Agentopia on a Consumer GPU: A Reduced-Scale Long-Horizon Port with an 8B Model

Luo Huan
Shenzhen University
Shenzhen, China

Abstract

Large language model (LLM)-based multi-agent social simulation has demonstrated compelling results, but Agentopia was evaluated with 100 agents over 10 simulated years using Qwen3.5-397B-A17B, leaving the behavior of reduced-scale deployments on consumer hardware unclear. In this paper, we implement and evaluate a reduced-scale Agentopia port on a single NVIDIA RTX 5070 Ti (12 GB VRAM) using Qwen3-8B-AWQ, a 4-bit quantized model. We introduce three structural adaptations for this setting: (1) system-managed layered memory compression, (2) four activity blocks per simulated day, and (3) explicit physical- and mental-health state variables. Across three independent stochastic runs, two runs completed 52 weeks and the third completed 50 weeks before reaching the context limit, totaling 154 system-weeks (770 agent-weeks). No agent died, and no threshold-based health warning was logged; activity records containing at least one `NO_RESPONSE` field occurred at rates of 10.15–10.29% across runs. A 52-week memory-off run tied L2/L3 artifact production to layered memory; a separate 10-week comparison associated four daily time blocks with 2.72× more finalized records and lower lexical duplication, but a higher missing-field rate. These comparisons do not support causal behavioral claims. We release validated configurations, derived audits, analysis scripts, aggregate figure data, and our implementation changes in a public fork; raw runs and initial persona data are excluded because their redistribution provenance is not fully resolved.

1 Introduction

Large language model (LLM)-based agents have demonstrated remarkable capabilities in individual tasks, but their behavior in long-term multi-agent social simulations remains underexplored. Agentopia (Wang et al., 2026) recently introduced a framework where 100 LLM-based agents live, interact, and learn in a shared society over 10 simulated years. Its primary experiments use Qwen3.5-397B-A17B and substantially more compute than a consumer laptop. This leaves a practical question unanswered: what scale and duration of Agentopia-style simulation are attainable on hardware accessible to individual researchers?

In this paper, we study a reduced-scale port with five active agents, one apartment world, and a target horizon of one simulated year. We deploy Agentopia on an NVIDIA RTX 5070 Ti (12 GB VRAM) using Qwen3-8B-AWQ, a 4-bit quantized model served via vLLM (Kwon et al., 2023). To operate within this setting, we add three structural adaptations: (1) system-managed layered memory compression, (2) four activity blocks per simulated day, and (3) explicit physical- and mental-health state variables. Across three independent stochastic runs, the system completed 52, 52, and 50 weeks, respectively.

The two full runs reached one year, while the third exposed a context-budget boundary at week 51; activity-record-level `NO_RESPONSE` rates were 10.15–10.29%. Disabling layered memory removed L2/L3 artifacts, establishing an implementation dependency but not a downstream behavioral effect. In a 10-week comparison, the archived M+T run contained 2.72× as many finalized records per agent-week as the archived M run and showed less lexical reuse but more missing fields; the unseeded, one-run-per-condition design prevents causal claims. No deaths or threshold-based health warnings occurred, but incomplete health ablations prevent attribution to that module.

We position this work as an empirical account of a reduced-scale, long-horizon agent society on a consumer GPU, not as a same-scale reproduction of the original Agentopia experiments.

2 Related Work

LLM social simulation systems. Generative Agents established a widely used architecture that combines memory, reflection, and planning in an interactive sandbox (Park et al., 2023). Subsequent systems have emphasized complementary goals: SOTOPIA formalizes interactive evaluation of social intelligence (Zhou et al., 2024b); Agentopia targets 100-agent, multi-year life simulation (Wang et al., 2026); and AgentSociety and GenSim focus on parallelized or general-purpose simulation infrastructure (Zhang et al., 2025; Tang et al., 2025). Rather than competing on population scale, our study asks what longitudinal evidence and failure boundaries can be obtained from an Agentopia-style system on a single 12 GB consumer GPU.

