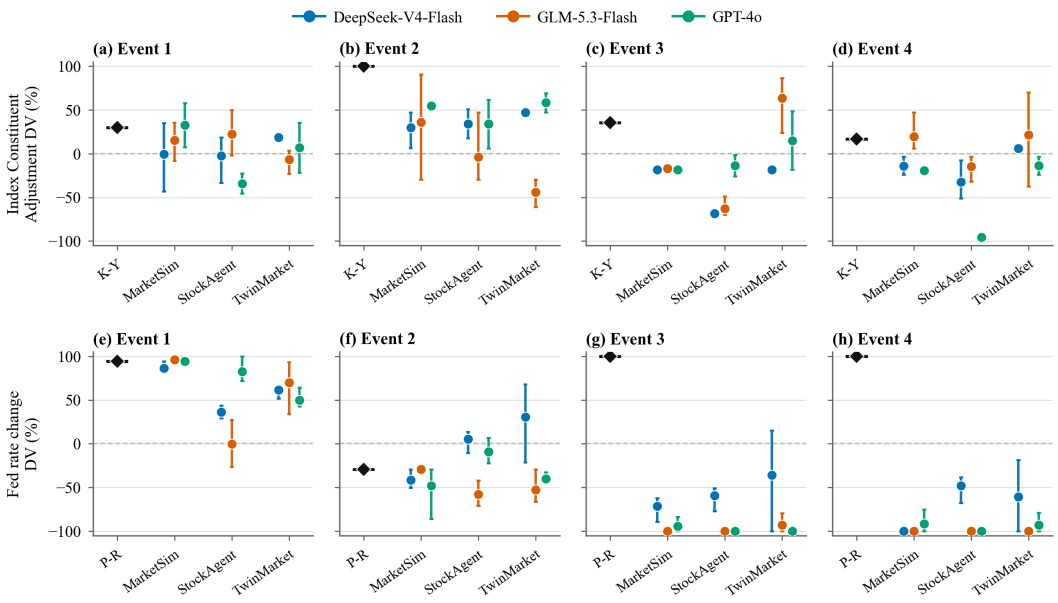


Figure 9: Event-level Directional Value across seeds and structural draws.

C.2 Resource Use

Table A2 gives input, output, and total LLM tokens for every baseline combination of simulator and backbone. Figure 10 relates token usage to Direction. The archived totals are aggregated per configuration.

Table A2: Token usage and directional performance.

Setting	Simulator	Backbone	Tokens (millions)		
			Input	Output	Total
Index	MarketSim	DeepSeek	38.387	53.723	92.109
	MarketSim	GLM	46.278	20.925	67.203
	MarketSim	GPT-4o	31.903	7.291	39.194
	StockAgent	DeepSeek	175.744	27.988	203.732
	StockAgent	GLM	185.422	1.245	186.667
	StockAgent	GPT-4o	164.466	0.356	164.823
	TwinMarket	DeepSeek	172.735	70.779	243.514
	TwinMarket	GLM	173.649	6.718	180.367
	TwinMarket	GPT-4o	170.213	4.346	174.559
Fed	MarketSim	DeepSeek	0.368	0.476	0.844
	MarketSim	GLM	0.382	0.082	0.464
	MarketSim	GPT-4o	0.353	0.074	0.426
	StockAgent	DeepSeek	5.580	8.720	14.300
	StockAgent	GLM	6.739	1.487	8.226
	StockAgent	GPT-4o	6.177	0.897	7.074
	TwinMarket	DeepSeek	7.202	11.288	18.491
	TwinMarket	GLM	7.771	1.721	9.492
	TwinMarket	GPT-4o	7.252	1.276	8.527

