

WonderEcon: An Interactive LLM-Agent System for Simulating Economies

Jing Qian^{1,2}, Boyu Liu², Wenyan Gu², Hang Ruan², Cheng Huo³, Junjie Yang²,
Mingyu Deng², Jiale Han^{1,†}, Benyou Wang^{1,2}

¹ Shenzhen Loop Area Institute

² School of Data Science, The Chinese University of Hong Kong, Shenzhen

³ City University of Hong Kong (Dongguan)

Abstract

How an economic shock propagates through the heterogeneous decisions of diverse economic roles and emerges as aggregate fluctuations is a central question in economics. Traditional agent-based models hard-code behavioral rules and cannot read language, while recent LLM-agent simulations often omit explicit belief formation, endogenous price formation, or mechanism-grounded state transitions, and rarely let a human step into the loop. We present **WonderEcon**, a configurable, LLM-driven yet economic-rule-grounded multi-agent simulation system in which households, firms, commercial banks, a central bank, and a government interact across five markets, forming a self-consistent macroeconomic loop. Every agent decides through a four-step natural-language pipeline, reading real news alongside structured data, with every step observable. A player mode, with dedicated scene and character modeling, lets a human take over one household agent, yielding a counterfactual comparison and the Economic Personality Type Indicator (EPTI). On an interest-rate hiking cycle, WonderEcon demonstrates directional plausibility with respect to six well-known economic principles across four LLMs, and 33 participants find the player system satisfying. The engine and interfaces are publicly available.¹

1 Introduction

How economic shocks such as an interest-rate hike pass through the heterogeneous, boundedly rational (Simon, 1955; Kahneman and Tversky, 1979) decisions of private agents and emerge as aggregate dynamics is a classic, still-open problem. LLMs now make it possible to instantiate economic agents that read news, weigh trade-offs (Wei et al., 2022;

Yao et al., 2023), and carry heterogeneous “personalities”. Yet existing systems fall short. Traditional agent-based models (Hamill and Gilbert, 2015) hard-code behavioral rules and have no language layer, so they cannot react to qualitative narratives such as news. Recent LLM-based simulations (Park et al., 2023; Li et al., 2024; Horton et al., 2023; Aher et al., 2023) often omit explicit belief formation, endogenous price formation, or mechanism-grounded state transitions, drive only part of the system with LLMs, and rarely let a human step into the loop.

We introduce **WonderEcon**, an interactive, configurable LLM-multi-agent economic simulator. It serves as both a research tool and a human-in-the-loop counterfactual yielding an EPTI profile (Figure 1). While driven by LLMs, agent behaviors and the environment are strictly grounded by underlying economic rules. Its core engine follows a natural-language **observation-belief-price-action** pipeline characterized by five properties: (i) **Completeness**: five agent types, five markets, and a shared information space. (ii) **Fidelity**: boundedly rational agents leveraging behavioral economics under incomplete information. (iii) **Heterogeneity**: households, firms, and banks spanning multiple tiers, risk types, and roles. (iv) **Closed macroeconomic loop**: conserved money flows that prevent arbitrary value creation. (v) **Flexibility**: configurable parameters/LLMs, human-in-the-loop control, and instantiable scenarios.

Built on this engine, a **configurable research interface** lets users customize macro policies and news, attach compatible LLMs, and monitor detailed environment metrics at every step. Additionally, a **player counterfactual system** allows a human to replay a frozen simulation by replacing a single agent’s decisions. It compares these outcomes against the original LLM and peer averages to generate a four-dimensional economic-personality profile (EPTI). Beyond engagement,

[†]Corresponding author: jialehan@slai.edu.cn.

¹Code, a live demo, and a how-to video: <https://github.com/Planet-300894/WonderEcon> (code under Apache-2.0; assets under CC-BY-4.0 except third-party outdoor-scene models; see Appendix F).