Validity and long-horizon evaluation. LLMs can reproduce selected behavioral-study outcomes or conditional response distributions (Aher et al., 2023; Argyle et al., 2023), but apparent success can depend strongly on information access and evaluation design (Zhou et al., 2024a). Surveys and methodological critiques therefore call for calibration, robustness checks, reproducible protocols, and explicit limits on population-level inference (Gao et al., 2024; Zeng et al., 2026). Recent Social-LLM work similarly audits multi-turn trajectories and separates autonomous action selection from empirical calibration (Star et al., 2026; Lazzaroni et al., 2026). We follow this evidence-bounded view by reporting completed horizons, artifact coverage, lexical reuse, and termination conditions without treating five agents as a representative population.

Memory and resource-constrained deployment. Long-horizon evaluations expose history-dependent challenges across repeated interactions (Goel and Zhu, 2025), while memory-management studies show that how experiences are retained and consolidated can affect later experience-following behavior and computational cost (Xiong et al., 2026; Zhang et al., 2026). Our system-managed L2/L3 pipeline is motivated by this problem, but we evaluate artifact production rather than claiming learning or a causal behavioral benefit. At the deployment layer, AWQ provides weight-only quantization for memory-constrained inference (Lin et al., 2024), and PagedAttention supports memory-efficient LLM serving (Kwon et al., 2023); these are components of our stack, not contributions re-evaluated by this study.

3 System

Figure 1 summarizes the inference stack, weekly simulation loop, and the three structural adaptations used in our reduced-scale port.

3.1 Environment and Baseline

We adopt the Agentopia framework (Wang et al., 2026) with five active agents cohabiting an apartment world for up to 52 simulated weeks. Our implementation retains the framework's plan, contact and scheduling, activity, review, and settlement operations; rewards are computed during settlement according to the configured period.

Our hardware setup consists of an NVIDIA RTX 5070 Ti (12 GB VRAM) running Ubuntu 22.04 under WSL2. We serve Qwen3-8B-AWQ (Yang et al., 2025; Lin et al., 2024), a 4-bit quantized 8-billion-parameter model, via vLLM (Kwon et al., 2023) with a server-side context limit of 24,576 tokens, role model output capped at 3,584 tokens, God model output at 3,840 tokens, temperature 0.7, and concurrency limited to one request (`max_num_seqs=1`) to avoid memory pressure. The archived run configuration records a client-side context value of 28,672 tokens; the effective limit was the lower 24,576-token server boundary. We progressively scaled the simulation through a ladder of tests from T1 (1 agent $\times$ 4 weeks) to T6 (5 agents with a target of 52 weeks). The three T6 runs were independent stochastic restarts, but no global run-level seed was fixed; we therefore do not describe them as controlled-seed replicates.

3.2 Layered Memory Compression

The original Agentopia provides recent weekly diaries together with persistent scratchpad files that agents manage themselves. Detailed diaries outside the recent window leave the prompt unless an agent has promoted salient information into a scratchpad, so long-term retention depends on agent-initiated memory management rather than a systematic compression schedule.

We add a system-managed three-layer memory architecture with progressively coarser representations. **L1** stores complete event logs for the five most recent weeks; agents do not generate these records themselves. **L2** holds $\sim$200-word summaries for intermediate history, produced by the God model from L1 logs. **L3** distills older blocks

