Worldlines in the Mean Field: LLM Social Simulation of a Real Town

Anonymous ACL submission

Abstract

Long-horizon LLM-based resident simulation requires belief updates that reflect agents' daily experience while remaining coherent with social context. Existing LLM-based social simulators primarily update agents from individual experience and memory, while recent mean-field LLM methods mainly target single-topic online settings.

We introduce mean-field (MF) conditioning for multi-topic LLM resident simulation, where agents revise beliefs using aggregate topic-level stance distributions and natural-language belief nuances. We instantiate the framework for Koshimizu, Hokkaido, compressing a town of approximately 4,300 residents into 100 synthetic agents with demographics, districts, social ties, routines, and seven policy-topic beliefs. Four years of dynamics are simulated over 28 days, with each week representing one year.

Across exogenous-event worldlines including earthquake/blackout, COVID, and school-closure events, the simulator produces event-conditioned population-level interest shifts that directionally align with municipal budget changes. MF conditioning reduces broad `positive`-label collapse while preserving event-relevant collective shifts, and improves propagation of blackout-related discourse beyond direct broadcast exposure. These results suggest that MF is a practical mechanism for adding structured social context to long-horizon, multi-topic LLM resident simulations.

1 Introduction

LLM-based social simulation is scaling from believable agents to city- and society-scale simulators (Park et al., 2023; Piao et al., 2025; Bougie and Watanabe, 2025; Ye et al., 2026). At the same time, surveys and validation studies warn that simulated populations should be treated as externally anchored and mechanistically diagnosable models, not as direct human outcomes (Mou et al., 2026; Anthis et al., 2025; Seshadri et al., 2026; Lin et al., 2026). This paper studies that issue in long-horizon resident belief revision, where agents must update beliefs from both daily experience and socially observed context.

We introduce mean-field (MF) conditioning for multi-topic LLM resident simulation. Each agent revises topic-specific beliefs from persona, daily experience, previous beliefs, and an optional MF social-context signal containing a stance distribution and representative belief nuances. We compare Local-MF, Town-MF, and Hybrid-MF, corresponding to neighborhood diffusion, town-wide public context, and an interest-weighted mixture.

We build a Koshimizu-grounded resident simulation testbed, abstracting a real town in Hokkaido, Japan, with approximately 4,300 residents into 100 agents. Four years are compressed into 28 simulated days, and event worldlines inject earthquake, town-center blackout, COVID-19, and school-closure events. Municipal budgets are used only as coarse administrative proxies, not as ground-truth labels of individual beliefs. Our contribution is a compact mechanism-and-evaluation package: observer-dependent MF conditioning, a real-municipality event-worldline testbed, and diagnostics of stance-label drift, discourse coverage, and graph-faithfulness.

2 Related Work and Positioning

LLM social simulation and validity. Generative Agents (Park et al., 2023) introduced memory- and reflection-based LLM agents, and AgentSociety, CitySim, and MobileCity scale this paradigm to large social or urban settings (Piao et al., 2025; Bougie and Watanabe, 2025; Ye et al., 2026). Related work on user simulation and grounded opinion dynamics motivates diagnosable belief states (Kolluri et al., 2025; Mehri et al., 2026; Li et al.,

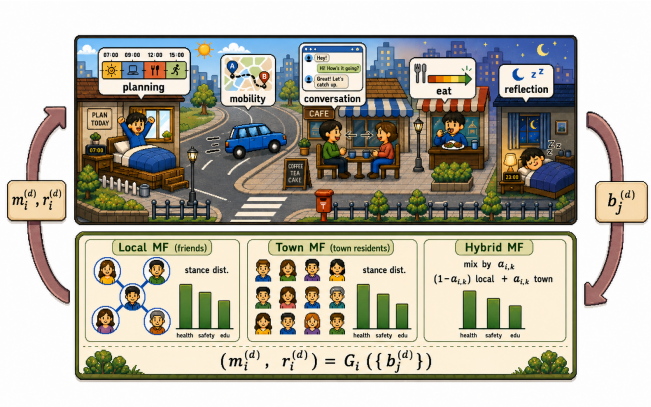

Figure 1: Overview of mean-field conditioning in the multi-topic LLM resident simulation. Agents revise beliefs from daily experience and MF social context, and the updated beliefs are aggregated into the next social signal.

