ProcureGym: A Multi-Agent MDP Framework for Modeling National Volume-based Drug Procurement

Anonymous ACL submission

Abstract

In this paper, we introduce ProcureGym, an open-source, data-driven multi-agent simulation platform that models China's National Volume-Based drug Procurement (NVBP) as a Markov decision process (MDP). Based on real-world data from 7 rounds of NVBP (covering 328 drugs and 2,226 firms), the platform establishes a high-fidelity simulation environment. Within this framework, we evaluate diverse agent models, including Reinforcement Learning (RL), Large Language Model (LLM), and Rule-based algorithms. Experimental results demonstrate that RL agents achieve superior winner alignment and profits. Further analyses show that maximum valid bidding price and procurement volume dominate strategic outcomes. ProcureGym thus serves as a rigorous instrument for assessing policy impacts and formulating future procurement strategies.

1 Introduction

Initiated in 2018, China's National Volume-Based drug Procurement (NVBP) represents a landmark reform in pharmaceutical pricing, achieving substantial cost reductions through centralized competitive bidding (Xinhuanet, 2023; Zhu et al., 2023, 2025). As the program expands to encompass more than 400 drugs and thousands of participating firms, the procurement process has evolved into a highly complex decision-making environment (Cao et al., 2024). Firms strategically determine bid prices to balance expected profits and winning probabilities under regulatory constraints such as government-specified procurement volumes and price ceilings. Accurate simulation is key to understanding strategic interactions, predicting policy responses, and refining procurement strategies.

While a growing body of research explores computational modeling of economic systems, existing studies largely focus on generic or macroeconomic scenarios. As summarized in Table 1, the current landscape centers on fiscal and taxation policy optimization (Zheng et al., 2021; Curry et al., 2022; Mi et al., 2024; Li et al., 2024; Ponse et al., 2025). Moreover, recent scholarship has extended to broader governance and market dynamics (Dwarakanath et al., 2025; Brusatin et al., 2024; Mi et al., 2025; Piao et al., 2025) and macroeconomic policy simulations (Mi et al., 2025; Piao et al., 2025). Consequently, research on simulation environments dedicated to specific vertical microeconomic domains, particularly centralized procurement, remains virtually unexplored.

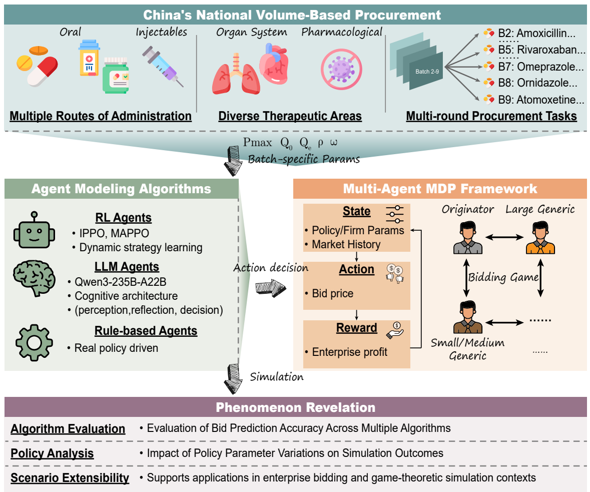

Figure 1: Overview of the ProcureGym Framework.

Unlike macroeconomic models that focus on system-level equilibrium, simulating such micro-level bidding games presents fundamental challenges: capturing firm-level heterogeneity, modeling dynamic strategic adaptation, and predicting behaviors under counterfactual interventions. This makes it non-trivial to apply existing models directly to this domain. To bridge this gap, we present **ProcureGym**, a Markov decision process (MDP)-based multi-agent simulation framework specifically designed for microeconomic scenarios, such as multi-firm bidding games in pharmaceutical procurement, as illustrated in Figure 1.

Platform	Year	Domain	Algorithms	Scenarios	LLM Support	Real Data	Open Source
AI Economist	2022	Tax policy	RL	2–3	✗	✗	✓
RBC model	2022	Tax policy	RL	2	✗	✗	✓
TaxAI	2024	Tax policy	RL	4	✗	✓	✓
EconAgent	2024	Macro-economy	LLM	1	✓	✓	✓
R-MABM	2024	Market competition	RL	4	✗	✗	✓
ABIDES-Economist	2024	Macro-economy	RL, Rule	4+	✗	✓	✗
AgentSociety	2025	Macro-economy	LLM, Rule	4	✓	✗	✗
EconoJax	2025	Tax policy	RL	4+	✗	✗	✓
EconGym	2025	Macro-economy	RL, LLM, Rule	25+	✓	✓	✓
ProcureGym (Ours)	**2025**	**Drug procurement**	**RL, LLM, Rule**	**360+**	✓	✓	✓

Table 1: Comparison with Related Economic Simulation Platforms.

Our contributions are summarized as follows:

- We propose a MDP-based multi-agent system for NVBP with a unified interface that supports heterogeneous agents, including Reinforcement Learning (RL)-, Large Language Model (LLM)-, and Rule-based policies.
- The platform accurately reproduces historical NVBP outcomes, with RL-based agents achieving 74.81% prediction accuracy, outperforming heuristic rule-based baselines by 10.80%.
- We conduct systematic analyses of key market parameters to enable counterfactual reasoning, providing quantitative evidence to support informed policy adjustments.