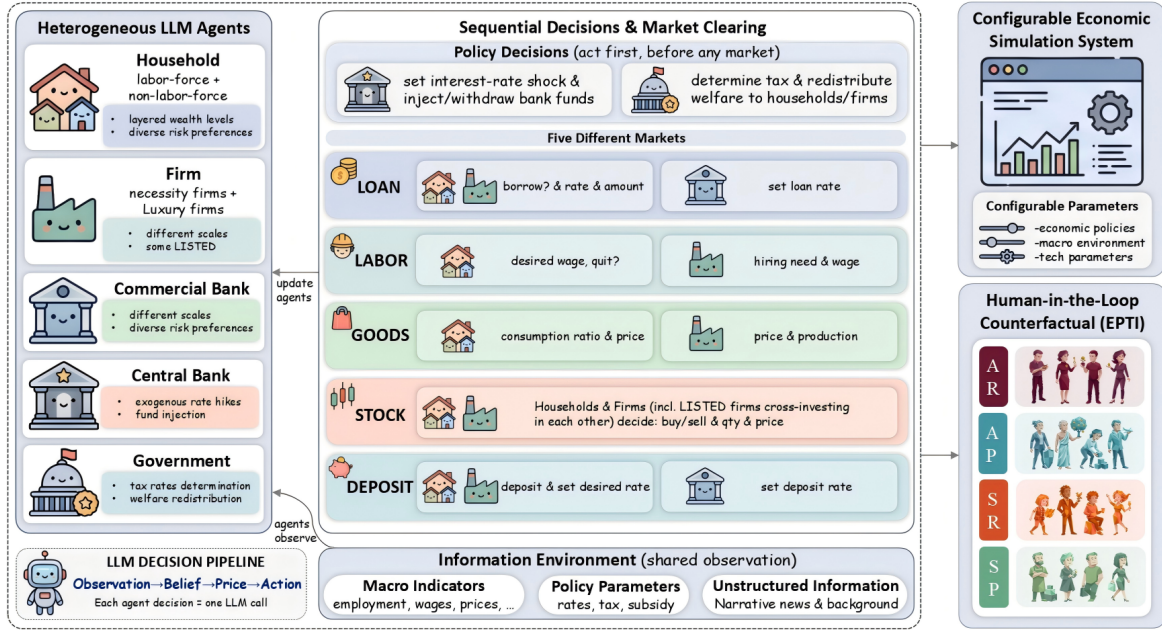


Figure 1: WonderEcon overview. Each agent decides through a four-step observation–belief–price–action pipeline; five markets then clear and settle, closing the macroeconomic loop before the next tick. The same engine drives two applications: a configurable economic simulation system, and a human-in-the-loop counterfactual.

the player mode yields data on players’ economic preferences that can be fed back into agents’ personas to make the whole economy and the EPTI test more faithful, forming a virtuous loop between human players and the simulated economy.

We evaluate WonderEcon from both economic and user perspectives. Across four LLM backends, the simulated economy consistently recovers expected responses to monetary tightening, while a 33-participant study demonstrates strong usability, engagement, and perceived EPTI accuracy. By transforming isolated LLM agents into interactive participants in a living economy, WonderEcon provides a practical foundation for economic experimentation, human–agent comparison, and AI-assisted education.

2 Related Work

LLM-multi-agent systems. A growing body of work orchestrates multiple LLM agents to solve complex tasks through specialization and interaction, spanning human-in-the-loop planning (Kim et al., 2025), collaborative annotation (Qin et al., 2025), multi-agent debate (Becker et al., 2025), and prompt optimization (Xiang et al., 2025). This paradigm is a natural fit for simulating an economy, whose participants are themselves many heterogeneous decision makers that specialize and interact. But an economic simulation must maintain more

than the agents. It also needs a rule environment that clears markets and settles accounts, and an information environment that decides what each agent can observe. This environmental layer is central to WonderEcon.

Economic multi-agent systems. A second line of work places agents inside economic environments. Rule-based and RL simulators provide scalable, calibrated economies but rely on hand-coded rules or reward optimization rather than language (Dwarakanath et al., 2024; Mi et al., 2024; Brusatin et al., 2024). LLM-based simulators instead let agents reason in natural language, reproducing macro regularities in a goods-and-labor economy (Li et al., 2024), trading on a real order-matching market (Gao et al., 2024), analyzing policy across income groups (Hao and Xie, 2025), or replicating expectation surveys (Lin et al., 2025). EconGym (Mi et al., 2025) combines LLM and RL agents in a modular testbed of many roles and 25+ tasks, among the most complete simulators to date. However, existing work shares one or more limitations. (i) Most drive only part of the system with LLMs and lack explicit belief formation, endogenous price formation, or mechanism-grounded state transitions. (ii) Most cover very few markets rather than a full multi-market economy with interlocking financial flows. (iii) Most runs are offline without step-by-step observability, and configuration is ei-

ther unavailable or, too complex to be user-friendly. (iv) No human-in-the-loop mechanism exists for a clean counterfactual or behavioral profiling. WonderEcon closes all four gaps at once with a single configurable and playable engine.

3 System Overview: The LLM Economic Engine

3.1 Components

WonderEcon instantiates a self-contained economy from three kinds of components: heterogeneous agents, a shared information environment, and five markets through which the agents interact. One tick equals one month, and concrete population sizes are given in Section 3.4.

Agents. Following the agent-based tradition in economics (Tesfatsion, 2006) and recent LLM-agent designs (Park et al., 2023; Wang et al., 2024), WonderEcon has five agent types. Household agents decide consumption, stock investment, and deposits, and labor-force households additionally decide borrowing, desired wage, and whether to quit. Firm agents decide borrowing, R&D, wages, production scale, pricing, stock investment, and deposits. Bank agents set deposit and loan rates under a finite lending pool. The central bank decides liquidity injection or withdrawal, while the government decides the tax-redistribution ratio.

Information environment. All agents share a central information environment, the single source of truth for what anyone can observe. It holds two kinds of information. The first is structured economic data, such as employment rate, goods prices, average wages, stock prices, and tax rates. The second is unstructured macroeconomic news, comprising the per-tick economic background and news text. Each decision reads only the minimal subset it needs (Xi et al., 2025), so agents act under local, incomplete information rather than global truth (Stiglitz and Weiss, 1981).

Five market environments. Agents interact through five markets: a labor market, two sector-specific goods markets, two sector-specific stock markets (each a call auction over one listed firm), a loan market, and a deposit market. The clearing mechanisms are detailed in Section 3.3.

3.2 LLM-Driven Decision Making

Letting LLMs make economic decisions allows heterogeneous agents to reason about their circum-

stances and to “consciously” make human-like, context-dependent economic decisions, instead of executing fixed behavioral rules. Because LLM agents interpret qualitative narratives, form explicit beliefs, and justify their choices in natural language, an LLM-driven economy moves beyond curve-fitted reactions toward deliberating agents. We view this as an important step toward economic twins, in which simulated economies align with real-world observations (Han et al., 2026).