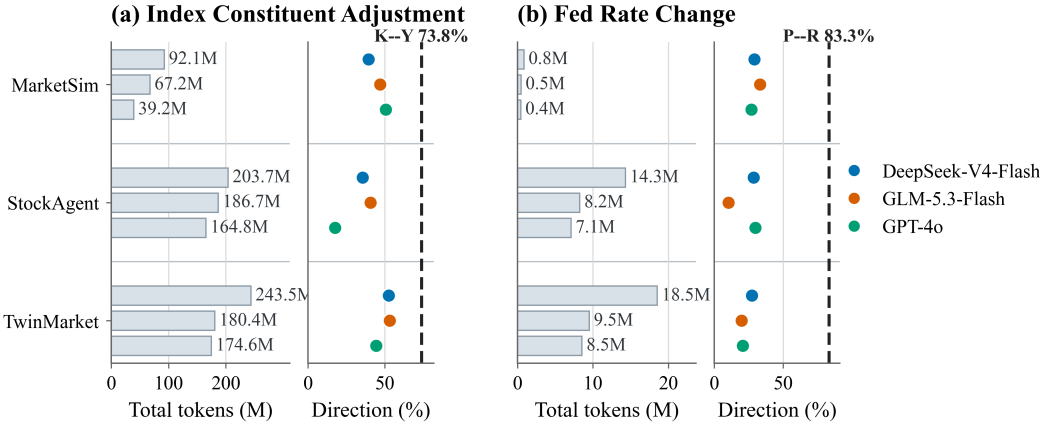


Figure 10: Token usage and directional performance.

D Structural Mismatch Diagnostics

D.1 Distance Construction

For each run r in event e, we summarize market responses in a two-dimensional vector $\widehat{\boldsymbol{\mu}}_r$. For Index Constituent Adjustment, the two entries are the mean log returns of added and deleted stocks. For Fed Rate Change, they are the SPY log return and the slope from regressing sector returns on their fixed pre-announcement market betas. Let $\boldsymbol{\mu}_e^\star$ and $\boldsymbol{\Sigma}_e^\star$ denote the mean and covariance of the structural responses for the same event. We measure deviation using the squared Mahalanobis distance:

$$D_{r,e}^2 = (\widehat{\boldsymbol{\mu}}_r - \boldsymbol{\mu}_e^\star)^\top (\boldsymbol{\Sigma}_e^\star)^{-1} (\widehat{\boldsymbol{\mu}}_r - \boldsymbol{\mu}_e^\star). \tag{17}$$

The joint response diagnostic in Equation 17 places each simulator run and its matched structural draws in the same response space with two dimensions. For index setting, the coordinates are the mean changes in log price for added and deleted stocks; all runs in an event share the same structural reference. For Fed setting, they are the SPY log return and the slope of sector responses on betas fixed before the announcement. Let $S_{r,e}^\star$ be the diagonal matrix of coordinate standard deviations and $R_{r,e}^\star$ their correlation matrix across the matched structural draws. The regularized covariance used in the distance is

$$\bar{\boldsymbol{\Sigma}}_{r,e}^\star = S_{r,e}^\star (R_{r,e}^\star + 10^{-4} I_2) S_{r,e}^\star. \tag{18}$$

The diagnostic threshold $c_{r,e}^{\text{diag}}$ is the empirical 95th percentile of the matched structural draws' squared distances from their own mean under this covariance. Dividing by that threshold makes $\log_{10}(D_{r,e}^2 / c_{r,e}^{\text{diag}})$ equal to zero at the reference boundary. This diagnostic and its threshold are example distinct from the SGBA constraint, which uses more dimensions in Equation 5. The structural reference density in Figure 3 is a χ_2^2 approximation scaled by the thresholds for each run.

D.2 Structural Deviation Groups

Each dot in Figure 4 represents one run identified by its simulator, backbone, event, and seed. Repeated simulations enter the diagnostic separately, preserving variation across runs in both structural deviation and directional accuracy.