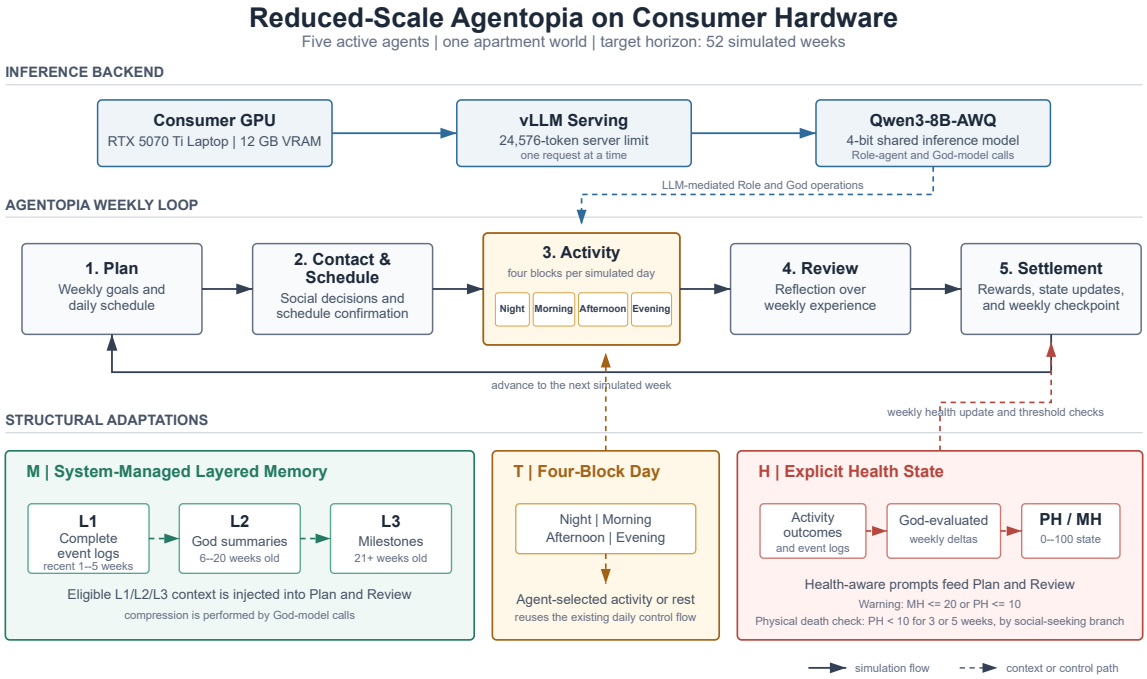

Figure 1: Architecture of the reduced-scale Agentopia port. A shared Qwen3-8B-AWQ endpoint served by vLLM executes Role-agent and God-model calls on a 12 GB consumer GPU. The weekly loop retains planning, contact and scheduling, activity, review, and settlement, while the M/T/H adaptations add system-managed layered memory, four activity blocks per simulated day, and explicit physical- and mental-health state. Solid arrows denote simulation flow; dashed arrows denote context or control paths.

of L2 summaries into ∼50-word milestone entries. When composing plan and reflection prompts, agents receive the currently eligible context from all three layers, providing a system-managed retention path for information that would otherwise depend on scratchpad updates.

In short preliminary runs with fixed review prompts, we observed that Reflection paragraphs froze into verbatim repetition after 3–5 weeks. After the review question set was randomized, the repetitions were no longer observed in subsequent short tests, suggesting that prompt structure may have contributed to the failure mode. The long-run experiments below evaluate whether the layered pipeline continues to produce nonempty, qualitatively varied memory artifacts; they do not isolate its effect on downstream behavior.

3.3 24-Hour Time Granularity

The original framework uses weekly plans with one solo-activity slot per day, and the same planned activity may recur across weekdays. This design limits the representation of within-day rhythms and activity variation.

We divide each day into four time blocks—Night (00:00–06:00), Morning (06:00–12:00), Afternoon (12:00–18:00), and Evening (18:00–24:00). Agents plan an activity or rest choice for each block, allowing different behavior within the same day. The adaptation reuses the existing weekly and daily control flow while iterating over four solo-activity blocks within each day. When disabled (via the ablation flag `time.enable_24h`), the system reverts to the original single-solo-activity-per-day behavior.

3.4 Health System

The original Agentopia models vitality and four fulfillment dimensions, including mood, but it does not maintain explicit physical- and mental-health trajectories or warning and death thresholds.