In the system, every priced action of agents is made by an LLM through a four-step natural-language pipeline (Figure 2), extending the observe–reason–act paradigm of recent language agents (Park et al., 2023; Yao et al., 2023) to priced economic decisions. (1) **Observation**: the agent reads its own attributes together with a minimal, action-specific subset of the information environment, including the economic-background and macro-news text, so it must interpret qualitative language, not just numbers. (2) **Belief**: a short free-text analysis, in the spirit of chain-of-thought reasoning (Wei et al., 2022), in which the agent reasons about its situation and the macro outlook, rather than the model-consistent forecasts assumed under rational expectations (Muth, 1961). (3) **Price**: a structured desired price, rate, or ratio, such as a desired loan rate or wage. (4) **Action**: a yes/no decision or a budget split, such as whether to borrow money or the consumption ratio.

Every step is a separate prompt that carries the full observation content plus the outputs of all previous steps, so beliefs ground prices and prices ground actions (Figure 2). Natural language here is the operating medium rather than a presentation layer: qualitative macroeconomic information, such as the economic background and news narratives, reaches agents only as unstructured text, so reacting to it requires language understanding rather than table lookup, and each priced decision is extracted from free-text reasoning into a structured value. Persistent self-descriptions like identity and risk preference are injected into every prompt, a prompt-as-persona technique that conditions behavior through language rather than hand-coded rules (Argyle et al., 2023; Serapio-García et al., 2025; Horton et al., 2023). The engine is built on the CAMEL framework (Li et al., 2023), so each agent’s brain is pluggable, allowing different models within one world. Same-type calls within a tick run asynchronously for throughput. An agent that cannot read would remain inert to news-driven

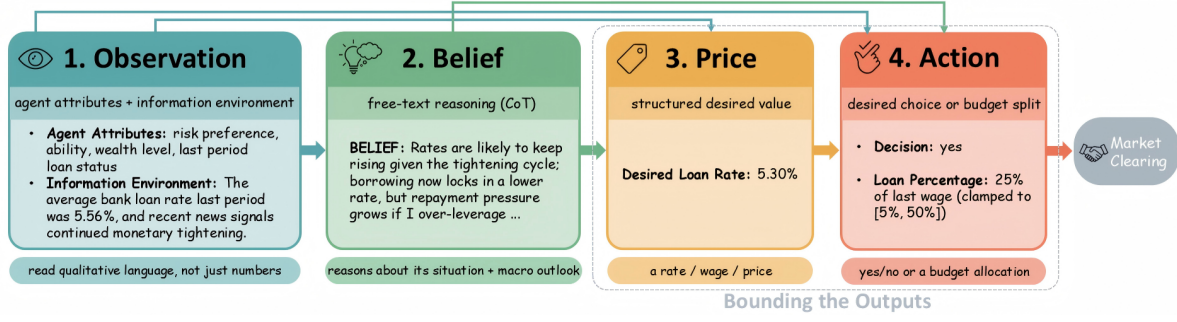


Figure 2: The four-step natural-language decision pipeline. For each priced action, the agent observes its inputs, forms a free-text belief, commits to a desired price, and takes an action. Each step is a separate prompt conditioned on the previous steps’ outputs, so beliefs ground prices and prices ground actions.

shocks: it is the language layer that lets the simulated economy respond to the qualitative narratives that move real economies. Full prompt templates are in Appendix B.

A free LLM can produce extreme or inconsistent numbers that would destabilize a coupled economy (Ross et al., 2024; Bini et al., 2026). We therefore clamp key numeric decisions, such as wage adjustments, pricing markups, and consumption ratios, to economically motivated ranges before they enter the markets. Unlike constrained decoding, which restricts generation token by token, clamping intervenes only after the agent has fully reasoned in natural language, so the textual chain from observation to belief to action is preserved and only the final numeric value is bounded. These bounds keep agent behavior realistic and low-variance, so emergent dynamics reflect the agents’ relative choices rather than occasional outlier generations.

3.3 Market Clearing and World Settlement

Once all agents have decided, the engine aggregates their bids into market outcomes through rule-based mechanisms and writes the results back into the world state, completing one monthly cycle.

Market clearing. Matching uses a logit discrete-choice rule whose attraction function adds a loss-aversion penalty inspired by Kahneman–Tversky prospect theory (Kahneman and Tversky, 1979). Goods and loan markets adopt a “lower-is-better” form in which a chooser i prefers a provider j offering a lower price or rate,

$$\text{Attr}_{ij} = \ln \frac{p_i}{p_j} - (\lambda - 1) \max\left(0, \ln \frac{p_j}{p_i}\right), \quad (1)$$

where p_i is i ’s desired price or rate, p_j is j ’s offer, and $\lambda \geq 1$ is i ’s loss-aversion coefficient penalizing unfavorable offers. Deposit and labor markets

use the dual “higher-is-better” form. The selection probability is $P_{ij} = \text{softmax}(\beta \text{Attr}_{ij})$, with β the agent’s ability score. It doubles as the logit temperature and a time-priority weight, so higher- β agents choose more sharply and are served earlier. The goods market additionally adds a brand bonus to the attraction, making larger firms more attractive, especially to wealthier households,

$$\text{bonus}_{ij} = \alpha \tilde{s}_j (1 + \gamma \tilde{w}_i), \quad (2)$$

where $\tilde{s}_j \in [0, 1]$ is firm j ’s normalized scale tier, $\tilde{w}_i \in [-1, 1]$ is household i ’s wealth tier relative to the median, and $\alpha, \gamma \geq 0$ set the brand strength and its wealth sensitivity. Stocks clear by call auction (Madhavan, 1992) at the volume-maximizing price; goods and labor clear by up to ten rounds of preference-ordered matching; loans are pool-constrained while all deposits are accepted.