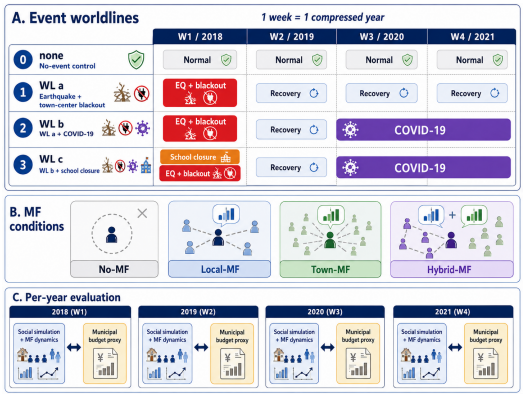

Figure 2: Experimental design. A 4-week simulation of 2018–2021 injected earthquakes, COVID-19, and school closures to compare four MF conditions (None/Local/Town/Hybrid). Evaluations used budget anchoring and internal diagnostics.

2025; Hashimoto et al., 2026). Validation studies warn that simulated users and digital twins can show distributional gaps, intervention-induced drift, systematic distortions, and strategic behavior unlike humans (Seshadri et al., 2026; Lin et al., 2026; Peng et al., 2026; Wang et al., 2026). We therefore treat outputs as mechanism diagnostics, not as direct measurements of human responses.

Mean-field and real-municipality anchoring. MF-LLM (Mi et al., 2025) and MF-MDP (Zhang et al., 2026) use aggregate population signals to stabilize LLM-agent opinion dynamics, mainly in online or scalar opinion settings. City- and society-scale simulators ground LLM agents in urban environments, demographic profiles, or mobility networks (Piao et al., 2025; Bougie and Watanabe, 2025; Ye et al., 2026), and social digital twins aim to calibrate simulated population responses to observable real-world signals (Koaik et al., 2026). Our setting differs by using a bounded municipality with public administrative records to diagnose multi-topic belief dynamics under event worldlines.

3 Method

We formulate belief revision as a socially conditioned update process. Each agent revises topic-specific beliefs from personal experience and an optional MF social-context signal. Local-MF represents neighbourhood-level diffusion, Town-MF represents broader public context, and Hybrid-MF interpolates between the two using topic interest. This design follows classical social-influence models, where private beliefs are updated under interpersonal or aggregate social influence (DeGroot, 1974; Latané, 1981; Friedkin and Johnsen, 2011).

Belief state. Each agent i maintains one belief per topic k and day d:

$$b_{i,k}^{(d)} = \left(I_{i,k}^{(d)}, s_{i,k}^{(d)}, u_{i,k}^{(d)}\right), \tag{1}$$

where $I \in [0,1]$ is interest, $s \in S$ is a stance label, and u is a natural-language nuance. We use $S = \{\text{positive}, \text{neutral}, \text{negative}\}$ in the main text.

Mean-field social context. For each topic, an MF signal consists of a stance distribution m and a textual summary r of representative belief nuances. Agent j is an eligible contributor for topic k on day d if

$$I_{j,k}^{(d)} \geq \tau, \qquad s_{j,k}^{(d)} \in S, \tag{2}$$

where τ is an interest threshold. Let $V_k^{(d)}$ be the eligible town population for topic k. With $n_k^{(d)}(s) = |\{j \in V_k^{(d)} : s_{j,k}^{(d)} = s\}|$, the town-wide MF is

$$m_k^{\text{town},(d)}[s] = \frac{n_k^{(d)}(s)}{\sum_{s' \in S} n_k^{(d)}(s')}, \quad s \in S. \tag{3}$$

