GAME: A GOVERNED AGENT MACROECONOMY AS A REALISTIC POLICY SANDBOX

Anonymous authors
Paper under double-blind review

ABSTRACT

Macroeconomic simulation has progressed from agent-based models to systems whose agents are driven by large language models (LLM), yet existing LLM-based frameworks focus on reproducing economic phenomena or individual decision-making while paying little attention to the fidelity and robustness of simulated outputs, and hardly any supports policy analysis that jointly embeds central-bank and fiscal interventions. We propose the Governed Agent MacroEconomy (GAME), which couples a stock-flow-consistent economy of households, firms, banks, a central bank, and a government with LLM household cognition. Parameters are calibrated on historical sequences via the simulated method of moments, household groups are condensed from survey microdata, and each simulation step runs a four-layer cycle: perception layer from memory and neighbor signals, decision layer with LLM over the assembled information set, governance layer that anchors decision parameters to realistic conditions, and settlement layer through labor, goods, and credit market equilibria, along with the policy authorities periodically survey the economy and intervene with their instruments. On US and Canadian macroeconomic time series, under a protocol that strips all temporal, monetary, and regional identity cues from every prompt, GAME attains state-of-the-art alignment with real series in statistical similarity and result robustness, acts as a policy sandbox whose responses to policy changes are correctly signed and dose-responsive, replicates exogenous shocks of distinct types in line with classical economic principles, and its component ablations confirm that every mechanism is necessary. We believe that GAME can advance deeper research in the social sciences.

1 INTRODUCTION

Macroeconomic forecasting has long been a prominent research topic. As early as Turvey (1950), academic papers began discussing the macroeconomic patterns of real-world social operations from an economic perspective. It was not until Tesfatsion (2006) that Agent-Based Models (ABM) were proposed and applied to macroeconomic simulations, shifting this field toward statistics and machine learning (Farmer & Foley, 2009; Dawid & Gatti, 2018; Haldane & Turrell, 2019; Poledna et al., 2023), with increasing attention paid to more diverse economic agents and complex operational parameters. The emergence of Large Language Models (LLM) has enabled more realistic simulations of complex systems such as social platforms and public policy governance (Yang et al., 2024; Gasztowtt et al., 2024), and this research area is no exception (Li et al., 2024; Zhang et al., 2026; 2025).

However, several critical shortcomings in current research on macroeconomic simulation with LLM remain insufficiently addressed. First, most existing studies focus primarily on capturing economic phenomena or modeling individual decision-making (Li et al., 2024; Feng et al., 2025; Taillandier et al., 2025), while insufficient attention has been paid to the fidelity and robustness of simulation outputs, the alignment between simulation results and real-world observations (Liu et al., 2026; Xia et al., 2026), and robustness modules such as decision drift constraints and collective expectation construction (Chang et al., 2025; Wang et al., 2025). Second, few research frameworks support policy simulation applications, and almost none simultaneously incorporate central bank and fiscal authorities' economic interventions (Chen et al., 2025; Gasztowtt et al., 2024). Even those that do support such analyses have not yet evaluated whether these interventions can generate emergent

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.

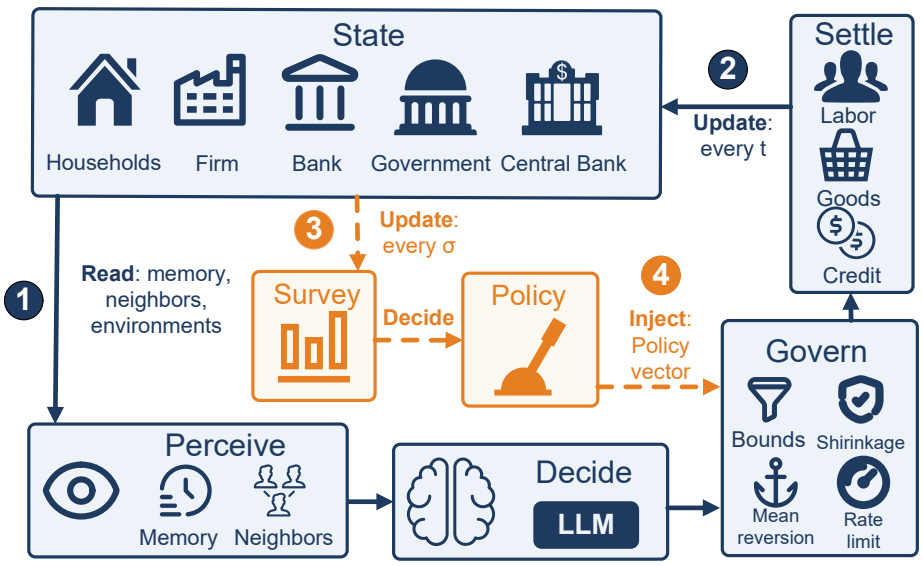


Figure 1: Overview of the GAME architecture. Household agents run every t perception–decision–governance–settlement cycle (blue panel) over the economy state, while the policy authorities run a σ survey-and-adjust cycle (orange panel) whose policy vector is injected into settlement and held for the intervening unit steps.

outcomes consistent with realistic scenarios, for instance whether an LLM-embedded legislature that passes bills reproduces the downstream macroeconomic responses observed around real legislation (Gasztowtt et al., 2024).

To address the above issues, we propose the Governed Agent MacroEconomy (GAME), a system that integrates parameterized interaction-based ABM with LLM agent decision-making. As shown in Fig. 1, the economy has five sectors, households, firms, banks, a central bank, and a government, and runs on two clocks: unit steps and survey periods. Parameters are calibrated by the simulated method of moments (SMM) on historical sequences, household groups are condensed from survey microdata, and a fixed network connects them. At each unit step, household agents run a four-layer cycle: **perception** assembles the information set from memory and neighbors, **decision** generates intentions through an LLM, **governance** anchors the intentions to realistic ranges, and **settlement** clears the labor, goods, and credit markets, whose outcomes feed the next step's memory. At survey boundaries, the central bank and government read a household questionnaire and reset their instruments.

We evaluate GAME on macroeconomic time series and household survey data from the United States and Canada, against recent pure ABM and LLM-based baselines, under a protocol that removes temporal, regional, and currency identifiers. Across these comparisons, GAME attains state-of-the-art statistical similarity and result robustness; within the policy sandbox it simulates real-world policies and reproduces the empirical responses to policy changes; at the macro level it reproduces distinct exogenous shocks in line with classical economic principles; and component ablations confirm that each module is necessary.

Our contributions:

- A more realistic ABM-LLM hybrid macroeconomic simulation framework, integrating the parameter calibration and interaction mechanisms of ABM with the decision-making flexibility of LLMs. To our knowledge, GAME is the first LLM-based macroeconomic simulator to align macroeconomic indicators with real time series, attaining state-of-the-art statistical similarity and robustness against a range of baselines.
- A reactive policy testing sandbox: GAME incorporates central-bank and government interventions with pluggable decision modules, and empirical experiments check the sign and dose-response of its transmissions and their consistency with classical economics.

2 Related Works

Quantitative agent based models. This line of work builds systems of many heterogeneous agents with pre-specified action rules and simulates the macroeconomy as a whole. Gualdi et al. (2017) studied the interaction between firms and banks in production and credit under interest-rate regulation; Reissl (2020) added household agents and let banks vary their lending rules across firms; and the same authors later added heterogeneous, changeable expectations and their interaction with decision-making (Reissl, 2021). A recent line builds larger multi-agent systems (Poledna et al., 2023; Coccia & Russo, 2025). These systems are stock-flow consistent (SFC): accounting identities discipline every transaction, which keeps the generated sequences stable and consistent with economic laws. Their behavior is nonetheless rigid, because the rules are written directly by researchers and lack the adaptive flexibility of LLM agents.

Generative economic and social simulation. Work of this kind constructs interactions among heterogeneous agents and lets LLMs take the corresponding decisions. EconAgent embeds LLM household cognition in a macro environment and checks macro regularities, but it omits fiscal and monetary policy and aligns loosely with the real world (Li et al., 2024); hierarchical frameworks add LLM policy planners, but whether the policy impact matches reality, and whether the policy learner helps, remain unclear (Chen et al., 2025); GPLab evaluates generative policy laboratories with consistency metrics, but only on micro-level behaviors and opinions, without macro-level variables (Zhang et al., 2026). A simulation is therefore needed that aligns with real-world conditions, yields robust outcomes, and supports the testing of macroeconomic policy.

3 The Game Architecture

3.1 Problem Formulation

For an economic system e, let the evaluation window be measured in unit step $t \in \{1, 2, \ldots, T\}$. Given the history of all macro indicator series $\mathcal{H}$, and current state S_0^e of e at $t = 0$, the problem is to construct an economic simulator $\Phi(\cdot)$ that generates a *time panel over the window*:

$$\hat{\mathbf{Y}}^e \;=\; \Phi(\mathcal{H}, S_0^e) \tag{1}$$

where $\hat{\mathbf{Y}}^e$ is the simulated panel over this window regarding macroeconomic (inflation, unemployment, etc.) or policy decision (the policy rate and fiscal instruments) indicators, and $\mathbf{Y}^e$ is the realized observation panel of the same indicators. The simulator should make $\hat{\mathbf{Y}}^e$ at every unit step as close as possible to $\mathbf{Y}^e$.

3.2 Overview of the Game System

Game represents an economy as five sectors, N weighted household groups, each endowed with LLM cognition, on one side, and four rule-based institutional functions (firm, bank, government, central bank) on the other, meeting in three markets, with a flow-of-funds kernel settling every transaction as matched changes of sectoral balance sheets. The global state S_t^e collects the household states, the institutional balance sheets, and the prevailing prices and rates.

Each unit step t runs one full perception–decision–governance–settlement cycle, while policy runs on σ interval. At every t, household groups update their spending, saving, borrowing, and expectations, and the three markets settle. Policy changes only at survey-period boundaries, where σ indexes consecutive blocks: the central-bank and government agents review the household survey and the macro conditions and reset their instruments, matching the real-world cadence at which policy reviews occur.

Definitions. In Game, the definition of these variables is crucial: the state S_t^e is the full economy settled at the end of step t; the information set $\mathcal{I}_{i,t}$ is what agent i reads at the head of step t; and the

policy vector $\mathbf{p}_\sigma$ is the instrument block held fixed within survey period σ:

$$S_t^e = \big(\mathcal{A}_t : \{s_{i,t},\, \mu_{i,t},\, \chi_i,\, \omega_i\}_{i=1}^N,\ \mathcal{B}_t,\ \mathbf{q}_t\big), \tag{2a}$$

$$\mathcal{I}_{i,t} = \big(\mu_{i,t},\ \bar{\mathbf{s}}_{i,t}^{\mathcal{G}},\ \mathbf{q}_{t-1},\ \mathbf{p}_{\sigma(t)}\big), \tag{2b}$$

$$\mathbf{p}_\sigma = \big(r_\sigma,\ \tau_\sigma^{\mathrm{inc}},\ \tau_\sigma^{\mathrm{c}},\ g_\sigma,\ z_\sigma\big). \tag{2c}$$

In Eq. 2a, $\mathcal{A}_t$ collects the household static profile χ_i, time-varying state $s_{i,t}$, rolling memory $\mu_{i,t}$, and aggregation weight ω_i of all N groups; $\mathcal{B}_t$ the balance sheets of the four institutional sectors; and $\mathbf{q}_t$ the environment vector of economic conditions. In Eq. 2b, the memory $\mu_{i,t}$ and the neighbor signal $\bar{\mathbf{s}}_{i,t}^{\mathcal{G}}$ are private to group i. In Eq. 2c, the five instruments are set at each survey-period boundary and held fixed for every σ period.

3.3 Unit-Step Mechanics

Unit-step execution. At t=0, $\mathcal{I}_{i,0}$ and $\mathbf{p}_0$ are initialized based on the initial economic state S_0^e, with the memory starting empty, $\mu_{i,0} = \varnothing$. At every unit step $t > 0$, the simulator runs all four layers in sequence:

$$\mathcal{I}_{i,t} = U\big(\mu_{i,t},\ \bar{\mathbf{s}}_{i,t}^{\mathcal{G}},\ S_{t-1}^e\big), \tag{3a}$$

$$\mathbf{x}_{i,t}^{\mathrm{raw}} = \mathrm{LLM}(h_i,\ \mathcal{I}_{i,t}), \tag{3b}$$

$$\mathbf{x}_{i,t} = G\big(\mathbf{x}_{i,t}^{\mathrm{raw}}\big), \qquad \mathbf{a}_{i,t} = \Pi\big(\mathbf{x}_{i,t}\big), \tag{3c}$$

$$S_t^e = \big(\mathrm{Credit} \circ \mathrm{Goods} \circ \mathrm{Labor}\big)\big(S_{t-1}^e,\ \{\mathbf{a}_{i,t}\}_{i=1}^N,\ \mathbf{p}_{\sigma(t)}\big). \tag{3d}$$

where h_i is the household i prompt template; $\mathbf{x}_{i,t}^{\mathrm{raw}}$ is the raw intention vector; $\mathbf{x}_{i,t}$ is its governed counterpart after the governance gates; and $\big(\mathrm{Credit} \circ \mathrm{Goods} \circ \mathrm{Labor}\big)$ is the settlement composite, which clears the labor, goods, and credit markets in execution order.

Initialization. Households with different economic circumstances often have heterogeneous perceptions and behavioral choices regarding the economy. At the same time, these households can make further judgments about future expectations based on the surrounding economic activities. Therefore, we designed a cluster of agents representing multiple N households at initialization and endowed them with anticipating expectations based on social networks. The N agents are condensed from survey microdata by weighted stratified binning on income and net worth: agent i is the weighted average of the survey records in its cell, ω_i is the cell's share of total survey weight with $\sum_i \omega_i = 1$. This design ensures that each heterogeneous class group is appropriately accounted for during the simulation process.

We further introduce a static network that enables agents to mutually observe one another's states, thereby forming expectations more consistent with the economic environment. A small-world network $\mathcal{G}$ connects the agents and carries peer exposure. Let distances $d(i,j)$ between all agent pairs i, j be

$$d(i,j) = \sum_F \big|\, [\chi_i]_F - [\chi_j]_F \,\big|, \tag{4}$$

where F index the standardized age, income, and wealth of the household, after which each agent is connected to its k nearest neighbors, with these edges preferentially established within the same income stratum. Then, at a specific reconnection probability p^*, these edges are randomly reassigned to other strata to satisfy the edge ratio mentioned in (Chang et al., 2025), and ultimately obtain the observation network $\mathcal{G}$. The full construction procedure is given in Appendix C.2.