Settlement and state update. The cleared trades update holdings, employment, inventory, and bank volumes. The engine then recomputes each agent’s endogenous quantities for the next tick. It settles the circular money flows in a stock-flow-consistent manner (Godley and Lavoie, 2007). Finally it refreshes the aggregate indicators exposed to the information environment, then logs a full state snapshot before the next tick begins.

3.4 Configuration and Execution

The default showcase world runs 156 agents (140 households, 12 firms, 2 commercial banks, a central bank, and a government) for 13 monthly ticks. Through a web wizard (Figure 3), a user can reconfigure the economic policy (tax rates and a clickable interest-rate timeline), the macro environment (the starting economic-background phase, editable per-tick news, and a brand-effect coefficient), and

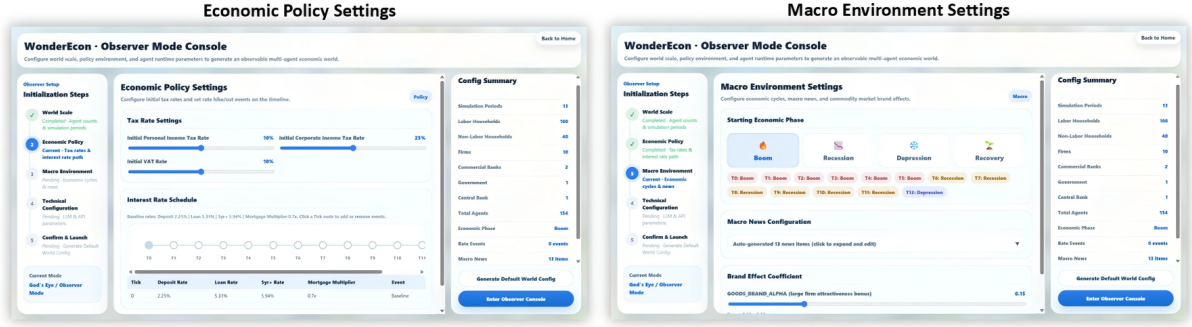


Figure 3: WonderEcon Configuration interface: a wizard sets the scenario.

the technical settings (the agent LLM model, temperature, and concurrency). The simulation horizon and the agent populations are currently fixed to their default values in the web interface, with in-interface scaling planned for a future release. Adjustments to the simulation duration or the scale of the economy should be implemented directly in the open-source backend engine code, while taking the stability of the system economy into account. Since the economic background is configurable, the hiking-cycle scenario is just one instance, and the same engine can instantiate many alternative financial scenarios.

Once the scenario is configured, the engine advances the world through the fixed monthly tick order, detailed as 25 steps in Appendix A, repeating until the horizon is reached. A state tracker then snapshots every agent and the environment to JSON at each tick, and the user enters an interactive, fully replayable economic map. There, for one agent at any tick, the user can watch its full decision process as an animation and inspect the same decisions as JSON. The user can also track how the macro variables, such as GDP and employment rate, evolve over the run.

4 Player Platform and Economic Personality (EPTI)

Whereas the engine runs the world forward under full LLM control (Section 3.4), the player platform lets a human intervene backward (Figure 4). A finished world is frozen and a person takes over a single agent, turning the simulation into a playable, counterfactual experience.

4.1 Player Mechanism

A previously simulated world is frozen into a map. The player picks a tick and a target profile, then is matched to one household agent and inherits

its full state. Across the five engine stations, the player overrides the agent’s decisions on loans, employment, consumption, stocks, and deposits. Considering that letting every agent re-decide live through LLM calls would render each interaction prohibitively slow under limited compute, while the player—being only one of over a hundred agents—exerts negligible influence on the aggregate economy, the other agents’ decisions stay unchanged during the interaction. This freezing yields a controlled counterfactual against the original LLM agent at the cost of suppressing endogenous responses from other agents. After the tick re-clears, the player sees their asset change and a happiness measure, compared against both the original agent and the same-tier average.

Owing to its strong interactivity, this player mode can be readily applied across a broad spectrum of scenarios. In public outreach and financial-literacy education, a user can experience how various economic decisions are made and, through the system’s feedback on happiness and asset changes, learn what consequences their own decisions would bring. In the classroom, an instructor can let students play household roles under a policy shock and contrast their decisions with those of LLM agents, turning abstract monetary transmission into an interactive exercise. For behavioral researchers, the frozen world serves as a standardized and cheap sandbox: every participant faces identical economic conditions, so preferences can be measured and compared across players at scale.

4.2 Economic Personality Type Indicator

Across a play session, WonderEcon maps the player’s decision fields to a four-dimensional, MBTI-style economic-personality code called EPTI, which is an exploratory engagement feature. This follows work that probes personality



Figure 4: Player platform: pick a month, build a character, get matched to one household agent, decide at five stations, then see the three-way comparison and the four-letter EPTI profile.

(Serapio-García et al., 2025; Pan and Zeng, 2023) and economic rationality (Chen et al., 2023) in LLMs, and it is grounded in revealed-preference inference from choices (Samuelson, 1938). Each dimension contrasts two poles, and one representative letter per pole forms a four-letter code. Time preference contrasts Spend (S) with Amass (A), read from the consumption ratio relative to the same-tier mean (Frederick et al., 2002). Risk attitude contrasts Risk-taking (R) with Prudent (P), read from loan intensity, stock intensity, a lower necessity-stock share, or quitting one’s job (Holt and Laury, 2002). Market stance contrasts Contrarian (C) with Trend-following (T), read from whether the player’s desired prices deviate from the recent reference in the same direction as same-tier peers (De Bondt and Thaler, 1985; Jegadeesh and Titman, 1993). Consumption taste contrasts Utilitarian (U) with Fine (F), read from the luxury share of consumption under the dual-sector design. The full set of per-dimension criteria is in Appendix C.