We introduce two separate health indicators: **physical_health** (0–100, initially 80) and **mental_health** (0–100, initially 70). Physical health is affected by sleep duration, nutrition, and illness events; mental health responds to social interactions—positive encounters (gratitude, recognition) increase it, while conflicts, rejection, and loss decrease it. The God model evaluates health deltas each week based on event logs and activity outcomes. Current scores are exposed to subsequent plan and review prompts through score-

dependent health-awareness text. At each weekly checkpoint, a mental-health score at or below 20 or a physical-health score at or below 10 is logged as a threshold-based health warning. Four consecutive weeks below a mental-health score of 20 additionally produce a severe-depression-risk warning but do not force death. For physical-health scores below 10, agents with low social-seeking are marked deceased after three consecutive weeks, whereas high social-seeking delays this outcome until five consecutive weeks. The profile-derived social-seeking score is the quantitative extraversion trait when available and otherwise $100 -$ introversion; scores of 60 or higher use the high-social-seeking branch.

To reduce template-like health evaluations, one prompt cue is sampled each week from a pool of 48 emotional cues. The health system is controlled via the ablation flag `health.enable`; its trajectories are stored separately from the original fulfillment-based subjective reward computation.

4 Results

We evaluate the full system (all three adaptations enabled, denoted M+T+H) across three independent stochastic restarts lasting 52, 52, and 50 weeks, respectively. All five selected agents are active participants, yielding 154 completed system-weeks and 770 agent-weeks.

4.1 System Stability

The first two runs reached the 52-week target. The third completed 50 weeks but stopped when the week-51 prompt reached the configured context boundary. Before termination, Run 3 wrote 16 finalized activity records for the incomplete week 51. We include these records in activity-record-level metrics because they are present in `activity.jsonl`, but exclude week 51 from completed-week totals and weekly-outcome analyses. No infrastructure-level termination occurred within a completed week; missing activity fields are quantified separately below. The weekly pipeline executed fully in every completed week. These results support feasibility at the tested five-agent scale and identify context growth as a long-horizon limitation observed in Run 3.

4.2 Observed Health Trajectories

Figure 2 shows the weekly physical- and mental-health trajectories, and Table 1 reports the scores at each run's final checkpoint. Across all three runs, **no agent died and no threshold-based health warning was logged**. Physical health remained at or above 93 for all agents at the final checkpoint; the lowest mental-health value recorded across all weeks was 53 (see Table 4 for per-run minima), above the mental-health warning threshold of 20. Because the health-disabled runs are incomplete, these observations characterize the enabled system but do not establish that the health module caused the absence of warnings.

Table 1: Final-checkpoint physical-health (PH) and mental-health (MH) scores. Both range from 0 to 100. Runs 1–2 end at week 52 and Run 3 at week 50; each cell reports PH/MH.

Agent	Run 1 PH/MH	Run 2 PH/MH	Run 3 PH/MH
Whitfield	100/93	95/70	96/71
Morales	100/97	100/79	100/89
Vale	93/92	96/77	100/96
Voss	100/79	100/77	100/99
Vieri	100/95	100/100	100/91

Table 2: Activity-record-level `NO_RESPONSE` incidence. Total records reports the number of finalized activity records, and Flagged records reports the number containing at least one `NO_RESPONSE` field; each record is counted at most once.

Run	Total records	Flagged records	Rate
Run 1	4,412	448	10.15%
Run 2	4,436	451	10.17%
Run 3	4,236	436	10.29%
Pooled	13,084	1,335	10.20%

We measure missing activity fields from the finalized `activity.jsonl` records rather than counting string occurrences in generation transcripts, where the same historical marker can be copied into many later prompts. The pooled activity-record rate in Table 2 is 10.20%, and the sample coefficient of variation across the three unrounded run-level rates is 0.75%. This is an activity-record metric, not a call-level failure rate, because one activity can contain multiple LLM calls.