2 The Structure of ProcureGym

NVBP is an N-price sealed-bid auction with explicit volume constraints. The procurement rule follows a deterministic lowest-price selection mechanism: the Top-x lowest bidders secure guaranteed volumes (ρQ_0), while non-winners access residual market demand ($Q_e - \rho Q_0$) through price-linkage policies. Due to environment uncertainty, firms optimize the trade-off between winning probability and corporate profit. The detailed mathematical derivations of the profit functions under winning (π_0) and non-winning (π_1) scenarios.

ProcureGym formulates the interactions among heterogeneous firm agents—characterized by distinct attributes such as originator-drug or generic-drug status, operational scale, and in-house Active Pharmaceutical Ingredient manufacturing license—as a Markov game. This MDP framework is constructed to align with real-world procurement scenarios, encapsulating policy regulations, firm profiles, drug specifications, and bidding mechanisms. See Table 2 for configurations and Appendix B for mathematical details. The modular MDP design readily extends to other game-theoretic settings by reconfiguring its components.

3 Dataset Description

To construct realistic NVBP simulation scenarios, this study aggregated real-world data from multiple authoritative sources: procurement documents were obtained from the National Healthcare Security Administration (NHSA), competitor information from the Center for Drug Evaluation (CDE), and data on enterprise type from the China National Pharmaceutical Industry Information Center

MDP Element	Notation	Formula / Definition	Meaning
State Space	S_t	$S_t = \{P_{max}, \rho, x, \omega_i, Q_0, Q_e, C_i, P^i_{t-1}, \Pi^i_{t-1}, t/T\}$	A 10-dimensional vector encoding: policy parameters, firm parameters, market history, and time encoding.
Action Space	A_t	$P^i_t = C_i + \frac{a_t+1}{2} \cdot (P_{max} - C_i)$	Normalized decision $a_t \in [-1, 1]$ mapped to bidding price $P^i_t \in [C_i, P_{max}]$.
Transition Probability	$P(s'\|s, a)$	$s^i_{t+1} = \left(P^i_t,\ \Pi^i_t(I^i_t)\right)$, where $I^i_t = \mathbb{I}(\text{rank}(P^i_t) \leq x)$	A binary indicator I^i_t (1 for winning) determines the realized profit Π^i_t, thereby updating the historical state (P^i_t, Π^i_t).
Reward Function	R_t	$R_t = I^i_t \cdot \pi_0 + (1 - I^i_t) \cdot \pi_1$	Profit conditioned on the winning status I^i_t, comprising the procurement profit π_0 (winning) and the linkage profit π_1 (non-winning).
Discount Factor	γ	$\gamma = 0.99$	Weighting factor for future rewards in the cumulative return.

Table 2: MDP Elements for Firm Agents in ProcureGym. This table summarizes the **State Space** S_t, encompassing maximum valid bidding price P_{max}, agreed procurement ratio ρ, number of winning bidders x, firm-specific price linkage coefficient ω_i, agreed procurement volume Q_0, actual procurement volume Q_e, unit production cost C_i, previous bidding price P^i_{t-1}, previous profit Π^i_{t-1}, and time information t/T; **Action Space** A_t, mapping the normalized bidding decision a_t to the actual bidding price P^i_t; **Transition Probability** $P(s'|s, a)$, governed by the binary indicator of winning status I^i_t, winning profit π_0, non-winning profit π_1.

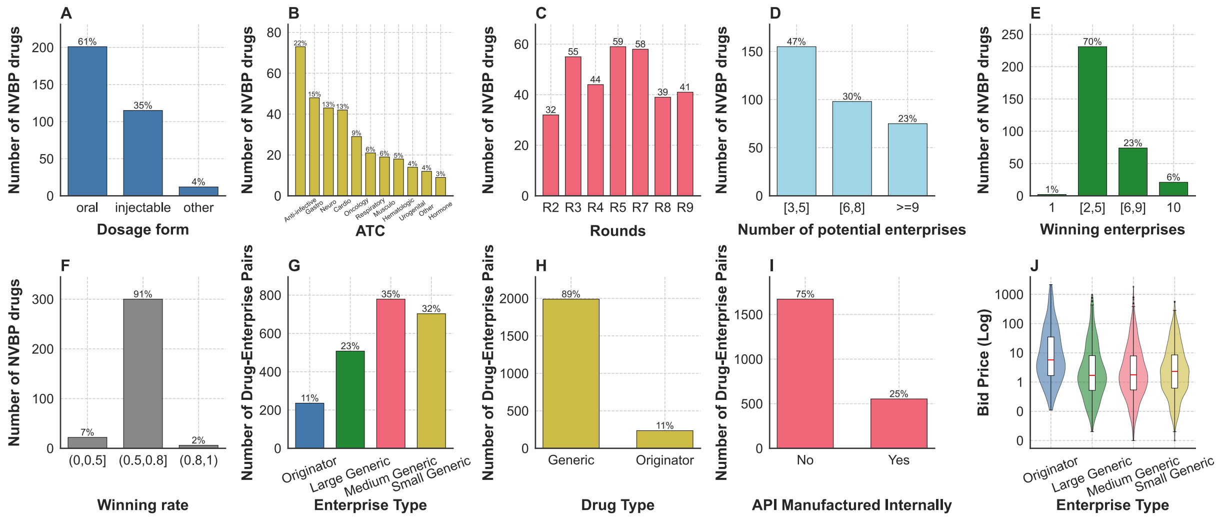

Figure 2: Summary of the research dataset. (A-C) Drug characteristics: dosage forms, Anatomical Therapeutic Chemical (ATC) categories, and drugs by procurement round. (D-F) Competition: potential bidders, winners per drug, and winning rates. (G–J) Enterprise attributes:enterprise type, originator versus generic status, in-house active pharmaceutical ingredient production, and the distribution of log-transformed bid prices by enterprise type.