For index constituent adjustments, let Δ_r^{ABM} denote the difference between addition and deletion responses in run r, and $\Delta_e^{\text{K–Y}}$ the corresponding difference between structural means for event e. We define the relative deviation as

$$d_r^{\text{Index}} = \frac{\left|\Delta_r^{\text{ABM}} - \Delta_{e(r)}^{\text{K–Y}}\right|}{\text{median}_{u \in \mathcal{R}_{\text{Index}}} \left|\Delta_u^{\text{ABM}} - \Delta_{e(u)}^{\text{K–Y}}\right|}, \qquad |\mathcal{R}_{\text{Index}}| = 108. \tag{19}$$

The left panel assigns these 108 values to empirical tertiles with low, medium, and high deviation.

For Fed rate changes, let x_r denote the simulated SPY log return and b_r the slope of sector responses within run r on betas fixed before the announcement. Let $\mu_{e,x}^{\text{P–R}}$ and $\mu_{e,b}^{\text{P–R}}$ denote the corresponding means across matched structural draws for event e. We define indicators of directional consistency

$$c_r^{\text{SPY}} = \mathbf{1}\{\text{sgn}(x_r) = \text{sgn}(\mu_{e,x}^{\text{P–R}})\}, \qquad c_r^{\text{slope}} = \mathbf{1}\{\text{sgn}(b_r) = \text{sgn}(\mu_{e,b}^{\text{P–R}})\}. \tag{20}$$

The Fed deviation score is the number of directional mismatches,

$$d_r^{\text{Fed}} = 2 - c_r^{\text{SPY}} - c_r^{\text{slope}}, \qquad d_r^{\text{Fed}} \in \{0, 1, 2\}. \tag{21}$$

The right panel maps $d_r^{\text{Fed}} = 0$, 1, and 2 to groups with low, medium, and high deviation, respectively. The one run with insufficient terminal quotes is excluded because neither output moment can be computed.

E SGBA

E.1 Method

E.1.1 Targets and Features

Structural moments characterize how demand, risk assessments, and valuations respond to changes in market conditions under investor preferences, beliefs, and portfolio constraints. For each setting, m_e^S stacks the selected numerical response coordinates, while Σ_e^S summarizes their joint variation across structural draws.

For index constituent adjustments, the K–Y structural model supplies six ordered moments. They comprise changes in the quantity held for net additions and deletions, substitution responses within industry, liquidity, and size groups, and the cash response. These coordinates are evaluated per unit change in log price around the baseline on the announcement date. For a trigger asset m and perturbation $\epsilon = 0.05$, the response in its own quantity is computed from executable portfolio weights as

$$r_m^{\text{own}} = \frac{w_m(+\epsilon)e^{-\epsilon} - w_m(-\epsilon)e^{\epsilon}}{2\epsilon}. \tag{22}$$

Let x denote the change in the trigger asset's log price and $w_m(x)$ its executable portfolio weight under that perturbation. With nominal wealth held fixed and prices normalized to the baseline, $w_m(x)e^{-x}$ is proportional to the quantity held. The central difference therefore measures a normalized quantity response to the change in log price.

For a relation set $R(m)$ defined by industry, liquidity, or size, the substitution response is

$$r_{m,R}^{\text{sub}} = \sum_{n \in R(m)} \frac{w_n(+\epsilon) - w_n(-\epsilon)}{2\epsilon}. \tag{23}$$

Within each event, aggregation uses fixed trigger weights. The cash coordinate applies the same central difference to the candidate's cash weight under price perturbations across the market.

For Fed rate changes, the P–R structural model supplies five ordered moments: responses of growth, inflation, and the surplus consumption level; the SPY risk premium component; and the slope of full valuation revisions across asset betas. The surplus consumption level summarizes the risk aversion state. For each asset, the full valuation revision equals its cashflow component minus the change in its required return. The last coordinate is the slope from a linear regression with an intercept of full valuation revisions on betas fixed before the announcement across the 11 sector ETFs, excluding SPY. The same betas are used for structural responses and candidate policy features.