The local MF is observer-specific. Let $F_{i,k}^{(d)}$ be the eligible social neighborhood of observer i for topic k. Local MF weights this neighborhood by intimacy w_{ij}:

$$m_{i,k}^{\text{local},(d)}[s] = \frac{\sum_{j \in F_{i,k}^{(d)}} w_{ij} \mathbf{1}[s_{j,k}^{(d)} = s]}{\sum_{j \in F_{i,k}^{(d)}} w_{ij}}, \quad s \in S. \tag{4}$$

Topics with no eligible observations are omitted from the textual social context rather than treated as evidence for any public opinion.

Hybrid MF and lagged updating. We use previous-day interest as a proxy for topic participation, $\alpha_{i,k}^{(d-1)} = I_{i,k}^{(d-1)}$. Hybrid MF mixes local and town-wide signals as

$$m_{i,k}^{\text{hyb},(d-1)} = (1 - \alpha_{i,k}^{(d-1)}) m_{i,k}^{\text{local},(d-1)} + \alpha_{i,k}^{(d-1)} m_k^{\text{town},(d-1)}. \tag{5}$$

Define $\mathbf{b}_i^{(d)} = \{b_{i,k}^{(d)}\}_{k=1}^{K}$. On the night of day d, agent i revises beliefs from its persona p_i, daily history $h_i^{(d)}$, previous beliefs, and the previous-day MF signal:

$$\mathbf{b}_i^{(d)} = F_{\text{LLM}}\Big(p_i, h_i^{(d)}, \mathbf{b}_i^{(d-1)}, m_i^{(d-1)}, r_i^{(d-1)}\Big). \tag{6}$$

The resulting beliefs are aggregated into the next social signal,

$$\Big(m_i^{(d)}, r_i^{(d)}\Big) = G_i\Big(\{\mathbf{b}_j^{(d)}\}_{j=1}^{N}\Big). \tag{7}$$

This one-day lag prevents circular dependence within a single reflection step.

Textual realization. The numerical component m is computed directly from aggregate categorical stances. The LLM is used only to generate the textual component r, which summarizes representative natural-language nuances from the same contributor pool while remaining consistent with the computed stance distribution. The contributor pool for r matches the corresponding numerical MF signal: local summaries emphasize socially close and topically engaged neighbors, town-wide summaries emphasize topically engaged residents, and hybrid summaries interpolate between the two sources using the same participation weight as Eq. 5.

4 Experiments

We evaluate whether mean-field (MF) conditioning changes collective belief dynamics in a town-scale LLM resident simulation.

Simulation setting. We construct a simulation of Koshimizu with 100 resident agents. Each agent has demographic attributes, household structure, residential district, social ties, daily routines, and initial beliefs over seven policy topics. The simulation runs for 28 days with 30-minute ticks. We treat each seven-day block as one compressed year corresponding to 2018–2021: the week preserves a weekday/weekend routine cycle, a standard unit in time-use and mobility analysis (Zerubavel, 1985; Jara-Díaz and Rosales-Salas, 2015; González et al., 2008). This is an experimental clock for year-level event injection and budget comparison, not a claim about real-time social change.

Worldlines and MF conditions. Figure 2 shows the event-worldline design. The `none` worldline is a no-event control; worldline `a` adds earthquake and urban-blackout broadcasts; worldline `b` adds COVID-related broadcasts; and worldline `c` further adds a school closure event. The blackout broadcast is restricted to town-center residents, enabling evaluation of propagation to outlying districts. For each worldline, we compare No-MF, Local-MF, Town-MF, and Hybrid-MF under matched personas, routines, conversations, event schedules, and nightly reflection settings.

External alignment. One simulation week corresponds to one year. For topic k and week w, we compare mean-interest change from seed with the 2017-baseline log-budget change:

$$\begin{aligned} \Delta I_{k,w} &= \overline{I}_{k,w} - \overline{I}_{k,\text{seed}}, \\ \Delta \log B_{k,y(w)} &= \log B_{k,y(w)} - \log B_{k,2017}. \end{aligned} \tag{8}$$

We compute external alignment only where the event–budget mapping is clear: disaster preparedness for earthquake/blackout, and health/sanitation for COVID.