Perception layer (step 1). At this layer, households can observe the current economic conditions and update their latest action memory. Through $\mathcal{G}$, each cognition receives the following signal $\bar{\mathbf{s}}_{i,t}^{\mathcal{G}}$ from neighbour states:

$$\bar{\mathbf{s}}_{i,t}^{\mathcal{G}} = \frac{1}{|\mathcal{N}_{\mathcal{G}}(i)|} \sum_{j \in \mathcal{N}_{\mathcal{G}}(i)} [s_{j,t-1}]_H \tag{5}$$

where H index the features of neighbour j's state, the employment status, the consumption relative to its own consumption anchor, the inflation expectation, and the perceived unemployment risk.

Such signal aggregation enables agents to perceive the economic expectations of other agents, which not only aligns with the formation mechanism of public expectations in reality but also mitigates the problem of decision drift among agents. The memory of step t is updated in the same spirit, by pushing the step's realized experience with current memory:

$$\mu_{i,t} = \mathrm{push}\big(\mu_{i,t-1},\, \mathbf{x}_{i,t-1},\, \mathbf{a}_{i,t-1}\big), \tag{6}$$

The current memory is thus constituted jointly by the memory from the preceding step, decision, and the action taken.

Decision layer (step 2). This step implements Equation 3b. The LLM prompt h_i integrates the information set to make decisions and responds in the manner of a household in a real-world context: monthly expenditure plans on a per-household basis, additional monthly savings, new borrowing in the current quarter, housing budget for the current quarter, expected price inflation, perceived employment risk, and work intensity, among others. This generates the raw intention vector $\mathbf{x}_{i,t}^{\mathrm{raw}}$. We considered the following aspects when designing the prompt. First, the household's circumstances are described using everyday terminology, that is, informing the relative prices of daily consumption and investment. Second, the prompt h_i contains no country names, time periods, or currency units to prevent the model from directly recalling economic data as much as possible from $\mathcal{I}_{i,t}$. The exact prompt templates are given in Appendix C.3.

Governance layer (step 3). To prevent counterintuitive decisions by large models or drift issues at both individual and group levels, we have designed a series of parameter calibration mechanisms. This step realizes Eq. 3c: the operator G screens every batch of raw proposals, through three ordered per-group gates and one cohort-level correction. *Gate 1 (economic bounds):* All components of $\mathbf{x}_{i,t}^{\mathrm{raw}}$ will be adjusted to conform to economic significance, echoing the empirical grounding of LLM agents in Xia et al. (2026). *Gate 2 (robust shrinkage):* For each candidate proposal field, the population-weighted median and standard deviation are calculated across all households, and outliers are corrected to within the median ± standard deviation range (Liu et al., 2026). *Gate 3 (rate limiting and smoothing):* Each agent's decision vector will be a weighted average of its previous decision, with the difference strictly limited Turvey (1950). *Gate 4 (mean reversion):* Finally, the weighted average vector of all households will be calibrated relative to the previous period, and this calibration will be applied to all decision vectors (Wang et al., 2025). The four gates compose into the governed intentions $\mathbf{x}_{i,t}$; their details are given in Appendix C.4.

The projection operator Π then maps the governed intentions $\mathbf{x}_{i,t}$ into feasible actions $\mathbf{a}_{i,t} = (c_{i,t}, \hat{b}_{i,t}, \ell_{i,t}, f_{i,t}, h_{i,t})$, indicating consumption demand, credit application, labor supply, financial-asset purchase, and housing purchase. Three economic truncations are enforced: consumption is capped by the liquid budget, credit applications are rationed by collateral at the loan-to-value limit, and the two asset purchases are jointly capped by the liquidity remaining after consumption, while labor supply passes through with its admissible intensity bounds, employment itself being rationed later by the settlement kernel. The exact constraint forms are given in Appendix C.4.

Settlement layer (step 4). After households take appropriate actions, the remaining agents in the system will clear the market according to household decisions to achieve the goal of system advancement. This step realizes Eq. 3d: the state is advanced deterministically by clearing the three markets in execution order, with the composition reading right to left and the market-by-market clearing rules and their institutional details formalized in Appendix C.5. In brief, Labor fills the employment quota by ranking groups on employment probability and intended hours, total labor aggregating the employed groups' intended hours (Mortensen & Pissarides, 1994); Goods clears total demand against capacity, output being the smaller of the two, with prices and wages carrying an inflation anchor that bounds drift (Gualdi et al., 2017); and Credit approves and rations credit applications at the loan rate, each granted credit being the approved application scaled by a funding-ration factor (Reissl, 2020).

Every cleared trade is then settled through the accounting circuit: each transaction is posted twice, payer's outflow, receiver's inflow, so sectoral net-worth changes sum to zero, an identity verified by paired accounting checks at every step (Copeland, 1949; Godley & Lavoie, 2007); stacking the updated sector tables yields S_t^e.

Readout. With both clocks in place, the simulated panel entry of Eq. 1 is determined at the end of every unit step by reading the target series off the settled state through a fixed readout map R:

$$\hat{\mathbf{y}}_t^e = R\big(S_t^e\big) \in \hat{\mathbf{Y}}^e \tag{7}$$

3.4 Survey Period Execution

Policy pipeline. At the boundary of each survey period σ, the central bank agent and the government agent run their own *policy cognition*: they survey the households, the survey function $Q(\cdot)$ compiles the weighted questionnaire aggregate of household expectations, perceived risks, and behavioral intentions from their states and memories, read the macro conditions off the same settled state S_{t-1}^e, and set the policy vector

$$\mathbf{p}_\sigma \;=\; P\big(S_{t-1}^e\big), \qquad \mathbf{p}_t = \mathbf{p}_{\sigma(t)}, \tag{8}$$

where P is the policy cognition, instantiated below as the composite of a decision rule D with bounded increments; the resulting vector remains in force for the θ unit steps of the period.

Questionnaire the survey function of Eq. 8 compiles, at each survey-period boundary σ, the weighted questionnaire aggregate from the household states as

$$Q(S_{t-1}^e) = \sum_{i=1}^{N} \omega_i \, \phi(s_{i,t-1}, \mu_{i,t-1}), \tag{9}$$

ϕ extracts expectation, intention, and stress fields from the governed state and memory; the weighted mean and the weighted 10%/90% percentiles are reported, so that the policy agents see both the central tendency and the disagreement among households.

Policy decision in the endogenous instantiations the operator P of Eq. 8 updates the policy vector by bounded increments,

$$\mathbf{p}_\sigma = P\big(S_{t-1}^e\big) = \mathrm{clip}_{\mathcal{P}}\big(\mathbf{p}_{\sigma-1} + \boldsymbol{\delta}_\sigma\big), \qquad \boldsymbol{\delta}_\sigma = D\big(Q(S_{t-1}^e),\ \mathbf{q}_{t-1}\big), \tag{10}$$

where D is the policy agent's decision rule: reading the survey $Q(S_{t-1}^e)$ and the environment vector $\mathbf{q}_{t-1}$, it proposes the five instrument increments $\boldsymbol{\delta}_\sigma$, and $\mathrm{clip}_{\mathcal{P}}$ enforces the per-period bounds that prevent discontinuous jumps. Three instantiations of D are used: an **LLM** policymaker, a **Rule** policymaker, and the **Observed** mode. Their concrete settings are specified with the experiment suites in §4.2.

4 Experiments

In this section, we conducted experiments around the following questions: **RQ1 (realism):** does GAME reflect real economic activity more faithfully than recalibrated agent-based baselines, in macro data? **RQ2 (policy sandbox):** can GAME produce policies close to the real world and accurately reflect their effects? **RQ3 (phenomena):** does GAME respond to simulated exogenous shocks and generate phenomena consistent with economic experience? **RQ4 (components):** is each specially designed component of the GAME effective?

4.1 Datasets

Two data sources drive the system. The macro side is a monthly panel of official series for each economy, policy rate, employment and unemployment, CPI, personal consumption and income, GDP and investment, fiscal aggregates, and banking series, from 2016 to 2025. The household side compresses national household-wealth surveys (the US Survey of Consumer Finances and the Canadian Survey of Financial Security) into the agent profiles, aligned with a one-period lag to reflect publication delays. Each record carries survey weights, income flows, balance-sheet stocks, and demographics. More details are given in Appendix D.2.

Table 1: Comparison across LLM cognition backends (main result; the full table is given in Table 8). For each run, the average value of all target sequences on the target metric among all countries (GDP and consumption) is taken as the final result. All data across 20 independent runs are presented as mean ± standard deviation for all evaluated performance metrics. Each block swaps the household LLM under an identical kernel (GAME, EconAgent, or GPLab). In each column the four best values are shaded green, with darker shades marking better performance. ↑ larger is better, ↓ smaller is better.

Model	Pearson (%) ↑	JSD (%) ↓	RMSE ↓	Terminal CV (%) ↓	Seed RMSE SD ↓
Pure ABM					
AB-SFC	−0.2±15.1	28.4±4.7	3.55±1.15	1.75±0.32	1.16±0.17
firm–bank ABM	+1.6±9.7	20.2±3.1	5.07±0.79	2.09±1.26	1.18±0.57
Minsky	+23.0±9.2	31.5±12.7	**2.69±1.53**	4.38±0.62	1.62±0.19
glm-5.3-flash (Zeng et al., 2026)					
GAME (ours)	+37.2±6.2	17.2±4.3	3.07±0.49	1.67±0.70	**0.54±0.36**
EconAgent	−13.6±7.4	24.4±9.1	3.84±1.09	2.65±0.89	1.30±0.36
GPLab	+23.3±5.5	34.4±11.2	4.63±0.82	**1.45±0.31**	0.88±0.15
qwen3.8-flash (Qiu et al., 2026)					
GAME (ours)	**+44.7±6.9**	**12.8±4.0**	2.79±0.36	1.98±0.85	0.69±0.26
EconAgent	−12.6±8.3	35.8±11.0	5.38±1.16	3.98±1.64	1.44±0.09
GPLab	+25.5±6.1	33.3±10.5	5.38±0.96	1.61±0.62	1.09±0.43

4.2 SETTINGS

Baselines. **Pure ABM models**: AB-SFC (Reissl, 2021), an SFC macro model with heterogeneous expectations; Minsky (Reissl, 2020), a model of Minsky-style financial instability; Firm–Bank ABM (Gualdi et al., 2017), a stylized firms–households–bank model with financial-fragility-mediated hiring and firing. **ABM-LLM mixture models**: EconAgent (Li et al., 2024), which embeds LLM household cognition in a macro environment; GPLab (Zhang et al., 2026), a generative-policy agent-based framework; and GAME (ours). Minsky and EconAgent are re-projected onto each country's historical series panel through the National Projection Bridge, and Firm–Bank ABM is extended with central-bank and government blocks. Further settings are given in the appendix.

Evaluation metrics. All metrics operate on month-on-month growth rates rather than levels, so a single over- or under-estimation at one time point cannot contaminate the whole sequence (Reissl, 2021; Dosi et al., 2015). Five metrics cover two families: statistical similarity, Pearson correlation, Jensen–Shannon divergence, and the root-mean-square error (**RMSE**) (Li et al., 2024; Yang et al., 2026; Xia et al., 2026), and result robustness, the cross-seed coefficient of variation of the terminal series (**Terminal CV**) and the cross-seed standard deviation of the per-seed RMSE (**Seed RMSE SD**) (Zhang et al., 2026; Poledna et al., 2023). Appendix B.2 gives the formal definitions and computation conventions of all metrics.

4.3 REALISM WITH AND WITHOUT IDENTITY INFORMATION (RQ1)

Setup. We evaluate realism on GDP and consumption across the US and Canada, comparing GAME with the baselines under normal conditions against actual data: 2016–2019 serves as the training set, 2022–2023 as the validation set, and 2024–2025 as the test set, excluding the pandemic period (2020–2021).

Discussion. Table 1 and Figure 2 present the comparison. (1) GAME achieved the best performance in both statistical metrics and result stability (see full table), indicating that our model can better simulate macroeconomic trends and changes in economic velocity. (2) GAME ranked among the top models across multiple LLM backbones of varying scales, demonstrating strong scalability of these favorable results. (3) Pure ABM models showed superior result robustness compared to baseline LLM frameworks, confirming the claim that ABM models offer greater stability while LLMs provide more flexible decision-making.

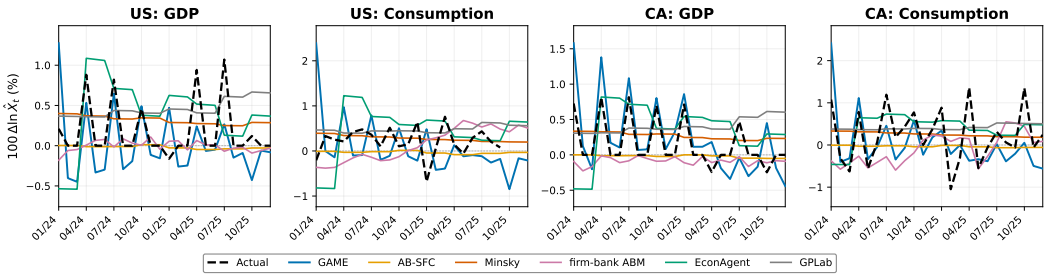


Figure 2: Monthly log growth rates of GDP and consumption for the US and Canada (2024-01–2025-12), chained month over month. Solid lines show seed means across twenty seeds; the dashed black line is actual data. Each LLM-bearing system uses its best-performing backend from Table 8.

Table 2: Comparison of endogenous policy decision agents with the observed central-bank and government decisions. Rule refers to a Taylor rule model; LLM to an LLM policymaker model; RL to a reinforcement-learning policymaker model; EconAgent to the endogenous policy rule of the external EconAgent sandbox (it has no expenditure lever, so that column is omitted). Rate ME is the mean signed error of the full monthly policy-rate series against the observed series; Terminal gap is the difference between the simulated and observed terminal policy rates; Expenditure ME is the mean signed error of the government expenditure series against the observed series, in percent of mean observed expenditure. Values are mean $\pm$ standard deviation over five runs; best performance per column in bold.

Policy agent	Rate ME (pp)	Terminal gap (pp)	Expenditure ME (%)
Rule	$+0.39 \pm 0.01$	$+1.30 \pm 0.03$	-38.8 ± 0.0
LLM	$+1.07 \pm 0.56$	$+2.42 \pm 0.90$	$\mathbf{-9.8} \pm 2.4$
RL	$\mathbf{+0.09} \pm 0.02$	$\mathbf{+0.87} \pm 0.06$	-23.9 ± 0.2
EconAgent	-1.04 ± 2.40	-1.69 ± 1.69	,

4.4 A Policy Sandbox: Decision Realism, Deviation Robustness, and Closed-Loop Optimization (RQ2)

Setup. We evaluate GAME as a policy sandbox along two dimensions. (1) Decision realism: whether endogenously generated policies align with the trajectory of real-world policymakers. We substitute the central bank and government with agents based on the Taylor rule, an LLM, and reinforcement learning (detailed in Appendix D.1), and compare their decision paths with actual historical policies. (2) Deviation robustness: whether the simulated economy responds plausibly to policies never observed in history. We perturb the policy rate and government expenditure, construct combined paths, and compare key indicators against the unperturbed scenario.