(PHIIC). The dataset comprises 328 NVBP drugs, predominantly oral formulations (61%) and anti-infectives (23%) (Figure 2A,2B), with the highest number of included drugs in Round 5 and the lowest in Round 2 (Figure 2C). Regarding competition, 48% of drugs faced 3–5 potential bidders, resulting in 2–5 winning firms for the majority (70%) of products and a winning rate concentrated in the 50–80% range (Figure 2D-2F). At the firm level, the study involves 2,226 drug–firm pairs, dominated by generics (89%) (Figure 2H)—especially small and medium-sized enterprises (67%) (Figure 2G)—with only 25% possessing in-house active pharmaceutical ingredient capabilities (Figure 2I) and distinct bid price distributions across enterprise types (Figure 2J).

4 Experiment

4.1 Settings

Modeling the NVBP scenario within an MDP framework, we incorporate three heterogeneous agent types: (1) **RL-based** agents comprise IPPO and MAPPO; (2) **LLM-based** agents are powered by the *Qwen3-235B-A22B-Instruct* model and employ a cognitive architecture characterized by *Perception-Memory-Decision-Reflection*; (3) **Rule-based** agents employ heuristic strategies formulated based on real-world government regulations and firm attributes. See Appendix C for details.

The experiments are conducted in a single-round setting. Evaluation metrics span three dimensions: **(1) Price Prediction Accuracy**: spearman correlation and coefficient of determination (R^2) between predicted and actual prices; **(2) Selection Prediction Accuracy**: alignment rates between predicted and actual winners under Top-K selection; **(3) Firm Profit**: profit distribution analysis validating learned bidding strategies.

4.2 Overall experiment results

Figure 3A illustrates the log-scaled actual vs. predicted bidding prices for four algorithms, all of which exhibit strong positive Spearman correlations ($\rho = 0.85$–0.88, all $p < 0.001$). The R^2 range from 0.76 to 0.79, with MAPPO demonstrating the highest explanatory power ($R^2 = 0.79$) and the Rule-Based method the lowest ($R^2 = 0.76$). Figure 3B presents RL algorithms achieve substantially higher accuracy rates (both 75%) compared to the LLM (66%) and Rule-Based (64%) methods. Finally, Figure 3C demonstrates that the RL algorithms learn profit-maximizing bidding strategies. Although real-world bids are not always optimal, RL maintains high predictive accuracy for selection and refines strategies to yield higher profits, underscoring practical advantages in strategy enhancement rather than mere replication.

4.3 Further analysis

Further analysis show clear effects of both policy and market factors. A higher procurement ratio and larger contractual volume suppress bidding prices and reduce profits (Figure 4A, 4B; Figure 4E, 4F), whereas a higher maximum valid bidding price and stronger market demand raise both bids and prof-

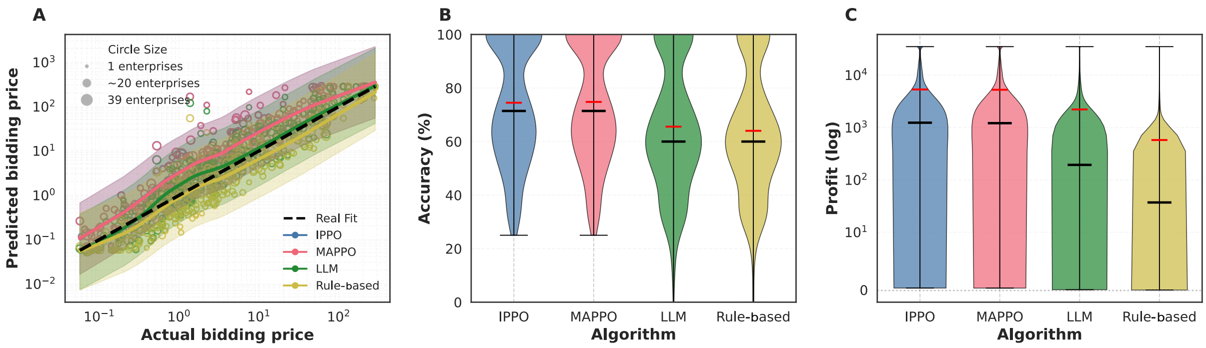

Figure 3: Evaluation of NVBP Simulation across 7 Rounds (Rounds 2-9, excluding Round 6(insulin-focused)). (A) Price Prediction Accuracy: Log-log scatter plot of predicted vs. actual bid prices; bubble size = number of firms per drug. Lowess smoothing curves with 95% confidence bands visualize trends; the black diagonal line ($y = x$) represents perfect prediction. (B) Selection Prediction Accuracy: Batch-wise winner alignment rate; Top-K lowest-price ranking predictions vs. actual outcomes. (C) Firm Profit: Log-scale profit distribution.

its (Figure 4C, 4D; Figure 4G, 4H). Demand is the most influential market driver of profitability, followed by contractual volume and then production costs; rising costs increase bids only modestly and compress margins, indicating limited cost pass-through (Figure 4I, 4J). RL methods maintain relatively stable bid-to-ceiling ratios and consistently outperform LLM and rule-based baselines across all settings, with profitability more sensitive to the ceiling price than to procurement-ratio changes.