Stance-label drift. We also measure within-run drift in topic-level stance labels. The internal `positive`/`neutral`/`negative` labels are the categorical stance states defined in Sec. 3; they do not indicate literal approval or disapproval of exogenous events. For each condition, we compare the seed and final-day distributions over these three stance labels and use the resulting drift as a diagnostic of whether nightly reflection collapses persona-conditioned heterogeneity.

Discourse propagation. Finally, we evaluate event discourse in natural-language belief nuances. Because scalar interest scores can miss whether a specific event context has propagated, we measure blackout-keyword penetration among the 50 outlying-district residents outside the broadcast scope.

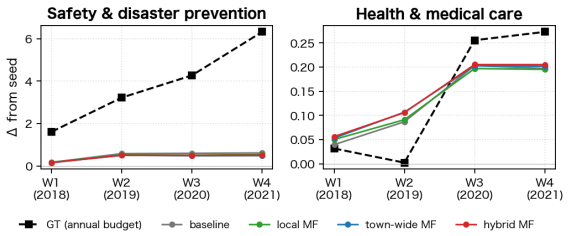

Figure 3: Weekly trajectory for worldline b. Simulated interest changes and municipal budget changes are shown in their raw units, so the comparison is directional rather than scale-matched. The health topic rises in week 3, matching the timing of the COVID-related budget response.

Condition	TVD	Δ positive share
baseline final	0.310	+0.301
local MF final	0.176	+0.124
town-wide MF final	0.194	+0.046
hybrid MF final	0.186	+0.024

Table 1: Stance-label drift across all seven topics in worldline b. MF conditioning reduces the broad `positive`-label collapse seen in the baseline condition.

5 Results

Event-conditioned external alignment The simulator produces topic-specific, event-conditioned shifts rather than uniform drift. Without the COVID broadcast, health interest in worldline a remains nearly flat (+0.03 to +0.08 by week 4). With the COVID broadcast in worldline b, health interest rises to about +0.20 in week 3, corresponding to 2020 and matching the timing of COVID-period health budget increases. We treat this as population-level validation rather than as ground-truth labeling of individual resident beliefs.

MF preserves persona-conditioned heterogeneity We diagnose whether nightly reflection overwrites persona-conditioned stance-label differences. For each topic, we compute the total variation distance (TVD) between seed and final-day distributions over `positive`, `neutral`, and `negative` labels, then average across all seven topics. Because event-relevant topics are included, TVD is used only as a diagnostic for broad positive-label collapse. Baseline has the largest drift: `positive` share increases by +0.301 on average. MF reduces this collapse, especially Town-MF and Hybrid-MF (+0.046 and +0.024), without freezing event-relevant collective shifts.

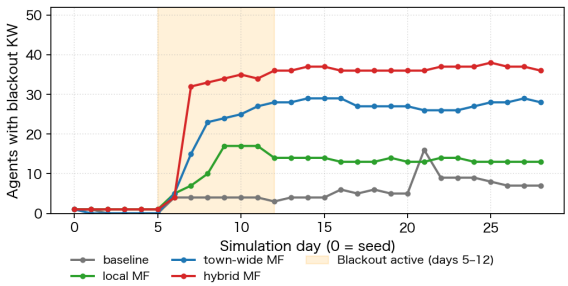

Figure 4: Blackout keyword penetration among the 50 out-of-scope agents. The shaded interval indicates the blackout-active period. Penetration follows No-MF < Local-MF < Town-MF < Hybrid-MF.

Reachability	n	No-MF	Local	Town	Hybrid
1-hop	17	3 (18%)	12 (71%)	11 (65%)	16 (94%)
2-hop	28	1 (4%)	5 (18%)	11 (39%)	16 (57%)
isolated	5	0 (0%)	0 (0%)	3 (60%)	3 (60%)

Table 2: Blackout keyword penetration by acquaintance-graph reachability.