Each play session produces data that reveal the player’s economic preferences, which are costly to collect through conventional surveys or experiments. These data can be fed back into agents’ personas, calibrating heterogeneous households with real human preferences and yielding more realistic agents, hence a more faithful simulated economy; a more realistic economy in turn enables more accurate EPTI analysis, forming a virtuous loop between human players and the simulated economy.

5 Evaluation

5.1 Macro Validity

The rate hikes are exogenous, so rising rates is not a result. Following prior economic simulators that validate against known macro regularities (Li et al., 2024; Dwarakanath et al., 2024), we instead test

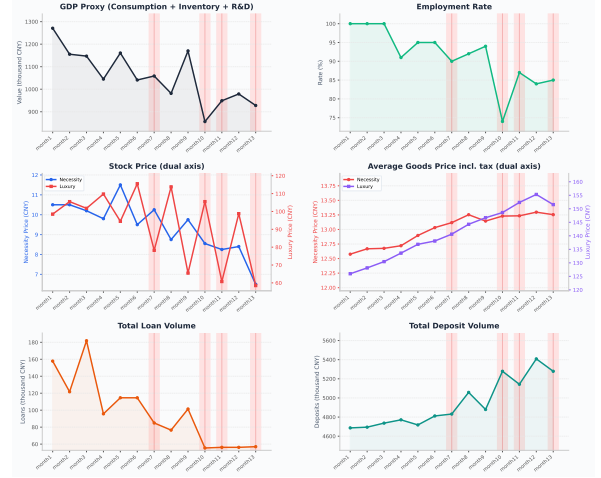


Figure 5: Core macro variables over the 13-tick show-case scenario for a single representative run (GPT-4o, one seed). Shaded bands mark the exogenous rate hikes as event-study windows.

whether the agents’ emergent responses match the stylized facts of monetary tightening. We verify six economic principles under our hiking scenario. Appendix D (Table 4) states them in full, and Figure 5 illustrates the relevant indicators.

We evaluate four pluggable LLM brains (Claude-Sonnet-4.6, GPT-4o, Gemini-3.1-Flash-Lite, Grok-4.1-Fast), each running the scenario ten times. Table 1 reports the number of matching runs out of 10 per principle and the overall score out of 60. All four brains score 56 to 60. The dividend-discount model, Euler substitution, and IS curve hold in all 40 runs, so the emergent asset-price, saving, and output responses are robust. The few deviations fall on Okun’s law and the Phillips curve, because the scenario stays near full employment and the small employment movements these two laws depend on occasionally miss the predicted direction. Overall, the emergent dynamics are directionally consistent with the stylized facts of monetary tightening.

Economic Principle	Claude-Sonnet-4.6	GPT-4o	Gemini-3.1-Flash-Lite	Grok-4.1-Fast
DDM	10/10	10/10	10/10	10/10
Okun’s law	8/10	10/10	10/10	9/10
Credit channel	10/10	10/10	10/10	9/10
Intertemporal substitution (Euler)	10/10	10/10	10/10	10/10
IS curve	10/10	10/10	10/10	10/10
Phillips curve	8/10	9/10	10/10	9/10
Overall	56/60	59/60	60/60	57/60

Table 1: Principle verification rate across models. Each cell shows the number of runs (out of 10) in which the emergent trend matches the theoretical prediction from Table 4.

Category	Measure	Score	Distribution
Usability	Navigation	4.45/5	
	Smoothness	4.36/5	
	Info. access	4.36/5	
Experience	Immersion	4.24/5	
	Usefulness	4.21/5	
	Replay	4.58/5	
EPTI Accuracy	Time preference	4.03/5	
	Risk attitude	4.09/5	
	Market stance	4.21/5	
	Consumption taste	4.21/5	
Overall	Satisfaction	4.70/5	

Table 2: Player-platform user study (5-point Likert; $N = 33$). The distribution column shows each dimension’s rating histogram over the 1–5 scale, with a red line at the mean. Full questionnaire in Appendix E.

These results validate the macroeconomic rationality of the complete hybrid system as a whole, which combines LLM decisions, explicit policy news, numerical clamps, and rule-based market clearing.

5.2 User Study

Following recent interactive-system demonstrations that assess their tools through user-centered questionnaires (Li et al., 2025; Xu et al., 2025; Zhao et al., 2025; Kar et al., 2025), we evaluate the player platform with 33 participants recruited at a technology fair, each completing one play session and receiving the outcome comparison and the EPTI profile. On a 5-point Likert scale they rate three usability dimensions, three experience dimensions, the perceived accuracy of the four EPTI axes, and overall satisfaction. Table 2 summarizes the results. All eleven dimensions score positively, with means between 4.03 and 4.70. Satisfaction is highest at 4.70, followed by willingness to re-

play at 4.58. The usability items sit around 4.4, so participants find the platform easy to use. Notably, all four EPTI axes score above 4 out of 5 (4.03–4.21), indicating that participants largely recognize themselves in the generated profiles and perceive the personality labels as accurate. Among the four axes, time preference scores the lowest (4.03), suggesting that the system still has room to improve in characterizing players’ saving-versus-consumption preferences. This perceived accuracy is encouraging for an exploratory feature, though it reflects self-assessed agreement from a single session rather than formal psychometric validation. We report the small sample honestly as a preliminary study.