Discussion. Table 2 presents the decision-realism comparison: (1) Realism splits by lever. On the interest-rate lever, RL tracks the observed path most closely, the Taylor rule is second, the LLM policymaker is furthest (persistently above the observed easing cycle), and the EconAgent rule is least stable, its small mean error reflecting cancellation between runs. On the expenditure lever the ordering reverses: every agent underspends the observed budget, and the LLM is the most realistic fiscal actor. (2) The transmission experiments of Table 9, with their full discussion, are deferred to Appendix E.2. In brief, every shocked variant responds in the textbook direction with the response growing in the shock size, investment is the most elastic margin so counterfactual policies propagate through the interest-sensitive investment channel, and the two levers partially substitute rather than stack. Taken together, our framework can provide a high-fidelity testing platform for designed policy approaches, and generates economically consistent responses to exogenous shocks.

4.5 Macro Phenomena (RQ3)

Setup. We evaluate the simulator as a macro-phenomena generator from two directions: active crisis scenarios that inject exogenous perturbations, and passive regularity checks against canon-

Table 3: Single-module ablations study under 5 runs. The Full GAME result is identical to Table 1. In each column the four best values are shaded green, with darker shades marking better performance.

Ablated component	Pearson (%) ↑	JSD (bits %) ↓	RMSE ↓	Terminal CV (%) ↓	Seed RMSE SD ↓
Full GAME	**+44.7±6.9**	**12.8±4.0**	**2.79±0.36**	1.98±0.85	0.69±0.26
w/o $\bar{s}_i^{\mathcal{G}}$ interaction	+7.4±11.3	18.3±3.1	7.94±1.64	5.64±4.78	1.94±1.31
w/o LLM cognition	+5.0±10.5	22.7±8.5	8.30±0.27	3.36±1.47	**0.38±0.26**
w/ random cognition	+4.2±7.6	20.9±4.7	9.74±1.31	6.38±5.39	1.76±0.69
w/o $G(\cdot)$ governance	+14.1±9.6	22.5±6.9	6.49±0.41	**1.76±0.23**	0.52±0.18
w/o π^* settlement	+7.4±5.9	23.0±5.5	8.33±0.83	3.37±1.32	1.69±0.47
w/o $\mathbf{M}_t$ feedback	+9.3±9.6	15.0±3.3	6.91±0.74	3.18±2.04	1.45±0.55
w/o Q_σ survey	+3.5±16.9	14.1±3.4	6.80±1.50	2.97±1.26	1.72±0.80

ical stylized facts. (1) The financial-crisis family applies a twelve-month credit crunch that rations credit supply and raises lending rates, and asks how the recession propagates endogenously to output, unemployment, and investment, and whether countercyclical policy cushions it. (2) The pandemic-proxy family injects the lockdown contraction at the settlement level through two calibrated channels, a household labor-income path offset and a labor-quota offset on firm hiring, with a policy-response variant that adds monetary easing and a fiscal expansion.

For passive regularity checks, we compare GAME and the baselines on the classical Phillips curve, the Taylor rule, and inflation-adjusted GDP.

Discussion. Figures 3, 4, and 5 present the results; the full discussion is deferred to Appendix E.3. In brief, the simulator reproduces both shock families through their distinct transmission channels, with employment reacting first and output following after a lag. It also tracks the canonical Phillips, Taylor, and GDP regularities more closely than the compared baselines, combining the stability of pure ABM models with the dynamism of LLM cognition.

4.6 COMPONENT ABLATIONS (RQ4)

Setup. To investigate the effectiveness of each component, we designed six single-mechanism ablation experiments: removing the neighbor information from the cognition inputs, replacing the LLM household cognition with the rule cognizer, substituting the same as random output, disabling the governance operator $G(\cdot)$ that caps and reverts cohort-level intention drift, switching off the settlement-layer inflation anchor, freezing the macro state in the household information set, and suppressing the quarterly survey aggregate fed to the policy prompt, from top to bottom, corresponding to the contents of the table. A seventh arm replaces the LLM household cognition with independent uniform draws inside the governance hard bounds, leaving every other layer unchanged; since the governed kernel is fully intact, this arm isolates the contribution of the LLM cognition itself.

Discussion. Table 3 shows that every component contributes to simulation statistical similarity. Ablating the cognition- and governance-related content slightly improves the certainty of results, but at the cost of prediction effectiveness. These results fully demonstrate the effectiveness of the model's components and their contribution to the stability of the system outcomes.

5 CONCLUSION & LIMITATIONS

We presented GAME, a macroeconomic simulator that couples governed LLM household cognition with an SFC settlement kernel under a strict no-leakage contract. Across four evaluation questions, the system attains state-of-the-art alignment with real macro time series in both statistical similarity and result robustness across LLM backends; serves as a policy sandbox whose endogenous decision agents bracket real policymaker behavior and whose transmission responses are correctly signed and dose-responsive; reproduces distinct exogenous shock families through their appropriate transmission channels; and every component proves necessary, with governance and the inflation anchor the most broadly load-bearing. We believe that GAME enables social science research to advance in a deeper direction.

However, there are several limitations for GAME. Our consideration of variables in public policy remains incomplete, and the content regarding interaction mechanisms still needs to be added. Moreover, due to limitations in available data, our experiment is confined to the period after 2016, and earlier scenarios have not been considered.

AI USE STATEMENT

LLMs were used to generate household and policy intentions in the reported simulations and to assist in drafting and organizing this manuscript. All numerical results were produced by the documented experiment pipeline. The authors have reviewed all AI-assisted content and take full responsibility for the final content of this work. No LLM is listed as an author.

REPRODUCIBILITY STATEMENT

We commit to release the reproduction code as soon as out paper is accepted.

REFERENCES

George A Akerlof, William T Dickens, George L Perry, Robert J Gordon, and N Gregory Mankiw. The macroeconomics of low inflation. *Brookings papers on economic activity*, 1996(1):1–76, 1996.

Bank of Canada. Bank of canada valet API. `https://www.bankofcanada.ca/valet`, 2026.

Ben S Bernanke and Michael Woodford. Inflation forecasts and monetary policy, 1997.

Board of Governors of the Federal Reserve System. Survey of consumer finances (SCF) 2022. Technical report, Board of Governors of the Federal Reserve System, 2023.

Alessandro Caiani, Ermanno Catullo, and Mauro Gallegati. The effects of fiscal targets in a monetary union: a multi-country agent-based stock flow consistent model. *Industrial and Corporate Change*, 27(6):1123–1154, 2018.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. Llms generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 341–371, 2025.

Zhixun Chen, Zijing Shi, Yaodong Yang, Meng Fang, and Yali Du. Hierarchical multi-agent framework for dynamic macroeconomic modeling using large language models. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems Aamas*, 2025.

Samantha Coccia and Alberto Russo. Inequality and transmission channels of monetary policy in a macro agent-based model. *Economic Modelling*, pp. 107316, 2025.

Morris A Copeland. Social accounting for moneyflows. *The Accounting Review*, 24(3):254–264, 1949.

Herbert Dawid and Domenico Delli Gatti. Agent-based macroeconomics. *Handbook of computational economics*, 4:63–156, 2018.

Giovanni Dosi, Giorgio Fagiolo, Mauro Napoletano, Andrea Roventini, and Tania Treibich. Fiscal and monetary policies in complex evolving economies. *Journal of Economic Dynamics and Control*, 52:166–189, 2015.

J Doyne Farmer and Duncan Foley. The economy needs agent-based modelling. *Nature*, 460(7256): 685–686, 2009.

Federal Reserve Bank of St. Louis. Federal reserve economic data (FRED/ALFRED). https://fred.stlouisfed.org, 2026. Mirrors of official BLS, BEA, and Federal Reserve statistical releases.

Yeqi Feng, Yucheng Lu, Hongyu Su, Yixin Tao, and Tianxing He. Simcity: Multi-agent urban development simulation with rich interactions. *arXiv preprint arXiv:2510.01297*, 2025.

Henry Gasztowtt, Benjamin Smith, Vincent Zhu, Qinxun Bai, and Edwin Zhang. Large legislative models: Towards efficient ai policymaking in economic simulations. *arXiv preprint arXiv:2410.08345*, 2024.

Wynne Godley and Marc Lavoie. Monetary economics: an integrated approach to credit, money, income, production and wealth. *New York*, 2007.

Stanislao Gualdi, Marco Tarzia, Francesco Zamponi, and Jean-Philippe Bouchaud. Monetary policy and dark corners in a stylized agent-based model. *Journal of Economic Interaction and Coordination*, 12(3):507–537, 2017.

Andrew G Haldane and Arthur E Turrell. Drawing on different disciplines: macroeconomic agent-based models. *Journal of Evolutionary Economics*, 29(1):39–66, 2019.

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, 2024.

Guangya Liu, Cheng Wang, Jiangtong Li, Huafei Wu, and Changjun Jiang. Do llm agents really mimic humans? diagnosing and aligning microeconomic behaviors in macro-abms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36102–36122, 2026.

Dale T Mortensen and Christopher A Pissarides. Job creation and job destruction in the theory of unemployment. *The review of economic studies*, 61(3):397–415, 1994.

Ornith Team. Ornith-1.5: From self-scaffolding to self-improvement, 2026. URL https://ornith.ai/ornith_1_5.html.

Sebastian Poledna, Michael Gregor Miess, Cars Hommes, and Katrin Rabitsch. Economic forecasting with an agent-based model. *European Economic Review*, 151:104306, 2023.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Severin Reissl. Minsky from the bottom up–formalising the two-price model of investment in a simple agent-based framework. *Journal of Economic Behavior & Organization*, 177:109–142, 2020.

Severin Reissl. Heterogeneous expectations, forecasting behaviour and policy experiments in a hybrid agent-based stock-flow-consistent model. *Journal of Evolutionary Economics*, 31(1):251–299, 2021.

Statistics Canada. Survey of financial security (SFS) 2023. https://www150.statcan.gc.ca, 2024.

Statistics Canada. Statistics canada data tables. https://www150.statcan.gc.ca, 2026.

Patrick Taillandier, Jean Daniel Zucker, Arnaud Grignard, Benoit Gaudou, Nghi Quang Huynh, and Alexis Drogoul. Integrating llm in agent-based social simulation: Opportunities and challenges. *arXiv preprint arXiv:2507.19364*, 2025.

John B Taylor. Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, volume 39, pp. 195–214. Elsevier, 1993.

Leigh Tesfatsion. Agent-based computational economics: A constructive approach to economic theory. *Handbook of computational economics*, 2:831–880, 2006.

James Tobin. Inflation and unemployment. In *Essential readings in economics*, pp. 232–254. Springer, 1995.

Ralph Turvey. Income, saving and the theory of consumer behavior, 1950.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks. In *Proceedings of the 31st international conference on computational linguistics*, pp. 3913–3923, 2025.

Chen Xia, Zexi Kuang, and Yuqing Hu. Empirical grounding improves the realism of llm agents simulating human behavior during disruptions. *arXiv preprint arXiv:2607.17437*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:63469–63519, 2026.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Shuhan Zhang, Zifan Peng, and Yinwang Ren. Gplab: A generative agent-based framework for policy simulation and evaluation. *Journal of Artificial Societies and Social Simulation*, 29(1), 2026.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

A Overview of the Appendix

This appendix is organized as follows.

- **Appendix B (Reference Material)** introduces the notation tables, the formal metric definitions, and the hyperparameter settings.
- **Appendix C (Method Details)** provides the state definitions, the network construction procedure, the household prompt templates, the exact governance functional forms, and the settlement rules.
- **Appendix D (Experimental Details)** documents the three policy agent instantiations together with the policymaker prompt, and the data sources.
- **Appendix E (Additional Results)** reports the full five-backend comparison, the transmission experiments, and the macro-phenomena discussion.

B Reference Material

B.1 Notation

Tables 4 and 5 summarize the notation of the main text and of the appendices. Each entry states the sector the object belongs to among the five sectors of the system (Households, Firm, Bank, Government, Central Bank; *System* marks cross-sector or economy-level objects), its mathematical kind (scalar, vector, set, graph, operator, or index), its role (a *variable* evolves during the simulation, a *hyperparameter* takes its value from Table 6, and *constant*, *index*, and *operator* label the remaining cases), and its meaning. Symbols local to a single derivation and the evaluation metrics of Appendix B.2 are kept in their defining context.

Table 4: Main-text notation.