Scope-restricted discourse propagation The strongest MF effect appears in the scope-restricted blackout event. Among the 50 outlying residents outside the broadcast scope, No-MF stays at 4–7 blackout mentions, Local-MF at 13–17, Town-MF at 26–29, and Hybrid-MF at 32–38.

Coverage and graph-faithfulness High coverage does not imply graph-faithful propagation. Local-MF leaves isolated outlying agents at 0% penetration, whereas Town-MF and Hybrid-MF reach 60% of isolated agents (Table 2). Thus, Local-MF models graph-faithful neighbourhood diffusion, while Town-MF and Hybrid-MF model broader public-context coverage.

6 Discussion and Conclusion

MF is best understood as a social-context mechanism rather than a scalar alignment optimizer. Municipal budgets provide only coarse external anchors, not labels of individual beliefs. Internally, MF mitigates broad stance-label drift and changes how blackout-related discourse spreads beyond direct broadcast exposure. Local-MF remains more graph-faithful, whereas Town-MF and Hybrid-MF provide broader public-context coverage. These results do not constitute full counterfactual policy evaluation, but they show that MF makes social context explicit and controllable in long-horizon, multi-topic LLM resident simulation.

Limitations

- Municipal budgets are coarse population-level external references, not ground-truth labels for individual resident beliefs.
- This study uses a 100-agent, four-year-compressed simulation of Koshimizu, so the results should be interpreted as directional validity rather than demographic prediction.
- This study does not perform full counterfactual policy evaluation; it provides a basis for future policy-worldline comparisons.
- Potential risks include over-interpreting simulated residents as real residents, using directional alignment as policy evidence without human review, and amplifying LLM or prompt biases through aggregate social-context signals. We therefore frame the simulator as an exploratory diagnostic tool, not as a substitute for surveys, participatory planning, or expert policy analysis.

References

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

Nicolas Bougie and Narimawa Watanabe. 2025. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 215–229. Association for Computational Linguistics.

Morris H. DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Noah E. Friedkin and Eugene C. Johnsen. 2011. *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. Cambridge University Press.

Marta C. González, César A. Hidalgo, and Albert-László Barabási. 2008. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Ryuji Hashimoto, Masahiro Kaneko, Ryosuke Takata, Takehiro Takayanagi, and Kiyoshi Izumi. 2026. From heard to lived opinions: Simulating opinion dynamics with grounded llm agents in economic environments. *arXiv preprint arXiv:2603.26701*.

Sergio Jara-Díaz and Jorge Rosales-Salas. 2015. Understanding time use: Daily or weekly data? *Transportation Research Part A: Policy and Practice*, 76:38–57.

Fatima Koaik, Aayush Gupta, and Farahan Raza Sheikh. 2026. LLM powered social digital twins: A framework for simulating population behavioral response to policy interventions. *CoRR*, abs/2601.06111.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. 2025. Finetuning llms for human behavior prediction in social science experiments. *arXiv preprint arXiv:2509.05830*.

Bibb Latané. 1981. The psychology of social impact. *American Psychologist*, 36(4):343–356.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Liang, Luis Alonso, and Kent Larson. 2025. Simulating society requires simulating thought. *arXiv preprint arXiv:2506.06958*.

Victoria Lin, Taedong Yun, Maja Mataric, John Canny, Arthur Gretton, and Alexander D'Amour. 2026. The illusion of intervention: Your llm-simulated experiment is an observational study. *arXiv preprint arXiv:2605.20767*.

Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tur. 2026. Measuring and mitigating the distributional gap between real and simulated user behaviors. *arXiv preprint arXiv:2605.07847*.