4.4 Unveiling the logic of bidding strategies

While RL agents demonstrate superior performance, LLM agents enhance the interpretability of the platform by providing natural language rationales that explicate the underlying strategic logic and intermediate decision-making steps. Agents exhibit distinct pricing dynamics stemming from cost variance. As for Batch 2, low-cost firms prioritized market share through aggressive low pricing (59.8% of bids < 30.0% P_{max}), whereas high-cost firms pursued higher profitability, maintaining higher average prices (56.7% vs. 36.2% P_{max}) and profit margins (60.6% vs. 34.0%). For detailed examples of LLM responses, please see Appendix D.

5 Conclusions

We introduce ProcureGym to simulate NVBP. RL outperforms LLM and rule-based algorithms in both profitability and prediction. Analysis identifies maximum bidding price and demand as primary policy and market drivers, respectively. These insights offer actionable strategies for navigating diverse policy and market landscapes.

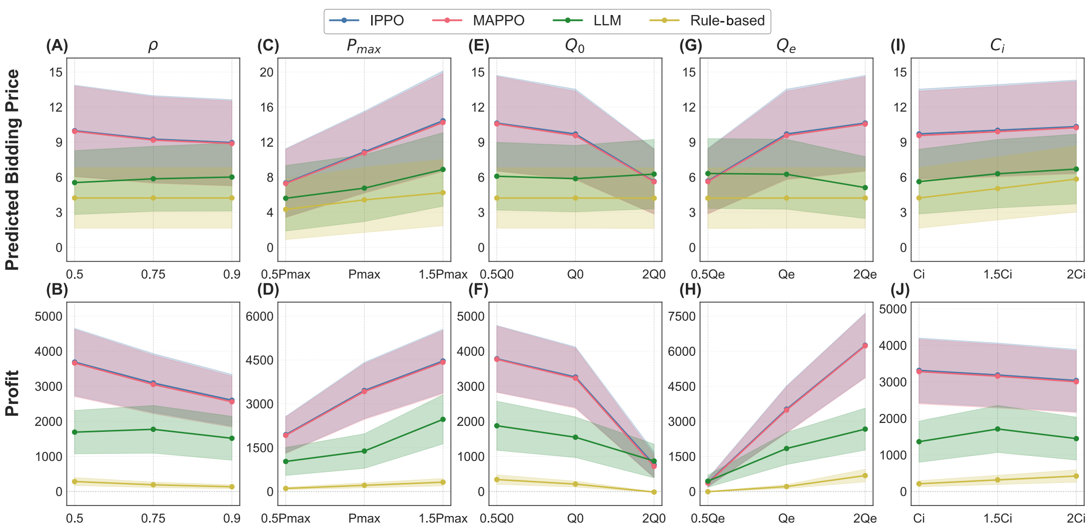

Figure 4: Further Analysis of Procurement Policies with respect to Variations in Key Parameters. Predicted bidding price and firm profit under varying: (A,B) argeed procurement ratios(ρ); (C,D) maximum valid bidding prices(P_{max}); (E,F) agreed procurement volume(Q_0); (G,H) actual procurement volume(Q_e); (I,J) unit production costs(C_i). Colored lines: four algorithms; shaded areas: 95% confidence intervals.

Limitations

While ProcureGym provides a simulation scenario for drug centralized procurement bidding, and the agents can learn optimal bidding strategies that maximize profits, this study still has several limitations. **First,** the centralized procurement scenario is designed based on China's NVBP. Although its logic aligns with that of most drug centralized procurement initiatives, caution should be exercised when extrapolating the findings to other drug procurement scenarios. **Second**, the model only simulates the bidding behavior of firms. As a public policy, however, the aforementioned complex system should also incorporate governments and medical institutions—this would enable the provision of insights for policy optimization.

Ethical considerations

This study does not involve human participants or animals. All data used in this research were used with permission from the data custodians for academic purposes, and the data sources comply with ethics review requirements. The data contain no personally identifiable information and were obtained from public sources.

References

Simone Brusatin, Tommaso Padoan, Andrea Coletta, Domenico Delli Gatti, and Aldo Glielmo. 2024. Simulating the economic impact of rationality through reinforcement learning and agent-based modelling. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 159–167. ACM.

Shengmao Cao, Lisa Xuejie Yi, and Chuan Yu. 2024. Competitive bidding in drug procurement: Evidence from china. *American Economic Journal: Economic Policy*, 16(3):481–513.

Michael Curry, Alexander Trott, Soham Phade, Yu Bai, and Stephan Zheng. 2022. Analyzing micro-founded general equilibrium models with many agents using deep reinforcement learning. *Preprint*, arXiv:2201.01163.

Kshama Dwarakanath, Tucker Balch, and Svitlana Vyetrenko. 2025. Abides-economist: Agent-based simulator of economic systems with learning agents. *Preprint*, arXiv:2402.09563.

Lars Holger Ehlers, Morten Berg Jensen, and Henrik Schack. 2022. Competitive tenders on analogue hospital pharmaceuticals in denmark 2017–2020. *Journal of Pharmaceutical Policy and Practice*, 15.

Natalia Ferreira and Carina Pizzinat. 2025. Framework agreements in uruguay: towards an optimal acquisition strategy in public procurement. Documentos de Trabajo (working papers) 0125, Department of Economics - dECON.

WHO Regional Office for the Western Pacific. 2002. Practical guidelines on pharmaceutical procurement for countries with small procurement agencies.

Warren Kaplan, Veronika Wirtz, Sabine Vogler, Aurélia Nguyen, and Richard Laing. 2016. Policy options for promoting the use of generic medicines in low-and middle-income countries. *Health Action Int*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. 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)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Jiaming Liu. 2025. The study on the supply early warning and guarantee mechanism of generic drugs based on immunity theory and multi-agent modeling. Master's thesis, Nanjing University of Chinese Medicine.

Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang, Shenghao Zhu, and Jun Wang. 2024. Taxai: A dynamic economic simulator and benchmark for multi-agent reinforcement learning. *Preprint*, arXiv:2309.16307.

Qirui Mi, Qipeng Yang, Zijun Fan, Wentian Fan, Heyang Ma, Chengdong Ma, Siyu Xia, Bo An, Jun Wang, and Haifeng Zhang. 2025. Econgym: A scalable ai testbed with diverse economic tasks. *Preprint*, arXiv:2506.12110.

Jing 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. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *ArXiv*, abs/2502.08691.

Koen Ponse, Aske Plaat, Niki van Stein, and Thomas M. Moerland. 2025. Econojax: A fast scalable economic simulation in jax. *Preprint*, arXiv:2410.22165.

Xinhuanet. 2023. Over 400 billion rmb saved: A closer look at the impact of china's drug procurement program. https://www.gov.cn/yaowen/2023-04/07/content_5750442.htm.

Ying Yang, Runhu Hu, Xin Geng, Lining Mao, Xiaotong Wen, Zhaolun Wang, Siyu Hao, Dan Cui, and Zongfu Mao. 2022. The impact of national centralised drug procurement policy on the use of policy-related original and generic drugs in china. *The International Journal of Health Planning and Management*, 37(3):1650–1662.

Jing Yuan, Z. Kevin Lu, Xiaomo Xiong, and Bin Jiang. 2021. Lowering drug prices and enhancing pharmaceutical affordability: an analysis of the national volume-based procurement (nvbp) effect in china. *BMJ Global Health*, 6.

Boya Zhao, Jing Wu, and Xing Lin Feng. 2024. Testing the unintended cost effects of health policies for generic substitutions: the case of china's national volume-based procurement (nvbp) policy. *Health Policy and Planning*, 40:194 – 205.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 2021. The ai economist: Optimal economic policy design via two-level deep reinforcement learning. *Preprint*, arXiv:2108.02755.

Zheng Zhu, Jiawei Zhang, Chao Gong, and Li Yang. 2025. Impacts of china's national volume-based procurement policy on the pharmaceutical industry: A systematic review. *Pharmacoeconomics and Policy*, 1(2):63–72.

Zheng Hua Zhu, Quan Wang, Qiang Sun, Joel R Lexchin, and Li Yang. 2023. Improving access to medicines and beyond: the national volume-based procurement policy in china. *BMJ Global Health*, 8.

A Related Works

A.1 Economic Simulation Modeling

Economic simulation modeling has evolved significantly, progressing from rule-based ABM to Multi-agent reinforcement learning (MARL) grounded in MDP, and more recently, to agents driven by LLM. Early ABM primarily relied on predefined heuristic rules to simulate interactions among heterogeneous entities. Adopting a "bottom-up" social modeling paradigm, these models defined individual-level behavioral rules to observe the emergence of macro-level social phenomena. However, the behavioral rules were typically handcrafted by researchers, resulting in a lack of learnable strategy optimization and an inability to adapt to drastic environmental variations. To incorporate rational decision-making, researchers began formalizing economic problems as Markov processes, enabling behavioral strategies to be learned via data-driven approaches rather than being manually prescribed. MARL frameworks, such as AI-Economist (Zheng et al., 2021) and TaxAI (Mi et al., 2024), allow economic agents to continuously learn and optimize strategies within dynamic games. These studies have demonstrated the superiority of RL over traditional frameworks (e.g., the Saez tax framework) in tasks such as tax policy formulation.

With the rise of generative AI, the simulation paradigm is shifting towards LLM-based agents. This approach employs "Homo Silicus" with the aim of complementing or substituting human-subject experiments. Compared to Rule-based and RL agents, LLM-based agents exhibit superior semantic understanding and human-like reasoning capabilities. For instance, EconAgent endows LLM-based agents with "perception-memory-decision" modules (Li et al., 2024). Experiments have demonstrated that these agents facilitate the emergence of macroeconomic laws, such as the Phillips curve, within ABM environments with greater accuracy than traditional models. Building upon these works, EconGym offers a scalable, theory-grounded testbed featuring four heterogeneous economic roles and over 25 cross-domain tasks (Mi et al., 2025). It supports RL, LLM, Rule-based, and hybrid agent modeling, with research demonstrating that integrating multi-paradigm agents often enhances system performance and robustness in complex coordination scenarios.

A.2 National Volume-Based drug Procurement Simulations

Centralized drug procurement serves as a pivotal policy instrument globally for containing healthcare costs and enhancing medication accessibility (Kaplan et al., 2016; for the Western Pacific, 2002). China's NVBP employs a "volume-based procurement" mechanism, mandating that firms secure market shares in specific regions through public bidding. Implemented to address long-standing issues of artificially inflated drug prices and redundancies in circulation channels, this policy has fundamentally reshaped the game-theoretic dynamics of the pharmaceutical supply chain (Zhu et al., 2023; Yuan et al., 2021).