Symbol	Sector	Kind	Role	Meaning
Clocks and indices				
t	System	index	index	unit-step month clock, $1, \dots, T$
T	System	scalar	constant	terminal step of the panel
σ	System	index	index	survey-period (quarter) index; $\sigma(t)$ the period containing t
θ	System	scalar	constant	unit steps per survey period (3, one quarter)
e	System	index	index	economy index (US, CA)
i, j	System	index	index	household-group index / neighbour or summation dummy
N	System	scalar	hyperparam.	number of weighted household groups (32)
f	System	index	index	profile-attribute index (age, income, net worth)
Problem and state (Eqs. 1, 2)				
$\mathcal{H}$	System	set	constant	observed history of macro indicator series
Φ	System	operator	operator	the simulator, $\hat{\mathbf{Y}}^e = \Phi(\mathcal{H}, S_0^e)$
$\mathbf{Y}^e$, $\hat{\mathbf{Y}}^e$	System	matrix	variable	actual / simulated monthly panel over the evaluation window
S_t^e	All	set	variable	settled economy state $(\mathcal{A}_t, \mathcal{B}_t, \mathbf{q}_t)$
$\mathcal{A}_t$	Household	set	variable	household atlas $\{\chi_i, s_{i,t}, \mu_{i,t}, \omega_i\}_{i=1}^N$
χ_i	Household	vector	constant	static group profile (age, income, wealth)
$s_{i,t}$	Household	vector	variable	time-varying group state (expectations, anchors, balances)
$\mu_{i,t}$	Household	set	variable	rolling memory of realized outcomes
ω_i	Household	scalar	constant	population weight, $\sum_i \omega_i = 1$

Table 4 (continued)

Symbol	Sector	Kind	Role	Meaning
$\mathcal{B}_t$	Institutions	vector	variable	institutional balance sheets (B_t^{firm}, B_t^{bank}, B_t^{gov}, B_t^{cb})
$\mathbf{q}_t$	System	vector	variable	environment vector (goods price, wage, loan rate)
$\mathcal{I}_{i,t}$	Household	set	variable	information set group i reads at the head of step t
Network (§3.3, Appendix C.2)				
$\mathcal{G}$	System	graph	constant	static small-world observation network
$\mathcal{N}_{\mathcal{G}}(i)$	System	set	constant	neighbours of group i in $\mathcal{G}$
$d(i,j)$	System	scalar	operator	homophily distance $\sum_f \lvert[\chi_i]_f - [\chi_j]_f\rvert$
k	System	scalar	hyperparam.	nearest neighbours per group
p^*	System	scalar	hyperparam.	rewiring probability, calibrated to the H benchmark
$\bar{\mathbf{s}}_{i,t}^{\mathcal{G}}$	Household	vector	variable	neighbour signal aggregated over $\mathcal{N}_{\mathcal{G}}(i)$
Unit-step pipeline (Eq. 3)				
U	Household	operator	operator	information-set update
h_i	Household	string	constant	prompt template of group i
$\mathbf{x}_{i,t}^{\text{raw}}$	Household	vector	variable	raw LLM intention (seven fields of Eq. 20)
$\mathbf{x}_{i,t}$	Household	vector	variable	governed intention after the four gates
G	Household	operator	operator	governance operator (Appendix C.4)
Π	Household	operator	operator	projection to feasible actions (Appendix C.4)
$\mathbf{a}_{i,t}$	Household	vector	variable	feasible action $(c_{i,t}, \hat{b}_{i,t}, \ell_{i,t}, f_{i,t}, h_{i,t})$
$\text{Credit}\circ\text{Goods}\circ\text{Labor}$	System	operator	operator	settlement kernel, executed Labor $\rightarrow$ Goods $\rightarrow$ Credit
push	Household	operator	operator	memory append of the step's realized experience
Policy pipeline (Eqs. 8–10)				
$\mathbf{p}_\sigma$	Gov/C.B.	vector	variable	policy instruments $(r_\sigma, \tau_\sigma^{\text{inc}}, \tau_\sigma^{\text{c}}, g_\sigma, z_\sigma)$
P	Gov/C.B.	operator	operator	policy cognition composing survey and macro conditions
Q	System	operator	operator	weighted questionnaire survey function
ϕ	Household	operator	operator	questionnaire field extractor from state and memory
D	Gov/C.B.	operator	operator	policy decision rule (LLM / Rule / RL instantiations)
$\boldsymbol{\delta}_\sigma$	Gov/C.B.	vector	variable	bounded instrument increments
$\mathcal{P}$, $\text{clip}_{\mathcal{P}}$	Gov/C.B.	set / operator	constant / operator	admissible policy box and its clip
Readout				
R	System	operator	operator	readout of the target series off the settled state S_t^e
$\hat{\mathbf{y}}_t^e$	System	vector	variable	simulated panel row at step t

Table 5: Appendix notation (governance, projection, settlement, network, policy agents).

Symbol	Sector	Kind	Role	Meaning
Governance (Appendix C.4)				
m	Household	index	index	intention-field index, the seven fields of Eq. 20
clip	System	operator	operator	clamp $\min\{b, \max\{a, x\}\}$
$\underline{x}_m$, $\overline{x}_m$	Household	scalar	hyperparam.	field admissible bounds; $\mathcal{X}_m$ the box, $\mathcal{X} = \prod_m \mathcal{X}_m$
$\mathrm{med}_{t,m}$, $\mathrm{MAD}_{t,m}$	Household	scalar	variable	population-weighted cross-sectional median / median absolute deviation of the batch
$\eta_{t,m}$	Household	scalar	variable	winsorization half-width
κ_{mad}	Household	scalar	hyperparam.	winsorizing multiplier
d_m	Household	scalar	hyperparam.	deadband: moves below it count as no move
ψ_m	Household	scalar	hyperparam.	pass-three smoothing weight
λ_m, $\bar{\lambda}_m$	Household	scalar	hyperparam.	per-group / batch quarterly rate limits
ρ	Household	scalar	hyperparam.	batch mean-reversion rate
$\bar{x}^{(3)}_{t,m}$, $\bar{x}_{t-1,m}$, $\bar{x}^0_m$	Household	scalar	variable	cohort means: pass three / previous step / origin-calibrated
$\Delta^{\mathrm{b}}_{t,m}$	Household	scalar	variable	batch-level correction shift
Projection (Appendix C.4)				
c^a_i	Household	scalar	constant	consumption anchor of group i
$B^{\mathrm{liq}}_{i,t}$, $B^{\mathrm{res}}_{i,t}$	Household	scalar	variable	liquid budget / liquidity remaining after consumption
$y_{i,t}$	Household	scalar	variable	annualized income
$y^{\mathrm{res}}_{i,t}$	Household	scalar	variable	current-step income left after consumption, floored at zero
$\hat{b}_{i,t}$	Household	scalar	variable	credit application
$f_{i,t}$, $h_{i,t}$	Household	scalar	variable	financial-asset / housing purchases
$f^{\mathrm{d}}_{i,t}$, $h^{\mathrm{d}}_{i,t}$	Household	scalar	variable	desired purchase amounts before the joint liquidity cap
$L^{\mathrm{acc}}_{i,t}$	Household	scalar	variable	accessible housing liquidity (deposits plus κ^{hs} of housing)
κ^{hs}	Household	scalar	hyperparam.	housing share in L^{acc}
$\underline{x}^{\mathrm{work}}$, $\overline{x}^{\mathrm{work}}$	Household	scalar	hyperparam.	work-intensity admissible bounds (0 and 1.5, one is normal effort)
λ_{LTV}	Bank	scalar	hyperparam.	loan-to-value limit
$V^{\mathrm{coll}}_{i,t}$	Household	scalar	variable	collateral value (housing plus liquid assets)
ϵ	Household	scalar	hyperparam.	consumption-propensity elasticity
L^{res}	Household	scalar	hyperparam.	liquidity reserve horizon (months)
Settlement: labor (Appendix C.5)				
$\tilde{D}_t$	System	scalar	variable	total demand (consumption + investment + government spending)
$\tilde{Y}_t$	System	scalar	variable	trend output, n^{tr}-month moving average
δ_t	System	scalar	variable	demand-pressure ratio $\tilde{D}_t/\tilde{Y}_t$, bounded; one means normal pace
e^*_t	System	scalar	variable	employment quota target
e_0	System	scalar	constant	origin employment rate anchoring the quota
κ_u	System	scalar	hyperparam.	labor-demand sensitivity
$e^{\min}, e^{\max}, \delta^{\min}, \delta^{\max}$	System	scalar	hyperparam.	quota and demand-ratio bounds
$\mathrm{score}_{i,t}$	Household	scalar	variable	employment score $\kappa_{\mathrm{emp}} p^{\mathrm{emp}}_{i,t-1} + \kappa_{\mathrm{work}} x^{\mathrm{work}}_{i,t}$
κ_{emp}, κ_{work}	Household	scalar	hyperparam.	employment-score weights (record / stated intensity, summing to one)

Table 5 (continued)

Symbol	Sector	Kind	Role	Meaning
ν	System	scalar	hyperparam.	quota-fill boundary tolerance (score points)
$p_{i,t}^{\mathrm{emp}}$	Household	scalar	variable	employment probability
Δe, $p^{\min}$, $p^{\max}$	Household	scalar	hyperparam.	employment-probability transition step and bounds
A_t	Firm	scalar	variable	labor productivity
L_t	Firm	scalar	variable	total labor, weight-aggregated intended hours of employed groups
Settlement: goods				
Y_t	Firm	scalar	variable	output $\min(\tilde{D}_t, A_t L_t)$
C_t	Household	scalar	variable	realized aggregate household consumption
ration_t	System	scalar	variable	consumption ration factor
g_t	Government	scalar	variable	government expenditure (monthly, along the policy path)
ζ	System	scalar	hyperparam.	monthly price-change clip
π^*	System	scalar	hyperparam.	inflation anchor target
ω	Household	scalar	hyperparam.	expectation reversion rate toward π^*
n^{tr}, n^{util}	System	scalar	hyperparam.	trend / utilization-benchmark window lengths (months)
Settlement: credit and the firm				
r_t^{ℓ}	Bank	scalar	variable	loan rate, set at the head of the step (last entry of $\mathbf{q}_t$)
β_0, β_1, β_2	Bank	scalar	hyperparam.	loan-spread markup / NPL loading / unemployment premium
u^{nat}	System	scalar	hyperparam.	natural-rate unemployment threshold
npl_t	Bank	scalar	variable	nonperforming-loan ratio
u_t	System	scalar	variable	simulated unemployment rate
$\beta_0^{\mathrm{c}}, \ldots, \beta_6^{\mathrm{c}}$	Bank	scalar	hyperparam.	credit-logit constants: intercept, DTI, liquidity, liquidity cap, employment, unemployment, approval cutoff
$\mathrm{dti}_{i,t}$, $\mathrm{liq}_{i,t}$	Household	scalar	variable	debt-to-income / deposits-to-income ratios
κ^{dep}, κ^{eq}, κ^{npl}	Bank	scalar	hyperparam.	bank-funding weights: deposits, equity, NPL penalty
κ^{cap}, κ^{nw}	Firm	scalar	hyperparam.	firm-application investment share / net-worth buffer
$\hat{b}_t^{\mathrm{firm}}$	Firm	scalar	variable	firm credit application
V_t^{firm}	Firm	scalar	variable	firm net worth
F_t	Bank	scalar	variable	loanable funds: deposits + equity − NPL deduction
φ_t	Bank	scalar	variable	funding-ration factor applied to approved applications
dep_t, eq_t	Bank	scalar	variable	bank deposits / equity
$\mathcal{J}_t$	Bank	set	variable	approved credit applicants
I_t^{d}, I_t^*, I_t^{base}	Firm	scalar	variable	desired / target / baseline investment
v_t	Firm	scalar	variable	investment response ratio, $v_t \in [v^{\min}, v^{\max}]$
v_t^{pull}	Firm	scalar	variable	demand-side investment pull (output gap + fiscal stance)
v_t^{drag}	Firm	scalar	variable	credit-cost investment drag (loan rate above reference)
κ_I, κ_γ, κ_ℓ, κ_G	Firm	scalar	hyperparam.	investment sensitivities: speed, output gap, loan rate, fiscal stance
r^{ref}	Firm	scalar	hyperparam.	investment reference loan rate
γ_{t-1}	System	scalar	variable	lagged simulated output gap

Table 5 (continued)

Symbol	Sector	Kind	Role	Meaning
$\Delta_{g,t}^{12}$	Government	scalar	variable	annual growth rate of government expenditure
Balance-sheet update				
δ	Firm	scalar	hyperparam.	capital depreciation rate
δ^{npl}	Bank	scalar	hyperparam.	NPL monthly decay factor
W_t	Firm	scalar	variable	wage bill
π_t^{firm}	Firm	scalar	variable	firm profit
K_t	Firm	scalar	variable	capital stock
$\boldsymbol{\xi}_t$	All	vector	variable	cleared transaction records feeding the sector tables
Network diagnostics (Appendix C.2)				
H	System	scalar	statistic	cross-stratum edge ratio; $H = 1$ random mixing, benchmark 0.528
HR_f	System	scalar	statistic	homophily ratio on attribute f; one is random mixing
$\mathcal{E}$	System	set	constant	realized edge set of $\mathcal{G}$
Policy agents (Appendix D.1)				
α^{rl}, γ^{rl}, ϵ^{rl}	Gov/C.B.	scalar	hyperparam.	Q-learning learning rate, discount, exploration

B.2 Metric Definitions

Define $\hat{\mathbf{X}}_c^e$ as the simulated sequence of indicator c (GDP, consumption) and $\mathbf{X}_c^e$ as the actual sequence. The monthly log growth transform gives the actual and simulated growth sequences $\mathbf{g}_c^e$ and $\hat{\mathbf{g}}_c^e$,

$$g_{c,t}^e \;=\; 100\,\ln\!\big(X_{c,t}^e / X_{c,t-1}^e\big), \qquad \hat{g}_{c,t}^e \;=\; 100\,\ln\!\big(\hat{X}_{c,t}^e / \hat{X}_{c,t-1}^e\big), \qquad t = 1, \ldots, T, \tag{11}$$

The adoption of logarithmic growth metrics is predicated upon the following considerations: (1) a single erroneous estimation will not compromise subsequent predictive performance through directional bias; (2) it eliminates spurious correlations inherent in trend-dominant models such as linear specifications, thereby incentivizing models to simulate the direction and rate of change in subsequent sequences; (3) multiple economic simulation studies have employed this object as a measurement target (Dosi et al., 2015; Gualdi et al., 2017; Reissl, 2021; Caiani et al., 2018; Poledna et al., 2023).

We report the mean and standard deviation of each metric for every sequence c and economy e as one data point. The evaluation metrics are defined as follows.

Pearson.

$$\rho_c \;=\; 100 \cdot \mathrm{corr}\big(\hat{\mathbf{g}}_c^e,\, \mathbf{g}_c^e\big). \tag{12}$$

The percent-scaled correlation between the simulated and actual growth sequences: it rewards matching the direction and timing of month-on-month change; larger is better.

JSD.

$$\mathrm{JSD}(P, Q) \;=\; \tfrac{1}{2}\, D_{\mathrm{KL}}(P \,\|\, M) + \tfrac{1}{2}\, D_{\mathrm{KL}}(Q \,\|\, M), \quad D_{\mathrm{KL}}(P \,\|\, M) = \textstyle\sum_b P_b \log_2 \frac{P_b}{M_b}, \tag{13}$$

where P and Q are the bin frequencies of $\hat{\mathbf{g}}_c^e$ and $\mathbf{g}_c^e$ on $B = 4$ equal-width bins over their combined range and $M = \frac{1}{2}(P + Q)$. It measures the distributional fidelity of the growth sequence.

RMSE.

$$\mathrm{RMSE}^{(k)} = \Big(\tfrac{1}{4} \textstyle\sum_{(c,s)\in\mathcal{GC}} \frac{1}{|T|} \sum_{t\in T} \big(\hat{X}_{c,s,t}^{e,(k)} - X_{c,s,t}^e\big)^2\Big)^{1/2}, \tag{14}$$

where $\mathcal{GC}$ collects the four GDP and consumption cells of the two economies and the indices carry the 2023Q4-mean origin of 100. Reported as the mean $\pm$ standard deviation over seeds, in index points, it measures level accuracy on the two series that matter most for the assessment of living standards.