Qirui Mi, Mengyue Yang, Xiangning Yu, Zhiyu Zhao, Cheng Deng, Bo An, Haifeng Zhang, Xu Chen, and Jun Wang. 2025. MF-LLM: simulating collective decision dynamics via a mean-field large language model framework. *CoRR*, abs/2504.21582.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2026. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Comput. Surv.*, 58(11):280:1–280:41.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Tianyi Peng, George Gui, Melanie Brucks, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J. Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, Hortense Fong, Elizabeth Friedman, Ariana Guevara, Mohamed Hussein, Kinshuk Jerath, Bruce Kogut, Akshit Kumar, Kristen Lane, Hannah Li, and 4 others. 2026. Digital twins as funhouse mirrors: Five key distortions. *Preprint*, arXiv:2509.19088.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *CoRR*, abs/2502.08691.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations. *arXiv preprint arXiv:2601.17087*.

Caroline Wang, Daniel Kasenberg, Kim Stachenfeld, and Pablo Samuel Castro. 2026. Discovering differences in strategic behavior between humans and llms. *arXiv preprint arXiv:2602.10324*.

Xiaotong Ye, Nicolas Bougie, Toshihiko Yamasaki, and Narimawa Watanabe. 2026. Mobilecity: An efficient framework for large-scale urban behavior simulation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 5: Industry Track, Rabat, Morocco, March 25-27, 2026*, pages 287–303. Association for Computational Linguistics.

Eviatar Zerubavel. 1985. *The Seven Day Circle: The History and Meaning of the Week*. Free Press, New York.

Yunyao Zhang, Yihao Ai, Zuocheng Ying, Qirui Mi, Junqing Yu, Wei Yang, and Zikai Song. 2026. Coupling macro dynamics and micro states for long-horizon social simulation. *CoRR*, abs/2604.05516.

A Implementation and Reproducibility Details

This appendix summarizes the implementation details required to reproduce the reported simulation conditions. Exact code paths are provided where appropriate; prompt templates are summarized to avoid duplicating long Japanese runtime prompts in the main paper.

License note. AgentSociety is distributed under the Apache License Version 2.0; see the repository LICENSE file for details. MF-MDP is distributed under the MIT License. Public statistics are used under the applicable e-Stat and Japanese government statistics terms of use.

Intended-use note. We use public statistics only for aggregate demographic and spatial grounding, and municipal budget records only as coarse annual alignment references. The external repositories listed above are used as architectural references, not as the exact implementation of the reported experiments. The artifacts created in this work, including synthetic personas, event-worldline configurations, MF-condition configurations, simulation traces, and analysis scripts, are intended for research reproduction and auditing of the reported experiments. They should not be interpreted as measurements of real residents or used outside research contexts for automated policy decisions or individual- level inference.

A.1 Run configuration

We report the configuration in Table 4, Table 5, Table 6, Table 7, Table 8, Table 9 and Table 10.

A.2 Worldlines and MF conditions

Important consistency check. All reported conditions are specified by YAML configuration files. Table 6 maps paper condition names to implementation code names. In particular, if Local-MF and the none worldline are reported in the main text, the released configuration set should include `friend` and none runs.

[1] https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200521&tstat=000001080615&cycle=0&tclass1=000001089055&tclass2=000001089057&tclass3=000001089058&cycle_facet=cycle&tclass4val=0&metadata=1&data=1

A.3 Belief state and stance labels

Each agent maintains one belief object per policy topic:

$$b_{i,k}^{(d)} = \left(I_{i,k}^{(d)}, s_{i,k}^{(d)}, u_{i,k}^{(d)}\right),$$

where $I_{i,k}^{(d)} \in [0,1]$ is an interest score, $s_{i,k}^{(d)} \in \{\texttt{positive}, \texttt{neutral}, \texttt{negative}\}$ is an internal stance label, and $u_{i,k}^{(d)}$ is a first-person natural-language nuance. If $I_{i,k}^{(d)} < 0.3$, the implementation coerces `stance` and `nuance` to `null`.