To audit output reuse separately from literal `NO_RESPONSE` markers, we examined solo records without that activity-level flag. Within each run and agent, normalized primary-text exact-duplicate rates were 14.48%, 17.18%, and 17.52% for Runs 1–3; comparison against the preceding eight weeks using 5-token shingle Jaccard $\geq$ 0.80

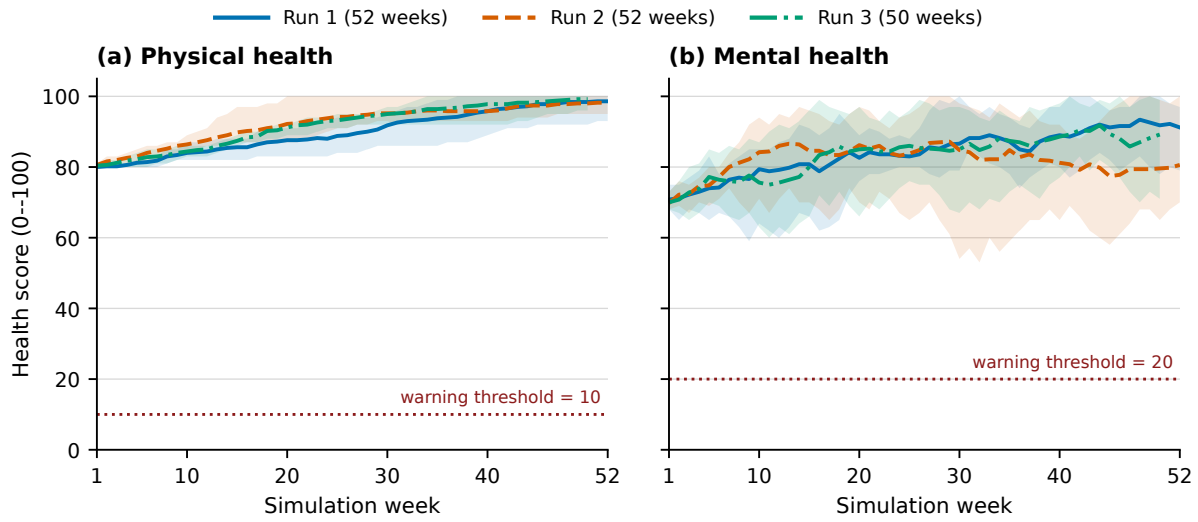

Figure 2: Weekly physical- and mental-health trajectories across the three full-system runs. Lines show the weekly mean over the five active agents, and shaded bands span the agent-level minimum and maximum. Dotted horizontal lines mark the physical-health warning threshold of 10 and mental-health warning threshold of 20. Run 3 ends after week 50.

yielded additional near-duplicate rates of 5.72%, 4.66%, and 4.40%. Normalization applied Unicode NFKC normalization, lowercasing, and punctuation removal, and comparisons were restricted to the same activity type. These lexical screens expose output reuse but do not measure semantic or behavioral diversity.

4.3 Memory Artifacts and a Short Time Comparison

Across the stochastic runs, Aaron Whitfield's week-40 L2 entries emphasized, respectively, deadline-driven family guilt, avoidance through writing, and peer conversations prompting a promised visit. In qualitative inspection, these entries appeared non-template-like and contextually grounded. Each active agent has 47, 47, and 45 L2 entries in Runs 1–3, respectively (Run 3 produces two fewer because it is two weeks shorter), for 695 L2 and 90 L3 artifacts in total. Given the configured five-week recent window, all 15 agent-run archives contained an L2 entry for every expected older week; this establishes continuous artifact production, not successful recall or downstream use. These examples illustrate that the stored summaries preserve run-specific histories; they do not establish an effect on downstream behavioral diversity.

We also audited two archived 10-week, five-agent conditions: layered memory alone (M) and memory plus four daily time blocks (M+T), with health disabled. Their saved configurations differ only in `world.time.enable_24h` and output directory; Table 3 reports activity volume, missing fields, and lexical reuse.

Table 3: Exploratory 10-week time toggle (one run per condition; health disabled). Rec./A-W is finalized records per agent-week; NR is activity-record-level `NO_RESPONSE`. Exact/near are lexical duplicate rates in unflagged solo primary texts.