6 Conclusion

WonderEcon unifies a complete, grounded, closed-loop, and flexible LLM-driven economic engine with a configurable, observable research interface and a player counterfactual mode with EPTI. Across four LLMs, the engine shows directional plausibility on six economic principles under monetary tightening, and 33 participants rate the player platform positively in a preliminary user study, together suggesting that natural-language decision pipelines can drive a coherent multi-market economy. The same engine supports diverse macroeconomic scenarios, and the interest-rate cycle serves as a foundational showcase. More broadly, we see WonderEcon as an early yet concrete step toward twin economy: its observable belief pipeline, rule-grounded markets, and human-in-the-loop play form the substrate on which more realistic economies can be built. Future work includes component ablations that isolate the contribution of the language pipeline, validation across further scenarios and against empirical data, unlocking fixed configurations, scaling to more agents, automatic scenario generation, live-data ingestion, and a formal psychometric validation of EPTI.

Limitations

Although WonderEcon provides a comprehensive framework, several limitations remain. First, the simulation currently lacks real-time calibration using live data feeds. Second, the language models driving the agents have neither been fine-tuned via reinforcement learning for enhanced economic analysis nor equipped with self-evolution systems. Third, the current environment includes only a limited variety of firms and stock types. Fourth, our macro evaluation covers a single interest-rate-hiking scenario and tests directional agreement with stylized facts; it validates the complete hybrid system as a whole, and component ablations (e.g., rule-based agents, belief-free prompting, or news-free observations) are needed to isolate the contribution of the language pipeline. Fifth, the player-mode counterfactual freezes all other agents' decisions; with sufficient compute, future versions could let agents re-decide live around the player. Sixth, the user study relies on a small technology-fair sample that skews young, highly educated, and toward computer-science backgrounds, and lacks a control condition. Finally, the EPTI formulas remain heuristic without formal validation.

Ethics and Broader Impact

WonderEcon is a research and educational tool, not a source of financial advice; we caution against misuse of its outputs for real decisions. The “economic personality” labels could reinforce stereotypes if LLM biases are not accounted for; we present EPTI as exploratory. Player-choice data are collected only with informed consent, anonymized, and reported in aggregate. We see positive value for economics education and public financial literacy.

Acknowledgments

This work was supported by Shenzhen Loop Area Institute under Grant 2026SFP004, Shenzhen Medical Research Fund (B2503005), NSFC grant 72495131, Major Frontier Exploration Program (Grant No. C10120250085) from the Shenzhen Medical Academy of Research and Translation (SMART), the International Science and Technology Cooperation Center, Ministry of Science and Technology of China under Grant 2024YFE0203000, Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001), Program

for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2023ZT10X044), Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing, and Shenzhen Science and Technology Program ZDSYS20230626091302006.

References

- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 337–371.
- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.
- Jonas Becker, Lars Benedikt Kaesberg, Niklas Bauer, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. [MALLM: Multi-agent large language models framework](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 418–439.
- Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin. 2026. [Behavioral economics of AI: LLM biases and corrections](#). Working Paper 34745, National Bureau of Economic Research.
- 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)*, pages 159–167.
- Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. [The emergence of economic rationality of GPT](#). *Proceedings of the National Academy of Sciences (PNAS)*, 120(51):e2316205120.
- Werner F. M. De Bondt and Richard Thaler. 1985. [Does the stock market overreact?](#) *The Journal of Finance*, 40(3):793–805.
- Kshama Dwarakanath, Tucker Balch, and Svitlana Vyetrenko. 2024. ABIDES-Economist: Agent-based simulator of economic systems with learning agents. *arXiv preprint arXiv:2402.09563*.
- Shane Frederick, George Loewenstein, and Ted O'Donoghue. 2002. [Time discounting and time preference: A critical review](#). *Journal of Economic Literature*, 40(2):351–401.
- Shen Gao, Yuntao Wen, Minghang Zhu, Jianing Wei, Yuhang Cheng, Qunzi Zhang, and Shuo Shang. 2024. Simulating financial market via large language model based agents. *arXiv preprint arXiv:2406.19966*.