A.4 Mean-field construction

For Local-MF, the numerical stance distribution is computed over the observer's eligible social neighbors, weighted by intimacy. For Town-MF, the numerical stance distribution is computed uniformly over the eligible town population. For Hybrid-MF, the numerical distribution is

$$m_{i,k}^{\text{hybrid},(d-1)} = (1-\alpha_{i,k}^{(d-1)}) m_{i,k}^{\text{local},(d-1)} + \alpha_{i,k}^{(d-1)} m_{k}^{\text{town},(d-1)}.$$

For textual summaries, Local-MF samples high-scoring neighbor nuances ranked by interest $\times$ intimacy, Town-MF samples town-wide nuances ranked by interest, and Hybrid-MF splits the K samples between local and town pools according to $\alpha_{i,k}$.

If neither local nor town eligible contributors exist for a topic, the implementation uses a uniform no-signal sentinel for the numerical MF distribution. This sentinel is not interpreted as evidence for a balanced public opinion.

A.5 Acquaintance graph construction

The acquaintance graph is generated at run time from `experiments/baseline/data/relationships/koshimizu\_relationships\_100.csv` using `persona\_generate/scripts/build\_acquaintances.py`. Relationship edges are treated as bidirectional. Family relationships are refined using household roles where possible.

A.6 Prompt templates

Nightly reflection prompt. At the end of each simulated day, each agent receives a prompt containing: (1) date, (2) persona summary, (3) the seven policy topics, (4) previous topic beliefs, (5) recent diary entries, (6) today's executed plan, (7) today's conversation summaries, (8) received

Item	Source or artifact	Role in this paper
Census statistics	e-Stat census data portal[1]	Grounding for demographic, household, occupation, and spatial attributes used to construct the synthetic Koshimizu residents.
Simulator reference implementation	https://github.com/tsinghua-fib-lab/agentsociety/	Used as an architectural reference for the LLM-agent simulation stack.
MF reference implementation	https://github.com/AI4SS/MF-MDP	Used as an architectural reference for the mean-field conditioning design.

Table 3: External data sources and implementation references.

Category	Value	Notes
Town	Koshimizu, Hokkaido, Japan	Real town abstracted into a 100-agent resident population.
Agents	100	Loaded from `experiments/baseline/data/personas/personas\_100\_v2\_with\_beliefs.json`.
Policy topics	7	Welfare/childcare, health/medical/sanitation, education, safety/disaster, infrastructure, local economy/employment/industry, administration.
Simulation length	28 days	Four 7-day weeks, interpreted as compressed years 2018–2021.
Tick length	30 minutes	`tick_sec=1800`.
Start time	2018-01-01 04:00	The seed belief state is treated as the pre-event 2017 baseline.
Conversation mode	nightly	Co-location opportunities are collected during the day and summarized before nightly reflection.
Conversation RNG seed	12345	Used for conversation-pair selection. This is distinct from the historical 2017 seed belief state.

Table 4: Main simulation configuration.

event broadcasts, and (9) the optional MF signal. The model is instructed to output strict JSON: namely a `diary` field and a `topics` object whose per-topic entries contain `interest_score`, `stance`, and `nuance`. The stance must be one of `like`, `neutral`, `dislike`, or `null`. The nuance is required to be first-person and grounded in the agent's own background or observations.

MF summary prompt. For each topic, the MF-summary prompt provides the computed stance distribution and up to K representative contributor nuances. The model is instructed to produce a short JSON object mapping each topic to a 2–3 sentence Japanese summary covering: stance distribution, main viewpoints, emotional tone, and topic salience. The prompt explicitly states that the summary must remain consistent with the numerical stance distribution.

A.7 Keyword metric for blackout discourse

Blackout discourse propagation is measured on outlying-district agents outside the direct urban blackout broadcast scope. An agent is counted as using blackout-related discourse if any topic nuance contains a keyword matching the blackout lexicon.