Previous studies primarily used econometric methods, such as difference-in-differences and interrupted time series analysis, to evaluate the actual effects of the policy (Yang et al., 2022; Ehlers et al., 2022; Zhao et al., 2024). Research modeling centralized procurement scenarios has evolved from static game analysis to dynamic system simulation. Early studies primarily utilized static game theory and reverse auction models to assess the impact of centralized procurement on social welfare, drug prices, and government expenditure (Cao et al., 2024; Ferreira and Pizzinat, 2025). In these studies, volume-based procurement is often abstracted as a multi-item sealed-bid or multi-unit reverse auction problem. Researchers derive the optimal bidding prices and equilibrium solutions for firms under complete or incomplete information within Bertrand competition or Stackelberg game frameworks. Subsequently, Liu's study has used ABM to explain supply shortages after the implementation of centralized procurement (Liu, 2025), viewing centralized procurement and drug supply as complex, evolving systems over time.

In short, existing research primarily focuses on econometric analysis and theoretical game models, with limited exploration of drug pricing simulations under centralized procurement policies. While RL and LLM have shown strong simulation capabilities in broader economic contexts, microeconomic competitive scenarios, particularly in national drug procurement, remain underexplored. This paper fills this gap by designing a multi-agent MDP procurement environment aligned with the NVBP mechanism, systematically comparing Rule-based baselines, multi-agent RL algorithms, and LLM agents in simulating the bidding process.

B MDP Modeling

State Space (S_t): We define the state space S_t for the firm i at time t as a 10-dimensional vector:

$$S_t = \left\{ P_{max}, \rho, x, \omega_i, P_{t-1}^i, \Pi_{t-1}^i, Q_0, Q_e, C_i, \frac{t}{T} \right\} \tag{1}$$

The state space is decomposed into four distinct categories: (1) **Policy parameters** $\Theta_t = (P_{max}, \rho, x, \omega_i, Q_0, Q_e)$, encompassing the maximum valid bidding price, agreed procurement ratio, number of winning bidders, firm-specific price linkage coefficient, and the agreed and actual procurement volume; (2) **Historical information** (P_{t-1}^i, Π_{t-1}^i), comprising the firm's previous bidding price and profit; (3) **Firm parameters** (C_i), representing the firm's production cost; (4) **Time information** (t/T), representing the time.

Action Space (A_t): The action space is defined as a bounded continuous scalar $a_t \in [-1, 1]$, representing a normalized bidding decision that is subsequently transformed to the actual bidding price through an mapping:

$$P_t^i = C_i + \frac{a_t + 1}{2} \cdot (P_{max} - C_i) \tag{2}$$

where bidding prices are strictly constrained to the economically feasible range $[C_i, P_{max}]$, with C_i denoting the unit production cost and P_{max} the maximum valid bidding price.

Transition Function ($P(s_{t+1}|s_t, a_t)$): State transition dynamics are governed by a deterministic rank-based selection mechanism that directly mirrors the actual NVBP procurement rules. Given the bidding prices submitted by all firms at time t, the allocation indicator for firm i is defined as:

$$I_t^i = \begin{cases} 1, & \text{if } \mathrm{rank}(P_t^i) \leq x, \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where rank(P_t^i) denotes the ascending price rank of firm i among all N competing firms, and x is the number of winning slots specified by the procurement policy. The allocation indicator I_t^i directly determines the profit structure for each firm. For winning firms ($I_t^i = 1$), the profit π_0 comprises both procurement and price-linkage components:

$$\begin{aligned} \pi_0 = (P_t^i - C_i)\frac{\rho}{x}Q_0 + \\ (P_t^i(1 + \omega_i) - C_i)(Q_e - \rho Q_0)\beta_i \end{aligned} \tag{4}$$

For non-winning firms ($I_t^i = 0$), the profit π_1 consists solely of the price-linkage component:

$$\pi_1 = (P_t^i(1 + \omega_i) - C_i)(Q_e - \rho Q_0)\beta_i \tag{5}$$

The instantaneous profit can thus be expressed as:

$$\Pi_t^i = I_t^i \cdot \pi_0 + (1 - I_t^i) \cdot \pi_1 \tag{6}$$

The state transition updates historical information (P_{t-1}^i, Π_{t-1}^i) based on the realized profits, while policy parameters Θ_t and market parameters (Q_0, Q_e) remain fixed throughout the episode.

Reward Function (R_t): The reward is defined as the instantaneous profit:

$$R_t^i = \Pi_t^i \tag{7}$$

This formulation creates a strategic trade-off: lowering bid prices secures winning status ($I_t^i = 1$) and access to the higher profit π_0, but reduces per-unit margins; while raising prices improves margins but risks exclusion from procurement profits, limiting returns to π_1.

Discount Factor (γ): The discount factor is set to 0.99.

C Algorithm Implementation

IPPO. In the IPPO framework, each firm operates an independent Actor-Critic network equipped with shared feature extraction layers (comprising a 2-layer MLP with 128 hidden units and tanh activation). The Actor generates a Gaussian policy with a learnable mean and standard deviation, while the Critic estimates state values. Critical hyperparameters are configured as follows: learning rate 5×10^{-5}, discount factor $\gamma = 0.99$, Generalized Advantage Estimation (GAE) parameter $\lambda = 0.95$, and clip ratio $\epsilon = 0.2$. To ensure stable training, we implement entropy coefficient annealing (declining from 0.005 to 0.001) and employ KL-based early stopping ($\delta_{KL} = 0.01$).

MAPPO. Adhering to CTDE paradigm, MAPPO incorporates decentralized actors paired with a centralized critic. Each actor exclusively observes its local state S_t^i, while the critic accesses global state information (constructed via the concatenation of all agents' states). The centralized critic provides per-agent value estimates, enabling coordinated learning while maintaining decentralized execution. Training parameters mirror those of IPPO, including $\lambda = 0.95$, $\epsilon = 0.2$, and the activation of value function clipping. To ensure stable training, we implement entropy coefficient annealing (declining from 0.005 to 0.001) and employ KL-based early stopping ($\delta_{KL} = 0.01$).