Condition	Weeks	Rec./A-W	NR	Exact	Near
M (T off)	10	6.36	1.57%	57.60%	16.40%
M+T	10	17.28	7.75%	10.06%	6.45%

M+T contains 2.72× as many records per agent-week as M, reflecting its larger activity budget rather than higher quality; all five agents have lower exact-duplicate and higher NR rates. With one unseeded, unpaired run per condition and different activity opportunity counts, these results are descriptive and do not establish improved semantic or behavioral diversity.

4.4 Cross-Run Summary

Table 4 aligns the main run-level metrics with each run's end condition. Across runs, minimum PH was 80 in all cases, minimum MH ranged from 53 to 60, and activity-record-level `NO_RESPONSE` rates ranged from 10.15% to 10.29%; Run 3 ended two weeks earlier at the context boundary. Table 5 summarizes coverage and completion status for the configurations discussed below.

Table 4: Run-level summary. NR is the activity-record-level NO_RESPONSE rate, D/W denotes deaths/threshold-based health warnings, and health minima cover all completed weeks. L2 entries/agent is the final-checkpoint count per active agent. Aggregate weeks are system-weeks, and aggregate NR is pooled over records.

Run	Weeks	End condition	Activity records	NR rate	Min. PH	Min. MH	D/W	L2 entries/agent
Run 1	52	Target reached	4,412	10.15%	80	59	0/0	47
Run 2	52	Target reached	4,436	10.17%	80	53	0/0	47
Run 3	50	Context boundary	4,236	10.29%	80	60	0/0	45
Aggregate	154	–	13,084	10.20%	80	53	0/0	–

Table 5: Completion status and observed outcomes for the full-system runs, ablation conditions, and 4-week all-off baseline. M, T, and H denote layered memory, four-block daily time, and health.

Configuration	Runs	Weeks completed	Status or termination	Observed outcome
Full (M+T+H)	3	52, 52, 50	Planned horizon completed (2); context boundary (1)	Zero deaths/warnings; L2/L3 artifacts present
−Memory (T+H)	1	52	Planned 52-week horizon completed	L2/L3 artifacts absent by configuration
−Health (M+T)	2	21, 38	Incomplete: week 21, resume-path slowdown; week 38, WSL I/O failure	No complete 52-week health comparison available
Time toggle (M vs. M+T)	1 each	10, 10	10-week comparison complete; 52-week M+H not run	2.72× records/A-W; lower reuse, higher NR; descriptive only
All off	1	4	Planned 4-week baseline completed	Longitudinal comparison unsupported; health warning/death mechanisms disabled

5 Discussion

5.1 Ablation Analysis

We completed one 52-week memory ablation, one 10-week time-toggle comparison, and two incomplete attempts to disable the health module; we did not complete a 52-week M+H leave-one-out comparison for the time adaptation. The memory-off run (T+H enabled) completed 52 weeks with zero deaths and zero threshold-based health warnings, showing that this configuration remained operational without layered memory. Physical health ranged from 98 to 100 across agents at the final checkpoint, and mental health from 74 to 98. As expected from the configuration, its L2 and L3 directories were empty and no compressed summaries were available to agent prompts. This confirms the implementation dependency between the module and the retained long-term memory artifacts. It does **not**, by itself, establish that layered memory causes greater downstream behavioral or narrative diversity; that claim requires a matched comparison of agent outputs under enabled and disabled conditions.

The short M versus M+T comparison exposes a temporal-granularity trade-off—more activity and lower literal reuse, but more missing fields—rather than improved behavioral diversity, because opportunity counts differ and each condition has one unseeded run.

The two health-off runs (M+T enabled) did not complete the full 52-week horizon: the first terminated at week 21 due to a resume-path performance degradation, and the second at week 38 due to a WSL I/O failure. Because disabling the module also removes its health-specific warning and death mechanisms, and both runs are incomplete, we do not compare their health outcomes with the full system. A complete matched ablation remains for future work.