- Wynne Godley and Marc Lavoie. 2007. *Monetary Economics: An Integrated Approach to Credit, Money, Income, Production and Wealth*. Palgrave Macmillan.
- Lynne Hamill and Nigel Gilbert. 2015. *Agent-Based Modelling in Economics*. Wiley.
- Jiale Han, Xiang Li, Jing Qian, Wenyan Gu, Pin Gao, Ye Luo, Hongyuan Zha, Dacheng Tao, Benyou Wang, and Lin William Cong. 2026. From economic agents to agentic economies: A systems blueprint for economic world models. *arXiv preprint arXiv:2608.06020*.
- Yuzhi Hao and Danyang Xie. 2025. A multi-LLM-agent-based framework for economic and public policy analysis. *arXiv preprint arXiv:2502.16879*.
- Charles A. Holt and Susan K. Laury. 2002. *Risk aversion and incentive effects*. *The American Economic Review*, 92(5):1644–1655.
- John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report w31122, National Bureau of Economic Research.
- Narasimhan Jegadeesh and Sheridan Titman. 1993. *Returns to buying winners and selling losers: Implications for stock market efficiency*. *The Journal of Finance*, 48(1):65–91.
- Daniel Kahneman and Amos Tversky. 1979. *Prospect theory: An analysis of decision under risk*. *Econometrica*, 47(2):263–291.
- Debanjana Kar, Leopold Böss, Dacia Braca, Sebastian Maximilian Dennerlein, Nina Christine Hubig, Philipp Wintersberger, and Yufang Hou. 2025. *MathBuddy: A multimodal system for affective math tutoring*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 821–838.
- Hannah Kim, Kushan Mitra, Chen Shen, Dan Zhang, and Estevam Hruschka. 2025. *AIPOM: Agent-aware interactive planning for multi-agent systems*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–96.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. *CAMEL: Communicative agents for “mind” exploration of large language model society*. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 51991–52008.
- Jiazheng Li, Artem Bobrov, Runcong Zhao, Cesare Aloisi, and Yulan He. 2025. *AERA chat: An interactive platform for automated explainable student answer assessment*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 536–545.
- 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 (ACL)*, pages 15523–15536.
- Jianhao Lin, Lexuan Sun, and Yixin Yan. 2025. Simulating macroeconomic expectations in survey experiments with LLM-based economic agents. *arXiv preprint arXiv:2505.17648*.
- Ananth Madhavan. 1992. *Trading mechanisms in securities markets*. *The Journal of Finance*, 47(2):607–641.
- 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*. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1390–1399.
- 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*. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- John F. Muth. 1961. *Rational expectations and the theory of price movements*. *Econometrica*, 29(3):315–335.
- Keyu Pan and Yawen Zeng. 2023. Do LLMs possess a personality? making the MBTI test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. *Generative agents: Interactive simulators of human behavior*. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22.
- Maosheng Qin, Renyu Zhu, Mingxuan Xia, Chenkai Chen, Zhen Zhu, Minmin Lin, Junbo Zhao, Lu Xu, Changjie Fan, Runze Wu, and Haobo Wang. 2025. *CrowdAgent: Multi-agent managed multi-source annotation system*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 925–942.
- Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. LLM economicus? mapping the behavioral biases of LLMs via utility theory. *arXiv preprint arXiv:2408.02784*.
- Paul A. Samuelson. 1938. *A note on the pure theory of consumer’s behaviour*. *Economica*, 5(17):61–71.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. *A psychometric framework for evaluating and shaping personality traits in large language models*. *Nature Machine Intelligence*, 7:1954–1968.

- Herbert A. Simon. 1955. [A behavioral model of rational choice](#). *The Quarterly Journal of Economics*, 69(1):99–118.
- Joseph E. Stiglitz and Andrew Weiss. 1981. Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3):393–410.
- Leigh Tesfatsion. 2006. [Agent-based computational economics: A constructive approach to economic theory](#). In Leigh Tesfatsion and Kenneth L. Judd, editors, *Handbook of Computational Economics*, volume 2, pages 831–880. Elsevier.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. [A survey on large language model based autonomous agents](#). *Frontiers of Computer Science*, 18(6):186345.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. [The rise and potential of large language model based agents: A survey](#). *Science China Information Sciences*, 68(2):121101.
- Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. 2025. [PromptSculptor: Multi-agent based text-to-image prompt optimization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 774–786.
- Xi Xu, Sumin Jo, Adam Officer, Angela Chen, Yufei Huang, and Lei Li. 2025. [BioGraphia: A LLM-assisted biological pathway graph annotation platform](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 480–486.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Runcong Zhao, Artem Bobrov, Jiazheng Li, Cesare Aloisi, and Yulan He. 2025. [LearnLens: LLM-enabled personalised, curriculum-grounded feedback with educators in the loop](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 625–633.

A Tick Order

Each monthly tick executes a fixed 25-step order, grouped into nine stages below. Steps marked [LLM] issue batched language-model calls; [MKT] steps run a rule-based market clearing (Section 3.3, Eq. 1 and its dual); the rest are deterministic bookkeeping.

Monthly tick order (25 steps, 9 stages)
1. Setup Step 1 update tick + economic-background/news; Step 2 apply exogenous interest-rate shock. ↓
2. Policy [LLM] Step 3 central-bank injection/withdrawal; Step 4 government tax redistribution. ↓
3. Settlement Step 5 compute endogenous quantities; Step 6 rank households (wealth) and firms (profit/scale); Step 7 environment settlement (mortgage, taxes). ↓
4. Loan Step 8 household loan [LLM]; Step 9 firm loan [LLM]; Step 10 bank loan rate [LLM]; Step 11 loan market [MKT]. ↓
5. Labor Step 12 household wage & quit [LLM]; Step 13 firm R&D/wage/production [LLM]; Step 14 labor market [MKT]. ↓
6. Goods Step 15 household goods beliefs/price/ratios [LLM]; Step 16 firm pricing [LLM]; Step 17 goods market, per sector [MKT]. ↓
7. Stock Step 18 household stock [LLM]; Step 19 firm stock [LLM]; Step 20 two call auctions, necessity then luxury [MKT]. ↓
8. Deposit Step 21 household/firm/bank deposit beliefs & rates [LLM]; Step 22 deposit market [MKT]. ↓
9. Close Step 23 refresh environment aggregates; Step 24 append history (incl. the GDP proxy); Step 25 snapshot all agent state for the next tick.

B Prompt Templates

Every economic action runs the same four-step observation→belief→price→action pipeline (Section 3.2). As a representative example we give the household loan action verbatim; the four steps share an identical observation block and differ only

in the system instruction and the appended prior-step outputs. Remaining actions follow the same structure and are in the released code.

Step 1 — Observation → Belief

System: You are a rational household economic agent in a simulated economy. Your goal is to maximize your expected total utility over time. Form a belief about loan interest rate direction and your borrowing strategy. Respond with a concise analytical paragraph labeled 'BELIEF:' (2–4 sentences).