Terminal CV.

$$\mathrm{CV}_c = \mathrm{SD}_k\big(\hat{X}_{c,T}^{e,(k)}\big) / \overline{\hat{X}}_{c,T}^{e}, \tag{15}$$

with $\overline{\hat{X}}_{c,T}^{e}$ the cross-seed mean of the terminal index: it measures how reproducible the simulated endpoint is across the twenty stochastic runs. The superscript (k) denotes the k-th result under different seeds.

Seed RMSE SD.

$$\mathrm{RMSE}_c^{(k)} = \Big(\tfrac{1}{|T_c^{\mathrm{obs}}|}\textstyle\sum_{t\in T_c^{\mathrm{obs}}}\big(\hat{X}_{c,t}^{e,(k)} - X_{c,t}^{e}\big)^2\Big)^{1/2}, \tag{16}$$

computed over all the observed months T_c^{obs}.

B.3 Hyperparameter Settings

Table 6 lists the hyperparameters of the governed GAME system, their notation as used in the text, and their instantiated values in the reported experiments.

Table 6: Hyperparameters of the governed GAME system.

Symbol	Description	Value	Source
N	number of household agents	32	compute budget
θ	unit steps per survey period	3	calendar quarter
κ_μ	rolling-memory length (steps)	24	design
n_{inc}	income stratification bins	4	design
n_{nw}	net-worth stratification bins	8	design
$\lvert\mathbf{x}_{i,t}\rvert$	governed intention fields per agent	7	design
k	nearest neighbours in $\mathcal{G}$	6	design
p^*	rewiring probability in $\mathcal{G}$	0.125	calibrated
$\bar{\lambda}_m$	batch rate limit per field	per field (0.003–0.05)	design
ρ	batch mean reversion rate	0.25	design
ψ_m	temporal smoothing weight (all fields)	0.50	design
κ_{mad}	cross-group winsorizing multiplier	4.0	design
$\underline{x}_m, \overline{x}_m$	governance hard bounds per field	per field (work $[0.25, 1.25]$)	design
$\lambda_m,\ d_m$	per-group quarterly rate limits; deadbands	per field	design
L^{res}	liquidity reserve horizon (months)	12	observed
$\underline{x}^{\mathrm{work}}, \overline{x}^{\mathrm{work}}$	projection-side work-intensity range	0, 1.5	design
κ^{hs}	housing share in accessible housing liquidity	0.02	design
$\kappa^{\mathrm{dti}}, \lambda_{\mathrm{LTV}}$	credit-capacity coefficients: income share, housing share	0.45, 0.50	design
ϵ	consumption propensity elasticity	0.25	design
$w_{\min}$	edge-weight floor in $\mathcal{G}$	0.05	design
β_w, β_n	wage / nonlabor pass-through	0.0, 0.75	design
$\beta_0, \beta_1, \beta_2$	loan spread: markup / NPL loading / unemployment premium	0.025, 0.35, 0.12	design
$\beta_0^{\mathrm{c}}, \ldots, \beta_6^{\mathrm{c}}$	credit logit: intercept, DTI, liquidity, liquidity cap, employment, unemployment, approval cutoff	2.20, 1.80, 0.30, 3.0, 0.50, 3.50, 0.35	design
$\kappa^{\mathrm{dep}}, \kappa^{\mathrm{eq}}, \kappa^{\mathrm{npl}}$	bank-funding weights: deposits, equity, NPL penalty	0.08, 0.04, 0.50	design

Table 6 (continued)

Symbol	Description	Value	Source
$\kappa^{\mathrm{cap}}, \kappa^{\mathrm{nw}}$	firm credit application: investment share, net-worth buffer	0.35, 0.20	design
$\underline{r}^{\ell}, \overline{r}^{\ell}$	loan-rate bounds	0.015, 0.24	design
u^{nat}	macro-risk unemployment threshold	0.06	design
r^{ref}	investment reference loan rate	0.045	design
κ_{γ}	investment gap sensitivity	US 0.10 / CA 1.04	SMM, per economy
κ_{ℓ}	investment rate sensitivity	US 3.85 / CA 2.52	SMM, per economy
κ_I	investment adjustment speed	US 0.15 / CA 0.96	SMM, per economy
$v^{\min}, v^{\max}$	investment response bounds	0.35, 1.5	design
κ_G	fiscal crowding-in elasticity	0.5	training-window median
δ	capital depreciation rate	0.008/mo	design
δ^{npl}	NPL monthly decay factor	0.97	design
π^*	inflation anchor target	2% p.a.	policy convention
ζ	monthly price-change clip	0.93%	calibration-window data
$n^{\mathrm{tr}}, n^{\mathrm{util}}$	trend MA / utilization-benchmark windows (months)	12, 36	design
ω	expectation reversion to π^*	0.093/mo	design
κ_u	labor-demand sensitivity (floor)	0.30	design
$e^{\min}, e^{\max}$	employment quota bounds	0.05, 0.98	design
$\delta^{\min}, \delta^{\max}$	demand-trend ratio bounds	0.85, 1.15	design
$\kappa_{\mathrm{emp}}, \kappa_{\mathrm{work}}$	employment-score weights: record, stated intensity	0.65, 0.35	design
ν	boundary tolerance of the quota fill (score points)	0.15	design
$\alpha^{\mathrm{rl}}, \gamma^{\mathrm{rl}}, \epsilon^{\mathrm{rl}}$	Q-learning: learning rate, discount, exploration	0.2, 0.95, 0.1	design
Δe	employment-probability transition	$\pm$0.08/step	design
$p^{\min}, p^{\max}$	employment-probability bounds	0.02, 0.98	design

The stratification grid size $N = n_{\mathrm{inc}} \times n_{\mathrm{nw}}$ and the neighbour count k are compute-budget parameters: each agent incurs one LLM call per quarter. The rewiring probability p^* is calibrated to the homophily benchmark, the three investment sensitivities ($\kappa_{\gamma}, \kappa_{\ell}, \kappa_I$) by SMM per economy, and the fiscal elasticity κ_G from the training-window fiscal-elasticity median.

C Method Details

C.1 Definitions

Here unfolds the two composite objects of the state definition (Eq. 2a) that the main text introduces compactly.

Institutional balance sheets $\mathcal{B}_t$.

$$\mathcal{B}_t \;=\; \big(\, B_t^{\mathrm{firm}},\;\; B_t^{\mathrm{bank}},\;\; B_t^{\mathrm{gov}},\;\; B_t^{\mathrm{cb}} \,\big), \tag{17}$$

where B_t^{firm} is the firm's balance sheet (capital, loans, net worth), B_t^{bank} the banking sector's (household loans, firm loans, government bonds, reserves, deposits, equity), B_t^{gov} the government's (debt, tax rates, expenditure and transfer obligations), and B_t^{cb} the central bank's (government bond holdings, bank lending, currency, reserves).

Environment vector $\mathbf{q}_t$.

$$\mathbf{q}_t \;=\; \big(\, P_t,\;\; w_t,\;\; r_t^{\ell} \,\big), \tag{18}$$

where P_t is the goods price, w_t the wage rate, r_t^{ℓ} the bank loan rate set at the head of step t.

C.2 Network Construction and Diagnostics

Diagnostics definitions. Four statistics describe the realized graph. The **rewiring probability** p^* is the share of candidate edges redirected to a uniformly random agent during construction. The **cross-stratum edge ratio** H is the share of edges that cross income quartiles, observed over its expected value under random mixing: $H = 1$ means connections are income-blind, $H < 1$ means agents preferentially connect within their own income stratum. The calibration sets p^* by bisection so that the realized H matches the empirical benchmark $H = 0.528$ reported for a large social network by Chang et al. (2025). The **small-world signature** is the joint occurrence of high clustering (the probability that two neighbours of a node are themselves connected) and short average path length (the mean minimum hop count over all node pairs). The **homophily ratio** of dimension $f \in \{\text{income}, \text{age}, \text{net worth}\}$ is

$$\mathrm{HR}_f \;=\; \frac{\frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} |\chi_{i,f} - \chi_{j,f}|}{\frac{1}{\binom{N}{2}}\sum_{i<j} |\chi_{i,f} - \chi_{j,f}|}, \tag{19}$$

where $\mathcal{E}$ is the realized edge set and $\chi_{i,f}$ the normalized attribute of agent i on dimension f: the numerator is the mean attribute difference between *connected* pairs, the denominator the same mean over *all possible* pairs, so $\mathrm{HR}_f = 1$ indicates random mixing and values below one increasingly strong same-type connection.

Realized structure. The calibration converges to $p^* = 0.125$, and the realized cross-stratum ratio $H = 0.52$–0.53 matches the 0.528 target. The graphs retain the small-world signature: mean degree 6.8–6.9, clustering coefficient 0.48–0.53, and average shortest path length 1.74–1.79. The homophily ratios are below one on all three dimensions in every economy, income 0.40–0.41, age 0.70–0.74, net worth 0.86–0.87.

Sensitivity of the distance weights. The equal-weight distance is left uncalibrated by design. The sensitivity check confirms the choice does not drive the network statistics: perturbing each dimension weight by a factor of two in either direction, with p^* recalibrated to the benchmark each time, leaves the diagnostics essentially unchanged (clustering 0.45–0.51, average path length 1.69–1.79, and the homophily ratios within 0.05 of the equal-weight values).

C.3 Agent Prompt Templates

This appendix reproduces the household LLM prompt templates exactly as implemented in the reported configuration, with runtime values shown as {placeholder} fields matching the code's f-string variables. Household prompts address one weighted household group: monetary amounts are rendered on the calibrated simulation accounting scale and are directly comparable across groups; rates render as percentages with fixed decimals, and missing values as "not available". Qualitative phrases are deterministic bucketings of the corresponding numeric fields.

C.3.1 Household System Prompt

Household system prompt (life-scenario)

```
You are one household deciding for your own family for the next
quarter, the way a family would at its kitchen table. Decide only
behavioral intentions; you cannot execute actions or change accounts.
No calendar date or country identity is disclosed. Treat behavior as
persistent: change your plan only when your situation justifies it,
and avoid alternating large corrections between similar months.
Reason about money the way a real family does - bills, buffers, debts,
prices. Return one JSON object with exactly this schema:
{"monthly_spending_units": "number: units per month you plan to
spend this quarter", "extra_saving_units": "number: extra units
put aside each month beyond your usual saving", "new_credit_units":
"number: units of new borrowing for the whole quarter (0 if none)",
"credit_purpose": "short phrase or none", "housing_budget_units":
"number: units into housing this quarter, renovation or down
```

```
payment (0 if none)", "inflation_expectation_pct": "number: your
guess for next year's price rise in percent", "job_risk_pct":
"number: chance you lose your job within a year, in percent",
"work_intensity": "number: 1.0 = normal hours; 1.1 = extra
shifts; 0.9 = cutting back", "plan": "one or two sentences in
your own words, like telling your family"}
Numbers must be finite and in range. Return only the JSON object.
```

Household-unit scaling. Monetary amounts in the life-scenario prompt are rendered in a normalized unit system: the group's normal monthly income (labor income after unemployment risk plus non-labor income) is mapped to a fixed 5,000 units, and all other amounts are scaled by the same factor. This makes the prompt readable as plain family money without disclosing any absolute scale. Smaller numbers also help the LLM compute accurately, since the group-scale budget is very large. These numbers will be mapped back to their original scale when generating the intent vector.

Intention generation. The raw intention vector is

$$\mathbf{x}_{i,t}^{\text{raw}} = \left(x_{i,t}^{\text{con}},\ x_{i,t}^{\text{save}},\ x_{i,t}^{\text{cred}},\ x_{i,t}^{\text{house}},\ \pi_{i,t}^{e},\ u_{i,t}^{e},\ x_{i,t}^{\text{work}} \right), \tag{20}$$

where $x_{i,t}^{\text{con}}$ is the consumption propensity applied to the anchor c_i^a (obtained as one plus the signed consumption deviation from the prompt), $x_{i,t}^{\text{save}}$ the share of post-consumption income added to financial assets, $x_{i,t}^{\text{cred}}$ the desired-credit rate on annualized income, $x_{i,t}^{\text{house}}$ the housing-investment rate on accessible liquidity, $\pi_{i,t}^{e}$ the twelve-month inflation expectation, $u_{i,t}^{e}$ the perceived unemployment risk, and $x_{i,t}^{\text{work}}$ the work intensity (1=normal effort) that scales the intended hours entering the employment-quota ranking. Out-of-range or non-finite fields raise a validation error and are counted in the reported LLM validation rates.

C.3.2 Household User Prompt Template

Household user prompt template (life-scenario)

```
YOUR HOUSEHOLD AND YOUR LIFE THIS QUARTER

You are one household deciding for the next few months. No calendar
date or country identity is disclosed; units are simply your
family's money.

YOUR FAMILY:
- You are about {age} years old, {family_text}, around income
  percentile {income_percentile} and wealth percentile
  {wealth_percentile}.
- Your job: about {employment_probability} chance you keep working.
- A normal month brings in about {normal_income} units ({labor_income}
  from work after job risk, {nonlabor_income} other income). THIS
  month you actually earned {actual_income} units ({income_change_pct}%
  versus a normal month).

WHAT LIFE COSTS YOU (typical prices in your area, at your standard
of living):
- Rent for a modest home like yours: about {rent} units/month;
  utilities and phone about {utilities}.
- Groceries: about {groceries} units/month (a year ago the same
  basket cost about {groceries_year_ago}; prices {inflation_pct}%
  over the year).
- Getting around (transit/fuel): about {transit} units/month.
- Other essentials (health, clothing, basics): about {other} units/month.
- Essential bills come to roughly {essential_total} units; your normal
  total spending is about {normal_spending} units, leaving about
  {discretionary} for discretionary things.
- For scale: a basic new car costs about {car_price} units; a
  refrigerator about {fridge_price}; a modest home in your area
  sells for about {home_price} units.

YOUR MONEY AND DEBTS:
```

```
- Checking account: {deposits} (about {liquidity_months} months of
  your normal spending).
- Other savings: {other_savings}. {home_ownership}
- Total debt: {debt} units - about {debt_months} months of income
  ({debt_pct}% of a year's income). Minimum interest payments about
  {interest_month} units/month. Credit recently denied:
  {credit_denied}.
- Interest rates: new loans cost about {loan_rate}% a year; the
  central bank rate is {policy_rate}%; your savings earn almost
  nothing.

HOW THINGS HAVE BEEN GOING:
- Last month you earned {last_income} and spent {last_consumption}.
  {recent_events}
- Neighbors like you: {neighbor_employment} working, spending
  {neighbor_consumption}x normal, expect inflation
  {neighbor_inflation}, money stress {neighbor_stress}.

THE ECONOMY YOU LIVE IN:
- Prices rose {inflation_yoy}% over the past year; the output gap is
  {output_gap}; unemployment in your region is {unemployment}.
- Central bank communication: {communication}

YOUR DECISION FOR THIS QUARTER - answer in plain money terms (the
simulation converts your answers into model parameters):
Decide your monthly spending (normal is about {normal_spending}
units), extra monthly saving, any new borrowing this quarter and its
purpose, any housing budget, your expected price rise and job risk,
and your work effort (1.0 = normal hours).
Return exactly one JSON object with this schema. Return only the
JSON object.
```

Durable and housing price anchors (car, refrigerator, home) are multiplied by the simulation's own price index relative to the origin month, so typical prices move with the model's inflation; the grocery year-ago line uses realized year-over-year inflation. The macro block reads the simulator's own lagged aggregates; qualitative phrases are deterministic bucketings of the displayed numbers.