Each coordinate of $\phi_e(c_{ik})$ measures the corresponding response implied by a selected policy. In the index setting, features are computed from executable target portfolios after feasibility screening. In the Fed setting, responses of growth, inflation, and the surplus consumption level provide the state coordinates. The negative of the change in SPY's required return provides the risk premium coordinate, and the regression described above provides the valuation coordinate across assets. All coordinates use the same conditioning states, units, and normalization as their structural counterparts. Equation 4 aggregates these features using the shares of total investable risk capital held by the LLM decision units at the event start.

E.1.2 CANDIDATE POLICIES

A candidate policy is a repeatable decision rule: at each decision point it maps the announced intervention and prices generated within the simulator to belief revisions and desired holdings. In the Fed setting, it specifies growth, inflation, and surplus-consumption responses as well as cashflow and required return revisions for each asset. Their difference gives the valuation revision used in the target weight. Candidate targets are screened for budget feasibility, the ban on short selling, and the simulator's cash, concentration, and turnover constraints before entering the support $\mathcal{F}_i$ in Equation 3.

Illustrative Fed policy. Consider a balanced investor initially holding 40% cash, 16% SPY, and 4% in each sector ETF. One synthetic candidate c_{i1} interprets an unexpectedly lower Fed target as favorable to growth and equity valuation. Let Δ be the rate surprise, $r_{j,t} = \log(P_{j,t}/P_{j,0})$ the return generated inside the simulator, and β_j a pre-announcement asset beta. Its conditional rule is

$$\begin{aligned} z_t &= -\Delta/(10\,\mathrm{bp}) + 5r_{\mathrm{SPY},t}, \\ (g_t, \pi_t, s_t) &= (0.00030, 0.00005, 0.00025)z_t, \\ \widetilde{w}_{j,t} &= w_{j,0} + 0.010\beta_j z_t - r_{j,t}. \end{aligned} \tag{24}$$

The first line combines the rate surprise with the simulated SPY price change; the second records the candidate's economic and risk-state responses. In the third, $0.010\beta_j z_t$ is the valuation revision, combining a $+0.006\beta_j z_t$ cashflow revision and a $-0.004\beta_j z_t$ required-return revision. Subtracting $r_{j,t}$ reduces the remaining adjustment when the simulated price has already moved. For $\Delta = -10\,\mathrm{bp}$ and $\beta_{\mathrm{SPY}} = 1$, the announcement-state signal is $z_0 = 1$, so the provisional SPY target is $0.16 + 0.01 = 0.17$. After a 1% simulated SPY log return, $z_t = 1.05$ and the same rule instead gives $0.16 + 0.0105 - 0.01 = 0.1605$. These are two outputs of one policy, not fixed holdings or guaranteed fills. Targets remain subject to the investor's long-only mandate, 10% cash floor, 25% per-asset cap, and 20% gross turnover budget, followed by the simulator's native order checks. The coefficients are synthetic.

E.1.3 ALIGNMENT AND ASSIGNMENT

The optimization in Equation 5 minimizes the change from the uniform reference distribution over feasible policies subject to the structural tolerance region. Its covariance $\bar{\Sigma}_e$ is the sample covariance of complete joint draws of structural moments with a relative diagonal ridge of 10^{-4}; c_e is the empirical 95th percentile of those draws' regularized squared distances. These targets have six coordinates for index adjustments and five for Fed decisions, unlike the diagnostic with two dimensions in Section D.1. Policy support remains fixed within each solve. When additional support is required, one prespecified expansion step generates further policies along dimensions such as price sensitivity, substitution, and risk taking. We then repeat feasibility screening, reinitialize the uniform reference distribution over the expanded feasible policy set, and resolve the alignment problem. The constrained assignment that minimizes cost selects one policy for each LLM decision unit from the final solution. We record the final feasibility status and the left side of the structural constraint in Equation 5, accepting a run only when it is at most c_e.