A.8 Simulation Cost

We report the computational cost in Table 11

A.9 AI Assistance Disclosure

We used AI assistants to support code refactoring, debugging, and language polishing. The authors reviewed and edited all AI-assisted outputs and are responsible for the final code, analyses, claims, and manuscript text. Synthetic personas and simulation traces were generated through the documented persona- construction and simulation pipeline; AI assistants were not used as an undocumented source of experimental observations or evaluation labels.

Worldline	Code schedule	Event content
none	none.json	No exogenous event; used as a negative-control worldline if included in the reported experiment matrix.
a	worldline_a.json	Earthquake broadcast to the town in Week 1, followed by an urban-blackout broadcast restricted to town-center districts and later recovery messages.
b	worldline_b.json	Worldline a plus COVID-related town-wide broadcasts beginning in Week 3, including a short emergency-declaration window and later recovery.
c	worldline_c.json	Worldline b plus a Week-1 school-closure event broadcast to the town.

Table 5: Event worldlines. Event scopes are encoded in the disaster schedule files as town or urban.

Paper name	Code name	MF signal
No-MF	baseline	none
Local-MF	friend	social-neighborhood MF
Town-MF	town	town-wide eligible-population MF
Hybrid-MF	combined	interest-weighted local/town mixture

Table 6: Mapping between paper condition names and code names.

Component	Setting	Notes
Main reflection model	openai/gpt-5.4-mini	Configured through AGENTSOCIETY_LLM_MODEL.
MF-summary model	openai/gpt-5.4-mini	Configured through AGENTSOCIETY_NANO_LLM_MODEL.
Nightly reflection temperature	0.5	The model outputs a strict JSON object containing a diary and seven topic beliefs.
MF summary temperature	0.1	The model summarizes representative belief nuances into a per-topic social-context summary.
Nightly reflection output schema	Strict JSON object	The output contains diary and per-topic fields for interest_score, stance, and nuance. The stance field is positive, neutral, negative, or null.
Retries	5	Nightly reflection retries on transient API failures and rate limits.

Table 7: LLM configuration used by the simulation.

Parameter	Value	Description
Interest threshold τ	0.3	Only topic beliefs with interest at least τ and a non-null stance contribute to MF aggregation.
MF warm-up days	1	MF signals are activated after the initial day in the N=100 four-year runs.
Top-K nuances	10	The textual MF summary is generated from up to K representative nuances per topic.
Self-intimacy	1.0	The observer's own previous-day belief is included in the numerical MF. The observer's own nuance is not quoted in textual summaries.
Hybrid weight	$\alpha_{i,k} = I_{i,k}^{(d-1)}$	Higher topic interest increases the town-wide component in Hybrid-MF.

Table 8: MF construction parameters.

Relationship type	Intimacy weight w_{ij}
family / spouse / parent / child / sibling	0.8
coworker	0.5
neighbor	0.3
unknown / other	0.03
self	1.0

Table 9: Relationship-to-intimacy mapping used for Local-MF.

Code path	Purpose
`experiments/run_from_yaml.py`	Parses YAML configs, regenerates acquaintances when enabled, and launches the simulation.
`experiments/run_simulation.py`	Runs the tick-based simulation and invokes the MeanFieldEngine at day boundaries.
`mf_sim/agent/skills/night_reflection/scripts/night_reflection.py`	Builds the nightly reflection prompt and writes belief snapshots.
`mf_sim/mean_field/scripts/run_mf_mdp.py`	Computes local, town-wide, and hybrid numerical MF signals and samples nuances for textual summaries.
`mf_sim/mean_field/util_state/update_prompt_state.py`	Generates per-topic textual MF summaries.
`persona_generate/scripts/build_acquaintances.py`	Builds the acquaintance graph from the relationship CSV.

Table 10: Primary code paths used by the reported experiments.

Model	Agents	LLM calls	Time	Cost
`gpt-5.4-mini`	100	$\sim$ 1,436/day	$\sim$ 14.8 min/day	$\sim$ \$1.64/day

Table 11: Representative runtime and API-cost estimate for the N=100 simulation. The estimate is computed under the `gpt-5.4-mini` configuration used in the experiments.