C.4 Governance Details

Governance G. This paragraph gives the exact functional form of the governance operator G of Eq. 3c. The operator screens the batch $\{\mathbf{x}_{i,t}^{\mathrm{raw}}\}_{i=1}^{N}$ field by field through three ordered per-group passes and one batch-level correction:

$$x_{i,t,m}^{(1)} = \mathrm{clip}\big(x_{i,t,m}^{\mathrm{raw}};\, \underline{x}_m, \overline{x}_m\big), \quad x_{i,t,m}^{(2)} = \mathrm{clip}\big(x_{i,t,m}^{(1)};\, \mathrm{med}_{t,m} \pm \eta_{t,m}\big), \tag{21a}$$

$$x_{i,t,m}^{(3)} = \mathrm{clip}\big((1-\psi_m)\, x_{i,t-1,m} + \psi_m\, x_{i,t,m}^{(2)};\; x_{i,t-1,m} \pm \lambda_m\big), \tag{21b}$$

$$\Delta_{t,m}^{\mathrm{b}} = (1-\rho)\,\mathrm{clip}\big(\bar{x}_{t,m}^{(3)};\, \bar{x}_{t-1,m} \pm \bar{\lambda}_m\big) + \rho\,\bar{x}_m^{0} - \bar{x}_{t,m}^{(3)}, \tag{21c}$$

$$\mathbf{x}_{i,t} = G\big[\{\mathbf{x}_{j,t}^{\mathrm{raw}}\}_{j=1}^{N},\, \mathbf{x}_{i,t-1}\big] = \mathrm{clip}_{\mathcal{X}}\big(\mathbf{x}_{i,t}^{(3)} + \boldsymbol{\Delta}_t^{\mathrm{b}}\big). \tag{21d}$$

Throughout, m indexes all the intention fields of $\mathbf{x}_{i,t}$. $\mathrm{clip}(x;\, a, b) = \min\{b, \max\{a, x\}\}$ clamps a scalar to the interval $[a, b]$. $\underline{x}_m$ and $\overline{x}_m$ are the economically meaningful lower and upper limits of field m, so $\mathcal{X}_m = [\underline{x}_m, \overline{x}_m]$ is the field-wise admissible box. $\mathrm{med}_{t,m}$ and $\mathrm{MAD}_{t,m}$ are the population-weighted cross-sectional median and median absolute deviation of the raw batch $\{x_{i,t,m}^{\mathrm{raw}}\}_{i=1}^{N}$ on field m; $\eta_{t,m} = \max\{\kappa_{\mathrm{mad}}\,\mathrm{MAD}_{t,m},\, 2d_m\}$ is the resulting winsorization half-width, where κ_{mad} is a constant fixing the band at that many deviations and d_m is the deadband of field m, the threshold below which a proposed move counts as no move. $\psi_m \in (0, 1]$ is the smoothing weight that splits pass three between this anchor and the new proposal; and λ_m is the per-field rate limit, the largest change field m may undergo per elapsed quarter. Cohort quantities are population-weighted with the group weights ω_j: $\bar{x}_{t,m}^{(3)} = \sum_{j=1}^{N} \omega_j\, x_{j,t,m}^{(3)}$ is the pass-three cohort mean, $\bar{x}_{t-1,m}$ is the previous period's governed cohort mean, $\bar{x}_m^{0}$ is the calibrated cohort mean at the origin, $\bar{\lambda}_m$ is the rate limit imposed on the cohort mean itself, $\rho \in [0, 1)$ is the mean-reversion

rate toward $\bar{x}^0_m$, and $\mathbf{\Delta}^{\mathrm{b}}_t$ stacks the per-field batch shifts $\Delta^{\mathrm{b}}_{t,m}$ into a single vector. Eq. 21a applies the economic bounds and the robust cross-group shrinkage: winsorizing at the median rather than the mean keeps the band itself uncontaminated by the outlier it is meant to catch, so one aberrant group cannot drag the cohort, and the $2d_m$ floor keeps the band from collapsing when proposals agree. Eq. 21b tempers each field by partial adjustment from its last governed value and caps the move per elapsed quarter, so a single cognition event can trigger neither a jump nor a burst. Eq. 21d is the definition of G: it reads the whole batch (the median and MAD of Eq. 21a are cross-sectional statistics) and returns the shifted, re-projected intentions. Eq. 21c is the batch-level correction, and its rationale is the correlated-drift failure mode of LLM cohorts: every group can move a little in the same direction, passing every per-group limit of Eq. 21b while still swinging aggregate demand. The shift caps the cohort-mean move at $\bar{\lambda}_m$, reverts it toward the calibrated mean $\bar{x}^0_m$ at rate ρ (so sustained one-directional drift cannot accumulate into a unit root), and applies the correction equally to all groups, capping batch drift without disturbing cross-group heterogeneity. Quantization to a fixed output grid and a deadband d_m that snaps residual moves back to $x_{i,t-1,m}$ complete the per-group pass (the governed cohort mean being $\bar{\mathbf{x}}_t = \sum_{i=1}^{N} \omega_i \, \mathbf{x}_{i,t}$).

Projection Π**.** The projection maps the governed intention vector $\mathbf{x}_{i,t}$ into the feasible action set:

$$\mathbf{a}_{i,t} = \Pi(\mathbf{x}_{i,t}) = \begin{pmatrix} c_{i,t} \\ \hat{b}_{i,t} \\ \ell_{i,t} \\ f_{i,t} \\ h_{i,t} \end{pmatrix} = \begin{pmatrix} \min\big[x^{\mathrm{con}}_{i,t}\, c^a_i,\ \ B^{\mathrm{liq}}_{i,t}\big] \\ \min\big[x^{\mathrm{cred}}_{i,t}\, y_{i,t},\ \ \lambda_{\mathrm{LTV}}\, V^{\mathrm{coll}}_{i,t}\big] \\ \mathrm{clip}\big(x^{\mathrm{work}}_{i,t};\ \underline{x}^{\mathrm{work}},\ \overline{x}^{\mathrm{work}}\big) \\ \min\big\{1,\ \frac{B^{\mathrm{res}}_{i,t}}{f^{\mathrm{d}}_{i,t}+h^{\mathrm{d}}_{i,t}}\big\}\, f^{\mathrm{d}}_{i,t} \\ \min\big\{1,\ \frac{B^{\mathrm{res}}_{i,t}}{f^{\mathrm{d}}_{i,t}+h^{\mathrm{d}}_{i,t}}\big\}\, h^{\mathrm{d}}_{i,t} \end{pmatrix}, \tag{22}$$

where the five rows give the five components of the action vector, namely consumption demand $c_{i,t}$, the credit application $\hat{b}_{i,t}$, labor supply $\ell_{i,t}$, and the financial-asset and housing purchases $f_{i,t}$ and $h_{i,t}$. The first row caps desired consumption $x^{\mathrm{con}}_{i,t}\, c^a_i$, the governed propensity applied to the anchor, at the liquid budget $B^{\mathrm{liq}}_{i,t}$, the group's after-tax labor, non-labor, and transfer income plus the deposits it may draw down this step, net of required debt service and its minimum deposit reserve. The second row caps the desired credit application, the governed rate $x^{\mathrm{cred}}_{i,t}$ applied to the group's annualized income $y_{i,t}$, at the loan-to-value limit λ_{LTV} times the collateral value $V^{\mathrm{coll}}_{i,t}$, defined as housing plus liquid assets; approval and funding rationing of these applications take place later, in the credit market of Appendix C.5. The third row bounds the work intensity to its admissible range $[\underline{x}^{\mathrm{work}}, \overline{x}^{\mathrm{work}}]$, one being normal effort. The last two rows enforce the joint liquidity cap on asset purchases: the desired amounts are $f^{\mathrm{d}}_{i,t} = x^{\mathrm{save}}_{i,t}\, y^{\mathrm{res}}_{i,t}$, with $y^{\mathrm{res}}_{i,t}$ the current-step income left after consumption and floored at zero, and $h^{\mathrm{d}}_{i,t} = x^{\mathrm{house}}_{i,t}\, L^{\mathrm{acc}}_{i,t}$, with $L^{\mathrm{acc}}_{i,t}$ the drawable deposits plus a share κ^{hs} of housing wealth. Both desired amounts are scaled by the common factor $\min\{1,\ B^{\mathrm{res}}_{i,t}/(f^{\mathrm{d}}_{i,t} + h^{\mathrm{d}}_{i,t})\}$, where $B^{\mathrm{res}}_{i,t} = (B^{\mathrm{liq}}_{i,t} - c_{i,t})^+$ is the liquidity remaining after consumption, so the two purchases together never exceed that remainder.

C.5 Settlement Details

This appendix specifies the deterministic settlement rules that Eq. 3d summarizes, in execution order: the setup that all three markets share, then the labor, goods, and credit clearing in turn, and finally the balance-sheet update. All coefficients are design constants of the released configuration.

Market-clearing setup. Some preparatory calculations of indicators are required before carrying out the three market clearing processes. At the beginning, the bank sets the loan rate r^{ℓ}_t, the price at which households and the firm obtain credit, at the head of the step, before cognition, so that it can enter the information set. The rate is the policy rate r_t plus a lending spread, subject to the bounds $[\underline{r}^{\ell}, \overline{r}^{\ell}]$; the spread combines a constant markup β_0, a credit-risk premium proportional to the nonperforming-loan ratio npl_{t-1} with coefficient β_1, and a macro-risk premium $\beta_2\,(u_{t-1} - u^{\mathrm{nat}})^+$ that activates only when the lagged simulated unemployment rate exceeds the threshold u^{nat}:

$$r^{\ell}_t = \mathrm{clip}\big(r_t + \beta_0 + \beta_1\, \mathrm{npl}_{t-1} + \beta_2\,(u_{t-1} - u^{\mathrm{nat}})^+;\ \underline{r}^{\ell},\ \overline{r}^{\ell}\big), \tag{23}$$

where $(x)^+ = \max(x, 0)$. Then the firm adjusts desired investment I_t^{d} toward a target I_t^* at speed κ_I. The target is baseline investment I_t^{base} scaled by a bounded response ratio v_t, which is constructed from two opposing signals: a demand-side pull that rises with the lagged simulated output gap γ_{t-1} (coefficient κ_γ) and the fiscal stance $\Delta_{g,t}^{12}$, the annual growth rate of government expenditure (coefficient κ_G),

$$v_t^{\mathrm{pull}} = \kappa_\gamma \, \gamma_{t-1} + \kappa_G \, \Delta_{g,t}^{12}, \tag{24}$$

and a credit-cost drag that activates only when the loan rate exceeds the reference rate r^{ref} (coefficient κ_ℓ),

$$v_t^{\mathrm{drag}} = \kappa_\ell \left(r_t^\ell - r^{\mathrm{ref}}\right)^+. \tag{25}$$

The response ratio nets the pull against the drag and bounds the result to the admissible interval $[v^{\min}, v^{\max}]$:

$$v_t = \mathrm{clip}\big(1 + v_t^{\mathrm{pull}} - v_t^{\mathrm{drag}};\ v^{\min},\ v^{\max}\big), \qquad I_t^* = I_t^{\mathrm{base}} \cdot v_t. \tag{26}$$

Desired investment then moves toward the target at adjustment speed κ_I:

$$I_t^{\mathrm{d}} = I_{t-1} + \kappa_I \big(I_t^* - I_{t-1}\big). \tag{27}$$

Total demand then aggregates the just-projected household actions with firm and government spending,

$$\tilde{D}_t = \sum_{i=1}^{N} \omega_i \, c_{i,t} + I_t^{\mathrm{d}} + g_t, \tag{28}$$

the population-weighted consumption demand plus the desired investment and government expenditure. Finally, the labor-market quota target is anchored to the employment rate e_0 at the simulation origin (which carries no test-window information) and responsive to labor demand. The construction proceeds in two steps. First, the demand-pressure ratio compares total demand to trend output $\tilde{Y}_t$ (an n^{tr}-month moving average of the simulator's own output), bounded to $[\delta^{\min}, \delta^{\max}]$:

$$\delta_t = \mathrm{clip}\big(\tilde{D}_t / \tilde{Y}_t;\ \delta^{\min},\ \delta^{\max}\big), \tag{29}$$

so $\delta_t = 1$ means demand equals the economy's normal pace, and values above (below) one signal excess (deficient) demand. Second, the quota target adjusts the anchored employment rate by this excess, with sensitivity κ_u, and bounds the result to the admissible interval $[e^{\min}, e^{\max}]$:

$$e_t^* = \mathrm{clip}\big(e_0 \cdot [1 + \kappa_u \, (\delta_t - 1)];\ e^{\min},\ e^{\max}\big). \tag{30}$$

In brief, the bank quotes a lending price first thing each step, before anyone decides anything, so borrowers can see it when planning. The firm then decides how much to build, based on how busy it was, how expensive credit is, and whether the government is spending more. Adding up household spending plans, firm investment, and government expenditure gives total demand, and hiring follows that demand: when demand runs ahead of the economy's normal pace, the employment target drifts up, but only gradually and within fixed limits, so one hot month cannot trigger a hiring spree.

Labor-market clearing. The quota target of Eq. 30 is anchored to the employment rate at the simulation origin and responsive to labor demand, so persistent deviations of demand from trend move hiring. Household-group employment probabilities evolve through an experience rule: a bounded probability step per settled step. This probability-evolution mechanism follows the search-and-matching tradition in which employment status responds gradually to realized outcomes (Mortensen & Pissarides, 1994). The quota is then filled by ranking household groups on employment probability and intended hours. Each group is scored by a weighted combination of its state and its action,

$$\mathrm{score}_{i,t} = \kappa_{\mathrm{emp}} \, p_{i,t-1}^{\mathrm{emp}} + \kappa_{\mathrm{work}} \, x_{i,t}^{\mathrm{work}}, \tag{31}$$

where $p_{i,t-1}^{\mathrm{emp}}$ is the group's employment probability and the weights satisfy $\kappa_{\mathrm{emp}} + \kappa_{\mathrm{work}} = 1$. Groups are ranked by the score with ties broken by the group identifier, and the quota e_t^* of Eq. 30 is filled in rank order: a ranked group is employed while its population weight fits the remaining quota, with a boundary tolerance of ν score points. Total labor is the weight-aggregated intended hours of the employed groups.