E.2 RESULTS

E.2.1 COMPLETE PERFORMANCE

Table A3 reports the aligned outcomes and changes in the complementary metrics for every combination of simulator and backbone in both intervention settings.

Table A3: SGBA Predictive performance.

Model	Backbone	Direction (%) ↑	DV (%) ↑	MAE (%) ↑	RMSE (%) ↑	Δ Spearman ↑
Panel A. Index Constituent Adjustment						
MarketSim	DeepSeek-V4-Flash†	62.5 (**+23.1**)	47.1 (**+59.7**)	**+6.4**	-11.6	-0.059
	GLM-5.3-Flash	62.5 (**+15.6**)	47.1 (**+59.6**)	-50.1	-69.2	-0.046
	GPT-4o	57.1 (**+6.5**)	43.8 (**+38.9**)	**+3.0**	-9.3	-0.250
StockAgent	DeepSeek-V4-Flash†	50.0 (**+14.4**)	-21.9 (**+12.4**)	-7.4	-6.9	-0.152
	GLM-5.3-Flash	60.0 (**+19.4**)	-4.3 (**+23.7**)	-4.2	-4.7	**+0.444**
	GPT-4o	30.0 (**+12.5**)	-65.9 (-24.5)	-11.4	-10.2	-0.154
TwinMarket	DeepSeek-V4-Flash†	40.0 (-12.5)	18.7 (**+5.4**)	**+40.1**	**+40.0**	**+0.364**
	GLM-5.3-Flash	60.0 (**+6.9**)	-6.6 (**+17.7**)	**+25.9**	**+19.8**	**+0.219**
	GPT-4o	70.0 (**+25.6**)	47.7 (**+46.3**)	-11.9	-25.6	**+0.421**
Panel B. Fed Rate Change						
MarketSim	DeepSeek-V4-Flash†	91.7 (**+62.5**)	94.3 (**+133.6**)	**+7.5**	**+16.0**	**+0.250**
	GLM-5.3-Flash	91.7 (**+58.4**)	94.3 (**+128.0**)	**+17.5**	**+24.7**	**+0.182**
	GPT-4o	75.0 (**+48.1**)	66.4 (**+114.4**)	**+7.2**	**+16.3**	**+0.288**
StockAgent	DeepSeek-V4-Flash†	41.7 (**+13.1**)	-0.6 (**+30.9**)	**+19.5**	**+25.1**	**+0.195**
	GLM-5.3-Flash	16.7 (**+6.3**)	-44.4 (**+30.0**)	**+21.3**	**+37.9**	-0.062
	GPT-4o	58.3 (**+28.4**)	43.8 (**+81.5**)	-2.8	**+5.6**	**+0.015**
TwinMarket	DeepSeek-V4-Flash†	66.7 (**+39.3**)	26.9 (**+69.3**)	**+0.8**	**+8.0**	-0.705
	GLM-5.3-Flash	70.0 (**+50.1**)	30.6 (**+88.7**)	**+12.8**	**+20.5**	-0.719
	GPT-4o	81.8 (**+61.0**)	74.5 (**+124.7**)	-5.4	**+4.6**	-0.493

E.2.2 Ablation

Table A4 complements Figure 8 by reporting the complete metric set for the controlled ablation, averaged across the three LLM backbones. Direction and DV are the primary outcomes visualized in Figure 8; MAE, RMSE, and Spearman correlation summarize complementary magnitude and ranking performance.

Table A4: SGBA component ablation.

Configuration	Direction (%) ↑	DV (%) ↑	MAE (bp) ↓	RMSE (bp) ↓	Spearman ↑
Original simulator	22.7	-50.2	41.74	52.53	**0.299**
Unaligned policies	50.0	8.4	**38.10**	**43.93**	-0.223
SGBA	**72.8**	**44.0**	40.48	46.59	-0.340