User:

=== Household Loan-Rate Belief Generation (tick 7 of 13, each tick = 1 month) ===

HOUSEHOLD ATTRIBUTES:

- type: labor force (has wage income); risk preference: conservative
- last tick: no loan; comprehensive ability β : 0.81 (range 0.0–2.0)

OBSERVATIONS:

- last period weighted avg bank loan rate: 5.56%
- economic background: Inflation is worsening significantly.
- macroeconomic news: Tick 7 of 13. MAJOR POLICY CHANGE: the central bank has just raised the 1-year loan rate 5.31%→5.56% and the mortgage multiplier 0.7x→0.85x, the first hike of the cycle. [...]

ENDOGENOUS ATTRIBUTES:

- wealth level: level 2 — 20%–40% (lower-middle class)
- No extra government subsidy this tick (standard subsidy only).

Based on the above, analyze the likely direction of loan rates next period, how it affects your borrowing strategy, and the loan rate acceptable to you this tick.

BELIEF:

Step 2 — Belief → Price (desired rate)

System: Based on your belief, output your maximum acceptable loan rate as EXACTLY one line: desired_loan_rate: X.XX% (annualized).

User: [same block as Step 1] + your Step-1 belief.

Output ONLY:

desired_loan_rate: X.XX%

Step 3 — Action (borrow?)

System: Decide whether to borrow this period. CONSTRAINT: loan $\leq$ 50% of last-tick wage, used entirely for consumption. CONSEQUENCES: the loan plus most available funds go to consumption and the rest fully to stocks, leaving you highly exposed to stock prices. Respond EXACTLY one line: decision: yes OR decision: no.

User: [same block] + your belief + desired loan rate 5.56%.

Output ONLY:

decision: yes/no

Step 4 — Action (amount, if borrowing)

System: Set the loan amount as a percentage of last wage. CONSTRAINT: in [5%, 50%] (max 4166.81 CNY). Respond EXACTLY: loan_percentage: XX%.

User: [same block] + your belief + desired loan rate + last-tick wage.

Output ONLY:

loan_percentage: XX%

C EPTI Criteria

EPTI scores a player against a peer group (household agents sharing the player's labor-force status

Dim. (poles)	Criterion (desired)	<i>w</i>
Time: S/A	consumption ratio $\uparrow \Rightarrow$ S	1
Risk: R/P	loan ratio $\uparrow \Rightarrow$ R	1
	stock ratio $\uparrow \Rightarrow$ R	2
	necessity-stock share $\downarrow \Rightarrow$ R	1
	quit job $\Rightarrow$ R	1
Market: C/T	necessity goods price dir.	1
	luxury goods price dir.	1
	necessity stock price dir.	1
	luxury stock price dir.	2
Taste: U/F	luxury consumption share $\uparrow \Rightarrow$ F	1

Table 3: EPTI criteria and weights. For the market-stance rows, opposing the peer-majority price direction votes C and agreeing votes T (an undefined majority contributes 0).

Economic Principle	Indicator(s)	Predicted
DDM	Stock price	stock price declines
Okun's law	Employment rate, GDP	employment and GDP both decline
Credit channel	Loan volume	loan volume contracts
Intertemporal subst. (Euler)	Deposit volume	deposit volume rises
IS curve	GDP	GDP slows or declines
Phillips curve	Goods price, Employment rate	inflation slows as employment falls

Table 4: Macro-validity targets. Each row verifies an economic principle under the interest-rate hiking scenario.

and wealth level at the played tick), using the desired values from the decision stage rather than post-clearing outcomes. Each dimension tallies the weighted votes of its criteria (Table 3) and takes the left pole (S/R/C/U) when the tally is non-negative and the right pole (A/P/T/F) otherwise, forming a code such as SRCF.

D Macro-Validity Principles

Table 4 lists the six principles verified in Section 5.1, each with its indicator(s) and predicted trend under the interest rate hiking scenario; per-model verification rates are in Table 1.

E User Study Questionnaire

The 33 participants voluntarily tried WonderEcon at a technology fair and filled out the questionnaire

for free. They were 67% male, mostly aged 18–34 (94%), highly educated (over 60% at the master's or doctoral level), and drawn largely from computer science and AI backgrounds.

Administered anonymously after one play session, the questionnaire's four rated sections map to Table 2 (the demographics and optional free-text questions are omitted as unanalyzed). Each item uses a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree; for Part III, 1 = not at all accurate, 5 = very accurate).

WonderEcon Player-Experience Questionnaire (rated sections)

Part I — Usability.

- U1. I can easily locate the functions and pages needed to make decisions without external help (navigation convenience).
- U2. The system responds smoothly during operation, without noticeable lag or waiting (interaction smoothness).
- U3. The economic data and information presented are easy to obtain and understand (information accessibility).

Part II — Experience.

- E1. While making economic decisions, I feel immersed in the simulated world (immersion).
- E2. This platform helps me understand how the economy works (perceived usefulness).
- E3. I would be willing to use this platform again for an economic decision-making experience (willingness to replay).

Part III — EPTI Accuracy. After the session, the system generates an Economic Personality Type Indicator (EPTI). Participants rate how accurately each dimension characterizes them.

- P1. EPTI accurately characterizes my *time preference* (immediate consumption vs. delayed gratification).
- P2. EPTI accurately characterizes my *risk attitude* (conservative vs. aggressive).
- P3. EPTI accurately characterizes my *market stance* (deliberative vs. conformist).
- P4. EPTI accurately characterizes my *consumption taste* (necessity-oriented vs. luxury-oriented).

Part IV — Overall Satisfaction.

- S1. Overall, I am satisfied with my experience on the WonderEcon player platform.

F Availability and Licensing

The engine, both web interfaces and the demonstration video are publicly available. Code is released under Apache-2.0; text, data, and our original visual assets (indoor-scene and character models) under CC-BY-4.0. The outdoor-scene models are third-party assets under their original licenses.