This mechanic shows that the job market has a speed limit: employment never jumps to whatever today's demand would justify; it is anchored to how many people had jobs when the simulation starts and moves only slowly toward demand, and each group's odds of working inch up when its members are employed and inch down when they are not, like a reputation that builds with experience and fades without it.

Goods-market clearing. Fiscal expansion and monetary easing both raise the investment target of Eq. 26, and its rate term inherits the policy vector through the credit market. Prices and wages carry an inflation anchor that bounds drift: the goods-price update is clipped to $\pm\zeta$ per month, with ζ the monthly equivalent of the peak annual inflation observed across the economies in the calibration window; wage and consumption anchors are indexed only to nonnegative inflation, capturing downward nominal wage rigidity (Tobin, 1995; Akerlof et al., 1996) and the consumption ratchet (Turvey, 1950); and every such anchor, namely the price clip, the wage and consumption indexation, and the household memory update that absorbs realized inflation, reads the simulator's own inflation path, never the observed test-window CPI. Household inflation expectations revert toward the annual target π^* (Bernanke & Woodford, 1997) at rate ω per month; and the demand gap entering the firm's response is measured against a trailing n^{util}-month utilization benchmark. Clearing sets output at the smaller of demand and capacity: capacity is $A_t L_t$ with A_t the firm's labor productivity, demand is $\tilde{D}_t$ as defined at Eq. 28, and $Y_t = \min(\tilde{D}_t,\, A_t L_t)$. When capacity binds, investment and government expenditure are treated as committed demand and are served first; household consumption absorbs the remainder,

$$C_t = \big(Y_t - I_t^{\mathrm{d}} - g_t\big)^+, \qquad \mathrm{ration}_t = \frac{C_t}{\sum_{i=1}^{N} \omega_i\, c_{i,t}}, \tag{32}$$

and each group's realized consumption is its demanded amount $c_{i,t}$ scaled by $\min(1, \mathrm{ration}_t)$. The cleared output Y_t then decomposes into realized consumption, investment, and government expenditure, generating the income flows of every sector: the wage bill, firm profit, tax revenue, and transfer payments.

To put in a nutshell, prices and wages can move, but never freely. These brakes stop the simulated economy from developing runaway inflation or a wage–price spiral.

Credit-market clearing. The loan rate of Eq. 23 is set before cognition, as noted above, so the clearing starts from the applications: the application of group i is determined in the projection layer (Appendix C.4) and denoted $\hat{b}_{i,t}$, and the firm applies for $\hat{b}_t^{\mathrm{firm}} = \min\big(\kappa^{\mathrm{cap}}\, I_t^{\mathrm{d}},\ I_t^{\mathrm{d}} - \kappa^{\mathrm{nw}}\, V_{t-1}^{\mathrm{firm}}\big)^+$ with V_{t-1}^{firm} the firm's net worth. Applications pass a logit approval rule,

$$\begin{aligned}
\mathrm{logit}_{i,t} = \ & \beta_0^{\mathrm{c}} && \text{(intercept)} \\
& - \beta_1^{\mathrm{c}}\, \mathrm{dti}_{i,t} && \text{(debt burden)} \\
& + \beta_2^{\mathrm{c}}\, \min\big(\beta_3^{\mathrm{c}}, \mathrm{liq}_{i,t}\big) && \text{(liquidity buffer)} \\
& + \beta_4^{\mathrm{c}}\, p_{i,t-1}^{\mathrm{emp}} && \text{(employment stability)} \\
& - \beta_5^{\mathrm{c}}\, \big(u_t - u^{\mathrm{nat}}\big)^+, && \text{(macro risk)}
\end{aligned} \tag{33}$$

where $\mathrm{dti}_{i,t}$ is the debt-to-income ratio, $\mathrm{liq}_{i,t}$ the deposits-to-income ratio, and u_t the simulated unemployment rate cleared in the same step, the unemployment term activating above the same natural-rate threshold u^{nat} as Eq. 23; an application is approved when the logistic transform of Eq. 33 is at least β_6^{c}. Approved applications are then rationed by bank funding. The bank's loanable funds combine deposits and equity net of a nonperforming-loan deduction:

$$F_t = \kappa^{\mathrm{dep}}\, \mathrm{dep}_t + \kappa^{\mathrm{eq}}\, \mathrm{eq}_t - \kappa^{\mathrm{npl}}\, \mathrm{npl}_{t-1}, \tag{34}$$

and the rationing factor divides these funds by the total approved borrowing demand:

$$\varphi_t = \min\Big\{1,\ \frac{F_t}{\sum_{i \in \mathcal{J}_t} \omega_i\, \hat{b}_{i,t} + \hat{b}_t^{\mathrm{firm}}}\Big\}, \tag{35}$$

where $\mathcal{J}_t$ collects the approved groups; household grants are the approved applications scaled by φ_t, as is the firm grant $\varphi_t\, \hat{b}_t^{\mathrm{firm}}$. The cleared grants then change household and firm debt against bank assets, one-for-one, through the accounting circuit of the settlement kernel.

So a loan application passes two gates. The first is a credit score: heavy debt hurts, cash and a job help, and a weak economy hurts. The second is the bank's own wallet: if approved applications add up to more than the bank can fund, every approved borrower gets the same fraction of what they asked for.

Balance-sheet update. The settlement of step t leaves the cleared transaction records $\boldsymbol{\xi}_t$, each stamped with its payer, payee, and amount. All updates in this step are updates of the balance-sheet object $\mathcal{B}_t$ of Eq. 17: each record is entered once as a debit on its payer and once as an equal credit on its payee, following the quadruple-entry principle on which stock-flow-consistent accounting rests (Copeland, 1949; Godley & Lavoie, 2007; Caiani et al., 2018); for households the table covers group states (employment probability, memory, anchors) alongside balances, for institutions it is the balance sheet. Households receive labor, non-labor, and transfer income net of taxes, consume, service debt, and hold the residual in deposits; each group's employment outcome moves its employment probability by $\pm\Delta e$ and appends to its rolling memory. The firm earns profit $\pi_t^{\mathrm{firm}} = Y_t - W_t - r_t^{\mathrm{firm}} - \delta K_{t-1}$, the output remaining after the wage bill W_t, loan interest r_t^{firm}, and depreciation at rate δ per month, and accumulates capital $K_t = (1-\delta)K_{t-1} + I_t$. The bank's nonperforming loans decay at a factor of δ^{npl} per month plus new defaults, it earns the loan-deposit spread, and equity is the balancing item. The government accumulates debt equal to its deficit (expenditure plus transfers minus tax revenue), and the central bank accumulates asset purchases. Every transaction record is checked against its counterpart sector before the state is committed (Section 4).

D Experimental Details

D.1 Policy Agent Details

This appendix details the three instantiations of the policy decision rule D of Eq. 10.

Rule policymaker. A Taylor-style reaction function (Taylor, 1993) over the same inputs the LLM reads: the policy-rate increment responds to expected inflation (the questionnaire mean) minus the 2% target and to the output gap, $\Delta r = \mathrm{clip}\big(0.35\,(\pi^e - 2\%) + 0.25\,\gamma;\ \pm 1\%\big)$ in percentage points; the fiscal increment is countercyclical, $\Delta g = \mathrm{clip}\big(0.55\,(u - 5.5\%)^+ + 0.30\,(\mathrm{stress} - 0.35)^+ - 0.20\,(\mathrm{debt}/Y - 1)^+;\ \pm 8\%\big)$, raising expenditure when unemployment or household financial stress runs high and trimming it when the debt ratio exceeds 100% of output, with transfers set to a fixed share of the fiscal move.

LLM policymaker. A large language model reads the weighted questionnaire aggregate, the macro conditions, and the government and central-bank action memories, and returns a JSON object of bounded instrument increments (the prompt template is given below); the deterministic kernel then projects and executes the result, exactly as for households. The system prompt fixes the advising role and the return schema, one JSON object of bounded increments over the two tax rates, government expenditure, discretionary transfers, the policy rate, and asset purchases, plus a communication string. The user prompt renders the same information set the households see, now at the economy level:

Policymaker user prompt template

```
POLICY ECONOMIC SITUATION

No calendar date or country identity is disclosed; use only the
economic information below.

Macroeconomic conditions:
- Unemployment is {unemployment}.
- Year-over-year inflation is {inflation}; {inflation_phrase}.
- The output gap is {output_gap}; {output_gap_phrase}.
- Government debt is {debt_ratio} of annual GDP.
- Monthly real output is {real_output}; monthly nominal output is
  {nominal_output}.
- Household consumption is {household_consumption}; investment is
```

```
  {investment}; government consumption is {government_consumption}.

Financial conditions:
- The policy rate is {policy_rate} and the bank loan rate is
  {bank_loan_rate}.
- Bank credit is {bank_credit}; central-bank assets are
  {central_bank_assets}.
- Household debt is {household_debt}; household deposits are
  {household_deposits}.

Recent government decisions, most recent last:
- ({record_source}): income-tax change {income_tax_delta} percentage
  points; consumption-tax change {consumption_tax_delta} percentage
  points; expenditure changed by {government_expenditure_delta} of
  baseline; transfer changed by {discretionary_transfer_delta} of
  baseline.

Recent central-bank decisions, most recent last:
- ({record_source}): policy-rate change {policy_rate_delta}
  percentage points; asset purchases changed by
  {asset_purchase_delta} of baseline.

Weighted household questionnaire:
- The survey covers {n_groups} household groups with total weight
  {weight_sum}.
- {questionnaire_field}: mean {mean}, p10-to-p90 range {p10} to
  {p90}.

Decision request:
Use the macro situation, financial conditions, and household
questionnaire jointly. Propose one coherent fiscal and monetary
package for the next simulation step. Stabilize output and employment
while respecting inflation, public-debt, and financial-stability
constraints. Return exactly one valid JSON object with the required
schema and no text outside the JSON object.
```

RL policymaker. A tabular Q-learning agent. The state is the discretized macro situation, inflation binned at $\{1.5\%, 2.5\%, 4\%\}$, unemployment at $\{4.5\%, 6\%\}$, and the output gap at $\{-2\%, 0, 2\%\}$, giving 48 states, and the action is a joint move of the policy rate (five Δr steps) and government expenditure (five Δg steps), giving 25 actions. The reward is the negative dual-mandate loss (inflation and unemployment distance from target); updates use $\epsilon^{\mathrm{rl}} = 0.1$ exploration, learning rate $\alpha^{\mathrm{rl}} = 0.2$, and discount $\gamma^{\mathrm{rl}} = 0.95$. The Q-table is trained on practice rollouts in the simulation environment and then frozen: at evaluation the agent acts greedily by table lookup, with no further updates.

D.2 Data Sources

Macro data. Each economy is described by a monthly panel of official series aligned to a common 2016-01 calendar. The United States draws on FRED/ALFRED mirrors of the official statistical releases (Federal Reserve Bank of St. Louis, 2026): the federal-funds policy rate, the unemployment rate and employment level, headline and core CPI, personal consumption and personal income, nominal and real GDP, gross private domestic investment, and government consumption expenditure, federal receipts and expenditures plus the debt-to-GDP ratio, banking series, Federal Reserve total assets, capacity utilization, industrial production, and corporate profits. Canada draws on Statistics Canada tables (Statistics Canada, 2026): expenditure-side nominal and real GDP with household consumption, government consumption, and business gross fixed-capital formation, monthly real GDP by industry, corporate wages and operating surplus, Labor Force Survey employment and unemployment , CPI, bank deposits and credit series, and government revenue, expenditure, and social benefits, with the policy rate from the Bank of Canada Valet API (Bank of Canada, 2026). Table 7 lists the series identifiers, their economic content, and their native frequencies.

Table 7: Macro data inventory: series identifiers, economic content, and native frequency. Alignment to the monthly calendar follows the rules below the table.

Indicator	Series ID	Economic meaning	Frequency
United States (FRED/ALFRED mirrors)			
Bank credit	TOTBKCR	total loans and leases of commercial banks	Weekly
Bank prime rate	MPRIME	prime loan rate charged by banks	Monthly
Consumer loans	CONSUMER	consumer credit held by banks	Weekly
Real-estate loans	B1026NCBA	real-estate loans of commercial banks (H.8)	Weekly
Central-bank assets	WALCL	Federal Reserve total assets	Weekly
Policy rate	FEDFUNDS	effective federal-funds rate	Monthly
Capacity utilization	TCU	industrial capacity utilization	Monthly
Corporate profits	CP	corporate profits with inventory valuation	Quarterly
Industrial production	INDPRO	industrial production index	Monthly
Investment	GPDI	gross private domestic investment	Quarterly
Federal expenditures	FGEXPND	federal government expenditures	Quarterly
Federal receipts	FGRECPT	federal government receipts	Quarterly
Government consumption	GCE	government consumption expenditures	Quarterly
Government debt	GFDEGDQ188S	federal debt as a share of GDP	Quarterly
Personal consumption	PCE	personal consumption expenditures	Monthly
Personal income	PI	personal income	Monthly
Average hourly earnings	CES0500000003	average hourly earnings, production workers	Monthly
Employment level	CE16OV	employment level, 16 years and over	Monthly
Hires	JTSHIL	total nonfarm hires (JOLTS)	Monthly
Job openings	JTSJOL	total nonfarm job openings (JOLTS)	Monthly
Nonfarm payrolls	PAYEMS	total nonfarm employment	Monthly
Participation rate	CIVPART	labor-force participation rate	Monthly
Quits	JTSQUL	total nonfarm quits (JOLTS)	Monthly
Unemployment rate	UNRATE	civilian unemployment rate	Monthly
Headline CPI	CPIAUCSL	consumer price index, all items	Monthly
Core CPI	CPILFESL	consumer price index, excluding food and energy	Monthly
Nominal GDP	GDP	nominal gross domestic product	Quarterly
Real GDP	GDPC1	real gross domestic product	Quarterly
Canada (Statistics Canada tables; policy rate from Bank of Canada Valet)			
Nonpersonal deposits	v41552792	nonpersonal term and demand deposits	Monthly
Personal deposits	v41552791	personal term and demand deposits	Monthly
Monetary base	v37145	monetary base of the Bank of Canada	Monthly
Policy rate	v39079	target for the overnight rate	Daily
Business investment	v62305764	business gross fixed-capital formation	Quarterly
Corporate surplus	v62295553	corporate net operating surplus	Quarterly
Operating surplus	v62295552	gross operating surplus	Quarterly
Real GDP by industry	v65201210	monthly real GDP by industry	Monthly
Government consumption	v62305762	government final consumption expenditure	Quarterly
Government debt securities	v106982439	government debt securities outstanding	Annual
Government expense	v106982382	total government expenses	Annual
Government revenue	v106982326	total government revenue	Annual
Social benefits	v106982396	government social benefits	Annual
Government taxes	v106982327	taxes less subsidies	Annual
Household consumption	v62305755	household final consumption expenditure	Quarterly
Wages and salaries	v62295550	household wages, salaries, and supplements	Quarterly
Employment level	v2062811	employment level (Labor Force Survey)	Monthly
Participation rate	v2062816	labor-force participation rate	Monthly
Unemployment rate	v2062815	unemployment rate (Labor Force Survey)	Monthly

Table 7 (continued)

Indicator	Series ID	Economic meaning	Frequency
Headline CPI	v41690973	consumer price index, all items	Monthly
Core CPI	v41691233	consumer price index, excluding indirect taxes	Monthly
Nominal GDP	v62305783	nominal gross domestic product, expenditure side	Quarterly
Real GDP	v62305752	real gross domestic product, expenditure side	Quarterly

Frequency alignment. All series are aligned to the common monthly calendar under three fixed rules. Monthly series enter directly. Weekly and daily series enter at their last observation of the month. Quarterly series, and the annual Canadian fiscal accounts, repeat their published value over the three (or twelve) calendar months of the period; rates, indexes, and price levels are kept exactly as published and are never summed or averaged.