The all-off configuration of our reduced-scale port completed the planned 4-week baseline. Its short horizon is useful as an implementation reference but is insufficient for a longitudinal reliability comparison, so we do not use it to claim that the three adaptations collectively reduce failures.

6 Conclusion

On a single RTX 5070 Ti (12 GB), our five-agent system completed 52, 52, and 50 weeks, totaling 154 system-weeks (770 agent-weeks), with no deaths or threshold-based health warnings. The memory-off run tied L2/L3 artifacts to layered memory, while the short time toggle exposed a trade-off among activity volume, lexical reuse, and missing fields. These observations do not establish

causal behavioral effects, and the health module's contribution remains unresolved.

Limitations

Several limitations warrant discussion. First, our experiments use a single model (Qwen3-8B-AWQ) for both agent roles and God evaluations; a multi-model setup might yield richer agent differentiation. Second, the concurrency limit serializes inference requests; at the tested five-agent scale, a 52-week run took roughly 33–40 hours. We did not measure scaling beyond this setting. Third, we evaluate only the apartment world setting; generalization to other Agentopia worlds (school, workplace) remains untested. The physical- and mental-health variables are simulator-internal state variables; they are non-clinical and have not been validated against medical or psychometric constructs. Fourth, limited compute prevented a matched 52-week baseline and exhaustive ablations; additionally, we did not fix a global run-level random seed. Fifth, our time-toggle comparison covers only 10 weeks, with one unseeded run per condition and health disabled; it is not a 52-week M+H leave-one-out ablation, so the time module's causal contribution to behavioral diversity remains unquantified. Sixth, our narrative evidence consists of stored-memory examples and counts rather than an automated or human-rated measure of downstream behavioral diversity. Seventh, the unflagged solo-output audit identifies lexical reuse but does not determine whether repeated wording corresponds to repeated actions, intentions, or social outcomes. Finally, the `NO_RESPONSE` statistic is defined over finalized activity records, not individual model calls, and therefore characterizes missing activity fields rather than serving as a direct inference-server failure rate. The public research artifact at `https://github.com/luo675/Agentopia/tree/paper` is maintained as a fork of the upstream Agentopia repository and contains our implementation changes, validated configurations, derived audits, analysis scripts, and aggregate figure data. It excludes raw run archives, initial persona data, model weights, and credentials; the upstream project and its attribution remain linked through the fork relationship.

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

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

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259.

Hitesh Goel and Hao Zhu. 2025. LIFELONG SOTOPIA: Evaluating social intelligence of language agents over lifelong social interactions. *arXiv preprint arXiv:2506.12666*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626.

Ruggero Marino Lazzaroni, Lorenz Prattes, and Jana Lasser. 2026. Calibrated but autonomous: Inference-time bayesian logit correction for LLM social simulations. SocialLLM Workshop at ICWSM 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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*, pages 1–22.

Michelle Star, Andrew Aquilina, and Yu-Ru Lin. 2026. Auditing support strategies in LLMs through grounded multi-turn social simulation. SocialLLM Workshop at ICWSM 2026.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2025. GenSim: A general social simulation platform with large language model based agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 143–150.

Xintao Wang, Sirui Zheng, Hongqiu Wu, Weiyuan Li, Jen-tse Huang, Minghao Zhu, Can Zu, Qi Deng, Jiawei Wang, Qianyu He, Heng Wang, Xiaojian Wu, and Yunzhe Tao. 2026. Agentopia: Long-term life simulation and learning in agent societies. *arXiv preprint arXiv:2606.07513*.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. 2026. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–645.

An Yang et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yongchao Zeng, Calum Brown, and Mark Rounsevell. 2026. Too human to model: the uncanny valley of large language models in simulating human systems. *npj Complexity*, 3:13.

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. 2026. Lightweight LLM agent memory with small language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12914–12929.

Jun Zhang, Yuwei Yan, Junbo Yan, Zhiheng Zheng, Jinghua Piao, Depeng Jin, and Yong Li. 2025. A parallelized framework for simulating large-scale LLM agents with realistic environments and interactions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1339–1349.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*.