Rule-based. This heuristic baseline utilizes a target profit margin strategy with dynamic adjustments. Base profit margins are assigned based on firm type: 20% for originators (Type A), 14% for medium-scale generics (Type B), and 8.6% for small-scale generics (Types C/D). The pricing adjustments account for four factors: (1) selection probability feedback (±12% based on historical win rates); (2) competitive positioning (applying a 2% discount relative to the mean opponent price); (3) temporal discounting (0.75–1.0× in later rounds); and (4) loss aversion (imposing an additional 10% reduction when facing losses).

LLM. Building upon the EconAgent framework (Li et al., 2024), LLM agents leverage structured prompts encompassing: (1) a role definition designating the agent as a firm pricing strategist; (2) firm characteristics (such as cost and market position); (3) NVBP mechanism parameters (maximum valid bidding price P_{max}, agreed procurement ratio ρ, agreed procurement volume Q_0, and number of winning bidders x); (4) a 3-round decision memory annotated with profit change explanations; and (5) the current market state, including price rankings and selection status. The agent is instantiated using *Qwen3-235B-A22B-Thinking-2507-FP8* with a temperature of 0.7 and a maximum token limit of 512, supporting both API and LLM inference modes (vLLM). Furthermore, a periodic strategy reflection mechanism (triggered every 5 steps) facilitates adaptive learning from historical outcomes.

The prompt architecture follows a *Perception-Memory-Decision-Reflection* cognitive framework: the perception module parses the 10-dimensional state vector into natural language market descriptions; the memory module maintains a sliding window storing the most recent 3 decision-outcome tuples (the bid price, profit and price rank of firm i in round t); the reflection module triggers every 5 steps to analyze cumulative performance metrics and prompt strategic reconsideration. The LLM output is constrained to structured JSON format {"reasoning": "<text>", "bid_price": <float>}. We present an illustrative example of the system prompt and the corresponding model responses in Figure 5. Note: For the purpose of anonymity, the specific enterprise name in the provided exemplar has been redacted and substituted with the placeholder "xxx".

All experiments conducted in this paper were run on 1-4 H200 GPUs.

D Cases of LLM Responses from Heterogeneous Firms

Through a detailed analysis of the LLM-generated responses, we observed that LLM-based agents display heterogeneous bidding behaviors driven by firm-specific costs during NVBP simulations. Please refer to Figures 6 and 7 for specific instances of outputs from low-cost and high-cost firms.

System Prompt You are the Pricing Strategist for xxx, participating in the National Centralized Volume-Based Procurement bidding for the drug 'Adefovir Dipivoxil Tablets'. Enterprise Characteristics: Medium cost, leading position.

[Procurement Mechanism]
- Ceiling Price (Pmax): 1.08 CNY
- Agreed Procurement Ratio (rho): 60%
- Agreed Procurement Volume (Q0): 2893 units
- Number of Winning Bidders (x): 2

[Review of Historical Decisions]
- Round 2: Bid 0.23, Profit 76.7319, Rank 1, Won 100%
- Round 3: Bid 0.23, Profit 115.6812, Rank 1, Won 100% (Profit Increased)
- Round 4: Bid 0.24, Profit 115.6812, Rank 1, Won 100%

[Strategic Reflection]
Analysis of performance in the past 3 rounds:
- Average Profit: 102.6981, Average Win Probability: 100%
- Price Trend: Rising (0.23 -> 0.24)
- Reflection: Is the current strategy effective? Does the bid need adjustment to increase profit or win probability?

[Profit Variation Analysis]
Profits are essentially flat; the current strategy is stable.

[Current Status] (Round 6, Progress 10%)
- Previous Round Bid: 0.2450 CNY
- Previous Round Profit: 115.68 CNY
- Relative to Market Average Price: -73.1%
- Price Rank: 1st/3 companies
- Selection Status: Selected
- Price Trend: Rising

[Private Information]
- Cost (C_i): 0.1886 CNY
- Market Share (beta_i): 26.50%

[Bid Range]
[0.1886, 1.0800]

Response
"reasoning": "The current bid of 0.2450 CNY remains within the safe margin, exceeding the cost of 0.1886 CNY, and historical data indicates this price level ensures winning while maintaining high profitability. With a 1st place price ranking and low competitive pressure, a minor price increase aids in expanding profit margins. Considering the leading market position and the 60% guaranteed procurement volume, aggressive price reduction is unnecessary. Maintaining a prudent strategy, the bid is slightly raised to enhance profitability.",
"bid_price": 0.25

Figure 5: Example of the LLM System Prompt and Response

CASE 1

Scenario Information
Pharmaceutical Product: Simvastatin Tablets
Firm Name: Firm A.
Bidding Range [Ci, Pmax]: [0.0167, 0.2600]
Bidding Headroom (Pmax-Ci): 0.2433
Bidding Price: 0.0167
Response
"reasoning": "The current quotation of CNY 0.0168 closely approximates the marginal cost of CNY 0.0167, indicating negligible scope for further price reduction. As a cost leader, the enterprise aims to secure its selection status and displace high-cost competitors by maintaining a marginal profit level. Given its top-tier price ranking and stable selection prospects, the strategy focuses on anchoring the bid to the cost floor to defend its market position.",
"bid_price": 0.0167