We repeat quarterly values rather than interpolating them, for three reasons. First, interpolation would invent intra-quarter dynamics that the source data never measured. Second, a smooth interpolant needs the next quarter's observation, so month t would depend on information dated after it. Third, evaluation operates on monthly log growth: repetition keeps growth at zero within the quarter and concentrates the full quarterly change at the boundary, whereas interpolation would smooth the growth distribution and inflate measured persistence. Repetition also matches practice, since monthly policy analysis in real time works with the last published quarter until the next one arrives.

Household survey microdata. Household agents are compressed from national household-wealth surveys: the US Survey of Consumer Finances, waves 2016, 2019, and 2022 (Board of Governors of the Federal Reserve System, 2023), and the Canadian Survey of Financial Security, waves 2016, 2019, and 2023 (Statistics Canada, 2024). Survey releases lag by about one year, so each simulation window uses the most recently available survey wave. Each record carries a population survey weight, income flows, balance-sheet stocks, demographics, education, employment, and housing status.

E Additional Results

E.1 Full Backend Comparison

Table 8 reports the complete five-backend comparison underlying the main-text Table 1, which retains the pure-ABM block and the two flash backends discussed in the main text.

Discussion. Table 8 presents the complete comparative simulation performance of the proposed model and the baseline models. (1) The pure ABM models still lead the LLM-bearing frameworks on distribution fidelity and robustness, because existing LLM systems lack alignment with real conditions and control over their outputs. (2) GAME attains the best performance on both similarity and robustness: it reproduces normal economic trends while combining the stable simulation that the fixed parameters of ABM provide with the dynamic, realistic decisions of LLM agents. (3) Across five LLM backends of very different parameter scales the metrics stay close, so the framework does not depend on a particular foundation, and the parameter size of the underlying model does not determine the results.

E.2 Transmission Experiments

Table 9 reports the transmission experiments: the average error between the model's relevant sequence after perturbation and that of the baseline unperturbed model. All models simulate for three years starting from 2024. For each series the highest and lowest of the five seeded runs are dropped, and the mean $\pm$ standard deviation of the remaining three is reported. Expansionary shocks (negative rate, positive expenditure) should raise output; contractionary shocks should lower it.

Table 8: Full comparison across LLM cognition backends. Conventions as in Table 1, with rankings and shading computed over all rows of this table.

Model	Pearson (%) ↑	JSD (%) ↓	RMSE ↓	Terminal CV (%) ↓	Seed RMSE SD ↓
Pure ABM					
AB-SFC	−0.2±15.1	28.4±4.7	3.55±1.15	1.75±0.32	1.16±0.17
firm–bank ABM	+1.6±9.7	20.2±3.1	5.07±0.79	2.09±1.26	1.18±0.57
Minsky	+23.0±9.2	31.5±12.7	2.69±1.53	4.38±0.62	1.62±0.19
glm-5.3-flash (Zeng et al., 2026)					
GAME (ours)	+37.2±6.2	17.2±4.3	3.07±0.49	1.67±0.70	0.54±0.36
EconAgent	−13.6±7.4	24.4±9.1	3.84±1.09	2.65±0.89	1.30±0.36
GPLab	+23.3±5.5	34.4±11.2	4.63±0.82	1.45±0.31	0.88±0.15
qwen3.8-flash (Qiu et al., 2026)					
GAME (ours)	**+44.7±6.9**	**12.8±4.0**	2.79±0.36	1.98±0.85	0.69±0.26
EconAgent	−12.6±8.3	35.8±11.0	5.38±1.16	3.98±1.64	1.44±0.09
GPLab	+25.5±6.1	33.3±10.5	5.38±0.96	1.61±0.62	1.09±0.43
deepseek-v4-flash (Xu et al., 2026)					
GAME (ours)	+42.4±7.1	17.6±5.3	4.07±0.61	2.05±0.90	0.81±0.41
EconAgent	−3.9±10.8	35.0±15.6	6.88±1.85	3.73±0.29	1.98±0.13
GPLab	+23.5±6.7	34.6±10.1	5.27±0.69	1.32±0.12	0.86±0.09
qwen-3.8-27b (Qiu et al., 2026)					
GAME (ours)	+43.1±9.3	14.6±3.6	**2.61±0.26**	**0.81±0.22**	**0.43±0.12**
EconAgent	−16.4±7.8	23.5±8.2	3.81±1.45	3.56±0.93	1.49±0.66
GPLab	+18.5±7.5	37.8±14.0	3.44±1.42	3.74±0.32	1.56±0.32
Ornith-1.5-35B-A3B (Ornith Team, 2026)					
GAME (ours)	+35.4±7.6	13.5±3.7	2.66±0.46	2.17±0.46	0.87±0.22
EconAgent	−6.6±5.3	22.4±6.7	3.64±1.33	3.76±0.39	1.52±0.25
GPLab	+20.4±6.6	30.6±13.6	4.88±0.95	1.65±0.31	1.02±0.19

Table 9: The average error between the model's relevant sequence after perturbation and that of the baseline unperturbed model. All models simulate for three years starting from 2024. Expansionary shocks (negative rate, positive expenditure) should raise output; contractionary shocks should lower it.

Constant shock	ΔGDP	ΔConsumption	ΔInvestment
Panel A. Interest-rate shocks			
−300 bp	+7.0 ± 1.4	+7.7 ± 1.3	+13.3 ± 7.7
−100 bp	+1.2 ± 3.6	+1.3 ± 2.0	+3.2 ± 2.0
+100 bp	−1.3 ± 2.6	−0.8 ± 2.9	−4.4 ± 2.6
+300 bp	−4.5 ± 2.9	−5.7 ± 3.1	−8.8 ± 5.6
Panel B. Expenditure shocks			
−20%	−4.4 ± 2.9	−4.1 ± 2.6	−3.0 ± 5.5
−10%	−3.2 ± 2.6	−2.4 ± 2.9	−2.7 ± 3.5
+10%	+2.7 ± 1.8	+2.7 ± 2.2	+4.5 ± 5.3
+20%	+3.8 ± 0.2	+4.3 ± 0.4	+7.4 ± 4.4
Panel C. Combined shocks			
Stimulus (rate −100 bp, exp. +10%)	+2.9 ± 1.9	+3.1 ± 2.3	+4.9 ± 3.9
Tightening (rate +100 bp, exp. −10%)	−2.2 ± 0.2	−2.3 ± 0.3	−4.5 ± 1.1

The discussion follows. (1) Transmission is correctly signed and dose-responsive. Every shocked variant responds in the textbook direction, and within each panel the response grows with the shock size, so the simulated economy reacts to policies never observed in history as macroeconomic theory predicts rather than drifting or breaking. The composition of the response is equally classical: consumption moves roughly one-for-one with output under every variant, while investment is the most elastic margin and amplifies monetary shocks most strongly, so counterfactual policies propagate through the interest-sensitive investment margin rather than through mechanical rescaling of the aggregate path. (2) The two levers partially substitute rather than stack. Both combined variants fall short of the sum of their single-lever effects across all three series, so combined stimulus and combined tightening deliver less than their parts would separately. The response is mildly asymmetric: monetary easing outweighs equal-size tightening, in the direction the savers' interest-income channel predicts, and fiscal cuts bite slightly harder than equal expansions.

E.3 Macro Phenomena Discussion

This appendix discusses the RQ3 results of Figures 3, 4, and 5. (1) The simulator reproduces the two economic phenomena through distinct transmission channels: Figure 3 shows that, under the simulated economic crisis scenario, output and investment decline sharply compared to normal conditions, while unemployment rises substantially. The indicators improve after government intervention, but the economy remains worse than under normal conditions. Once the crisis window ends, all indicators recover quickly and follow a trend close to the baseline scenario. Figure 4 shows largely the same pattern; in the initial phase of the pandemic shock, seasonal effects briefly improve the economy, but it deteriorates again under sustained pressure, a trajectory consistent with real-world observations. The shock sensitivities also line up with actual dynamics: employment responds first and output follows with a lag. The model therefore reproduces both the overall trends and the transmission dynamics of these scenarios.

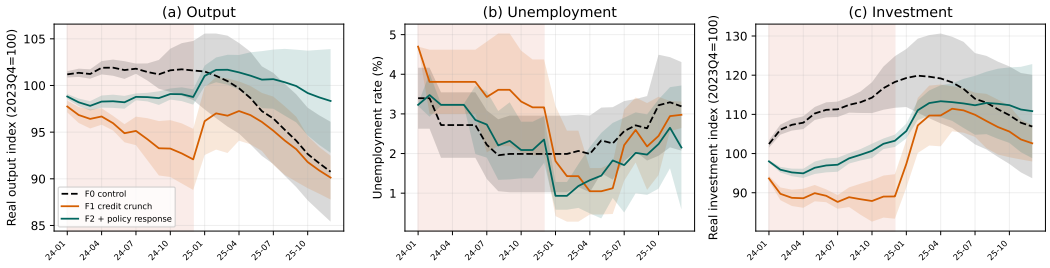


Figure 3: Simulated financial-crisis scenarios on US, 2024-01–2025-12: control without intervention(F0), simulated credit crunch with +300bp on loan spread and -60% in credit supply(F1), and F1 plus policy response with -450bp on policy interest rates and +10% government expenditure(F2); curves stand for mean results; colored bands are ± standard deviation across runs, and the pink background marks the crunch application interval. All results are shown under 5 runs.

(2) The simulator roughly reproduces the canonical economic regularities at three distinct grades. (a) The Phillips panel: The Phillips curve implies an inverse relationship between inflation and unemployment. The baseline points sit loosely, whereas GAME's lie closer to the actual slope, and away from the outliers its locations follow the classical empirical pattern. (b) The Taylor panel: Under the Taylor rule the policy rate should track the Taylor-rule indicator, the level shown in Figure 5 (2). GAME sits closer to the actual series than the baseline; it underestimates the policy rate, but stays nearer the diagonal, in line with the theoretical conclusion. (c) The GDP panel: The pure ABM baselines move stably, while the other LLM-bearing systems fluctuate strongly and drift away from the actual path. GAME stays closest to the real path over the whole window and still retains its underlying volatility, which indicates that it combines the stability of pure ABM models with the responsiveness of LLM agents.

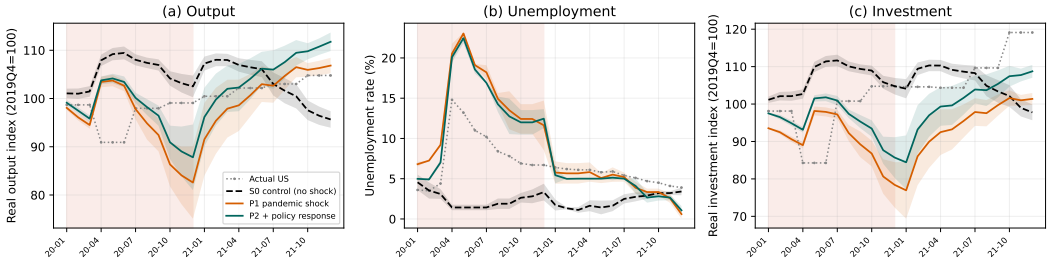


Figure 4: Simulated pandemic scenarios on US, 2020-01–2021-12: no-shock control (S0), income and employment-quota shock peaking at -20% in early 2020 (P1), and P1 plus policy response with -300 bp on policy interest rates and $+20\%$ government expenditure (P2), against the actual US series; curves stand for mean results; colored bands are $\pm$ standard deviation across runs, and the pink background marks the 2020 shock application interval. All results are shown under 5 runs.

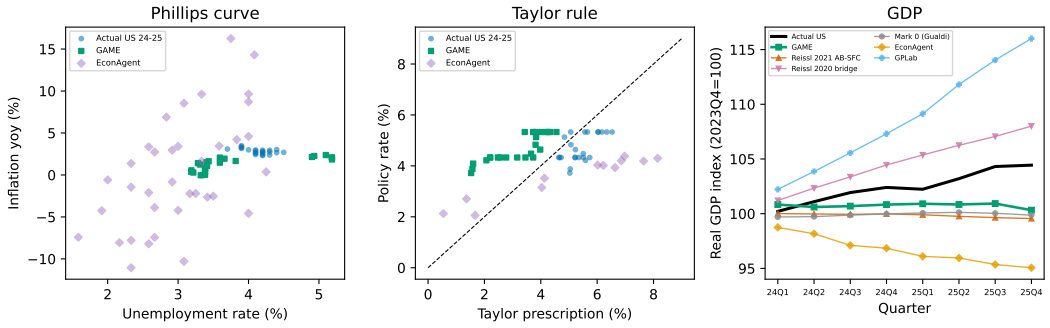


Figure 5: Macro-phenomena gallery, simulated vs. actual. (1) and (2) Scatter plots of the corresponding empirical indicators for each simulation system and the real economy. (3) GDP trends excluding inflation rates under each simulated system and in the real economy.