CASE 2

Scenario Information
Pharmaceutical Product: Amoxicillin Capsules
Firm Name: Firm B.
Bidding Range [Ci, Pmax]: [0.0149, 0.1000]
Bidding Headroom (Pmax-Ci): 0.0851
Bidding Price: 0.0150
Response
"reasoning": " The current bid of CNY 0.0150 remains slightly above the cost threshold of CNY 0.0149. Currently ranked 3rd, the enterprise possesses a distinct competitive advantage. In light of improving market conditions and anticipated margin recovery, the rationale is to maintain a price point near the cost floor. This ensures both bid eligibility and a modest profit surplus, thereby stabilizing long-term returns.",
"bid_price": 0.0150

CASE 3

Scenario Information
Pharmaceutical Product: Isoniazid Tablets
Firm Name: Firm C.
Bidding Range [Ci, Pmax]: [0.0300, 0.2300]
Bidding Headroom (Pmax-Ci): 0.2000
Bidding Price: 0.0450
Response
"reasoning": "The previous quotation of CNY 0.0350 fell below the cost line, resulting in unsustainable margins and an adverse profit trend. As a medium-cost enterprise, the priority is to restore financial viability while ensuring bid selection. Benchmarked against the price ceiling of CNY 0.2300 and the prevailing competitive landscape, a strategic upward adjustment to CNY 0.0450 is proposed. This "price repair" strategy aims to balance selection probability with margin optimization.
",
"bid_price": 0.0450

CASE 4

Scenario Information
Pharmaceutical Product: Indapamide Tablets
Firm Name: Firm D.
Bidding Range [Ci, Pmax]: [0.0118, 0.1100]
Bidding Headroom (Pmax-Ci): 0.0982
Bidding Price: 0.0250
Response
"reasoning": "The enterprise failed to secure a bid in the previous round with a quote of CNY 0.0275 (ranked 7th). Given a marginal cost of CNY 0.0118, there is significant latitude for downward price adjustment. To enter the winning bracket (top 5 firms), the bid is revised to CNY 0.0250. This tactical reduction, informed by historical winning benchmarks around CNY 0.03, is designed to enhance the probability of selection in a competitive environment.",
"bid_price": 0.0250

Figure 6: Cases of LLM Responses from Low-cost Firms

CASE 1

Scenario Information
Pharmaceutical Product: Paclitaxel for Injection (Albumin-bound)
Firm Name: Firm E.
Bidding Range [Ci, Pmax]: [221.7060, 2400.0000]
Bidding Headroom (Pmax-Ci): 2178.2940
Bidding Price: 2320.0000
Response
"reasoning": "The incumbent quotation of 2,350.00 CNY is positioned at the higher decile of the pricing spectrum; although it exceeds the market mean, it remains within the permissible threshold. Given the firm's significant cost-leadership advantage (unit cost at 221.71 CNY) and its current tier-three winning status, a marginal downward adjustment is recommended. This strategy aims to solidify market share and optimize profit margins. By balancing competitive intensity with selection stability, a moderate price reduction will enhance the firm's strategic positioning.",
"bid_price": 2320.0

CASE 2

Scenario Information
Pharmaceutical Product: Abiraterone Acetate Tablets
Firm Name: Firm F.
Bidding Range [Ci, Pmax]: [9.0348, 101.6700]
Bidding Headroom (Pmax-Ci): 92.6352
Bidding Price: 53.6000
Response
"reasoning": "As a dominant cost-leader, the firm has achieved successive price increments with its current quotation of 52.80 CNY while maintaining the top ranking, leading to a steady appreciation in profitability. Despite substantial pricing elasticity afforded by the low marginal cost (9.03 CNY), a strategic marginal markup is advised. Considering market price sensitivity and the prevailing competitive landscape, this adjustment will further bolster net margins without compromising the probability of bid success. With the current fulfillment rate at 58%, the objective is to optimize profitability while sustaining a superior competitive lead.",
"bid_price": 53.6

CASE 3

Scenario Information
Pharmaceutical Product: Ambrisentan Tablets
Firm Name: Firm G.
Bidding Range [Ci, Pmax]: [12.7724, 80.0000]
Bidding Headroom (Pmax-Ci): 67.2276
Bidding Price: 72.0000
Response
"reasoning": "The current quotation of 76.80 CNY resulted in a fourth-place ranking, falling outside the selection threshold. A price reduction is imperative to enhance competitive viability. Given the marginal cost of 12.7724 CNY, there remains sufficient buffer for tactical discounting. To secure a winning position within the top two rankings, a moderate price concession is proposed. This approach seeks to capture market share while simultaneously mitigating excessive margin erosion.",
"bid_price": 72.0

CASE 4

Scenario Information
Pharmaceutical Product: Tadalafil Tablets
Firm Name: Firm H.
Bidding Range [Ci, Pmax]: [17.4441, 90.7500]
Bidding Headroom (Pmax-Ci): 73.3059
Bidding Price: 65.5000
Response
"reasoning": "With a current ranking of fifth at 72.60 CNY, the firm has failed to secure a winning bid, necessitating a downward price adjustment. Characterized as a "medium-cost follower" (cost at 17.4441 CNY), the firm should benchmark its strategy against the pricing trajectories of market leaders. The objective is to converge toward the price ceiling while maintaining a reasonable profit margin to improve the probability of procurement inclusion. This round mandates a strategic price reduction to facilitate an upward shift in competitive ranking.",
"bid_price": 65.5

Figure 7: Cases of LLM Responses from High-cost Firms