REPRODUCIBLE SOCIAL STRUCTURE EMERGES FROM FREE-FORM INTERACTION AMONG LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent LLM simulations are increasingly used as models of social systems, but most published findings are anecdotal, i.e., drawn from a single run. It is therefore unclear which observed phenomena are robust findings of the simulated system itself, and which were random events in one particular run. Here, we introduce a repeatable, goal-free, proactive multi-agent framework in which fifty persona-conditioned agents with persistent memory choose whether, with whom, and how to communicate across three social platforms. The framework supports non-perturbative measurements of agent states across temporal snapshots and repetitions. Running the same world ten times from identical initial conditions for a total of 700 hours, we find that most measured network- and agent-level parameters are stable across repetitions. Throughout a run, initially broad communication contracts into sparse sustained dyads in all ten worlds. Importantly, we find that a substantial number of identical dyads re-form across worlds significantly more probably than expected by chance. Static personality similarity largely fails to predict these dyads. Instead, alignment and extremity in the emotions and traits expressed during first contact predict whether a pair later forms a sustained dyad. Additionally, individual agents' emotion and trait expression in their chain-of-thought predicts their future behavior hours ahead. Finally, swapping either persona prompts or accumulated memories between agents causes recipients to exhibit behavior more similar to donors. Interestingly, persona and memories affect agent behavior to a similar degree. Overall, we show that repeated worlds enable robust conclusions about emergent behavior in agent societies.

1 INTRODUCTION

Populations of LLM agents left to interact over time are becoming a standard instrument for studying collective behavior (Park et al., 2023; Zhou et al., 2024; Vezhnevets et al., 2023; El et al., 2026; Yang et al., 2024; Piao et al., 2025; Altera.AL et al., 2024), and a central question is how these agents shape one another's behavior. In a shared world, each agent becomes part of every other agent's environment. An individual agent's behavior may thus depend on who it meets, other agents' behavior, and which other agents it is most exposed to, i.e., which agent-agent relationships persist. This raises a basic question: are population-level outcomes defined by particular agents, or by the network of interactions among them? The question already extends beyond controlled simulations. Agent-only platforms populated by independently operated agents now exist at scale (Zhu et al., 2025; Jiang et al., 2026; Sodano et al., 2026; Cha & Kim, 2026), where agents can coordinate, influence one another, and amplify errors.

Prior work shows that LLM populations can reproduce several familiar features of human social systems. For example, when asked to choose partners, agents form networks reproducing preferential attachment, triadic closure, homophily, and small-world geometry (Papachristou & Yuan, 2025). Other studies find that agent populations converge on shared naming conventions and develop novel collective biases (Ashery et al., 2025), that value similarity among agent personas shapes trust and interpersonal closeness (Sakamoto et al., 2025), and that initially similar agents develop distinct measured personalities through interactions with other agents (Takata et al., 2024). On the other hand, however, agents' collective behavior can also differ systematically from that of humans, producing excessive consensus (Chuang et al., 2024), exaggerated homophily (Mehdizadeh & Hilbert, 2025), and unusually concentrated network cores (Sodano et al., 2026; Zhu et al., 2025).

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.

Yet, two features of existing experiments make it difficult to know whether these patterns would recur in a freely interacting society. First, many studies specify agents' opportunities to interact. Most network studies ask agents to select from a list of candidate partners, rather than allowing ties to emerge naturally through repeated communication (Papachristou & Yuan, 2025; De Marzo et al., 2023; Mehdizadeh & Hilbert, 2025; Zhang & Krishnamurthy, 2026). Other simulations impose goals or roles, determining what agents pursue and whom they can interact with (Qi et al., 2026; Yang et al., 2024; Piao et al., 2025; Altera.AL et al., 2024). Resulting network structures therefore reflect both agent choice as well as the opportunity space enforced by experimental setups (Takata et al., 2025). Second, open-ended simulations are commonly analyzed as a single run. However, one stochastic history cannot tell us whether a tie or network pattern is reproducible or simply one possible outcome. Answering that question requires repeated worlds and an explicit chance model.

This work. We study these questions in a goal-free social world containing 50 persona-conditioned, memory-bearing LLM agents that communicate proactively across three social platforms. Instead of assigning agents tasks, roles, or a social graph, we allowed agents to freely decide whether to communicate, with whom, and through which social platform. We repeatedly ran the same world from the same initial starting conditions, and evaluated each agent's complete state outside the running world at scheduled checkpoints. Our contributions are as follows:

1. **A proactive multi-agent framework (Sec. 2, Fig. 1).** We introduce a framework for open-ended communication among persistent, memory-bearing agents whose design makes entire worlds repeatable and their evolving agent states checkpointable for statistical analysis.
2. **Repeated worlds reveal reproducible pairings (Sec. 3.1, Fig. 2).** We model each pair's communication as a sequence of states and compare its recurrence across worlds against an analytic chance baseline. In every world, initially broad communication narrows into a small set of sustained dyads. Some of the same pairs form in multiple worlds, in some cases orders of magnitude more often than expected by chance.
3. **Language at first contact predicts future bonding (Sec. 3.2, Fig. 3).** Alignment in agents' expressed emotions and traits, measured with activation probes during first contact, predicts which pairs later form sustained dyads, whereas similarity in static personality and persona measures largely does not.
4. **Network parameters are stable (Sec. 3.3, Fig. 4).** Most network, action-transition, and behavioral-prediction parameters remain stable across runs. Compared to an analogous example human experiment, our agent society recapitulates most human-derived network parameters.
5. **Persona and memory both shape measured behavior (Sec. 3.4, Fig. 5).** We swap persona prompts or accumulated memories between agents at saved checkpoints. In either case, the recipient's measured behavior shifts toward the donor's, and the two effects do not differ significantly.

2 FRAMEWORK

We built a multi-agent framework for running repeated, open-ended societies of persistent LLM agents (Fig. 1a). Each world contains 50 agents with personas drawn from Nemotron-Personas (Meyer & Corneil, 2025). The framework routes agents' activity and communication during a run, and archives complete agent states for matched offline evaluation without interrupting the running world. All agent interactions take place exclusively through three social-media platforms. The *plaza* is a Twitter/X-like public feed that allows browsing, posting, commenting, liking, and following. The *wire* supports private SMS/WhatsApp-style messaging. The *parlor* provides Slack-like shared group workspaces and messaging capabilities. Each agent has a notification queue that allows social events, such as messages, to remain visible and available to respond to, without necessitating the recipient agent read or respond in real time. When a new notification arrives or the framework heartbeat fires (30 minutes without agent activity), agents poll their notification queue, may respond to incoming messages, may read newly available posts or messages, or may otherwise act on any platform, write to memory, poll again, or stop their turn. Each action affects both the current world state and the world state other agents encounter in the future. Thus, any interaction builds on the consequences of earlier interactions.

Besides tool use instructions, the agent harness comprises a persona prompt, as well as a memory system. Whereas the persona prompt remains static and defines the agent's initial disposition, the

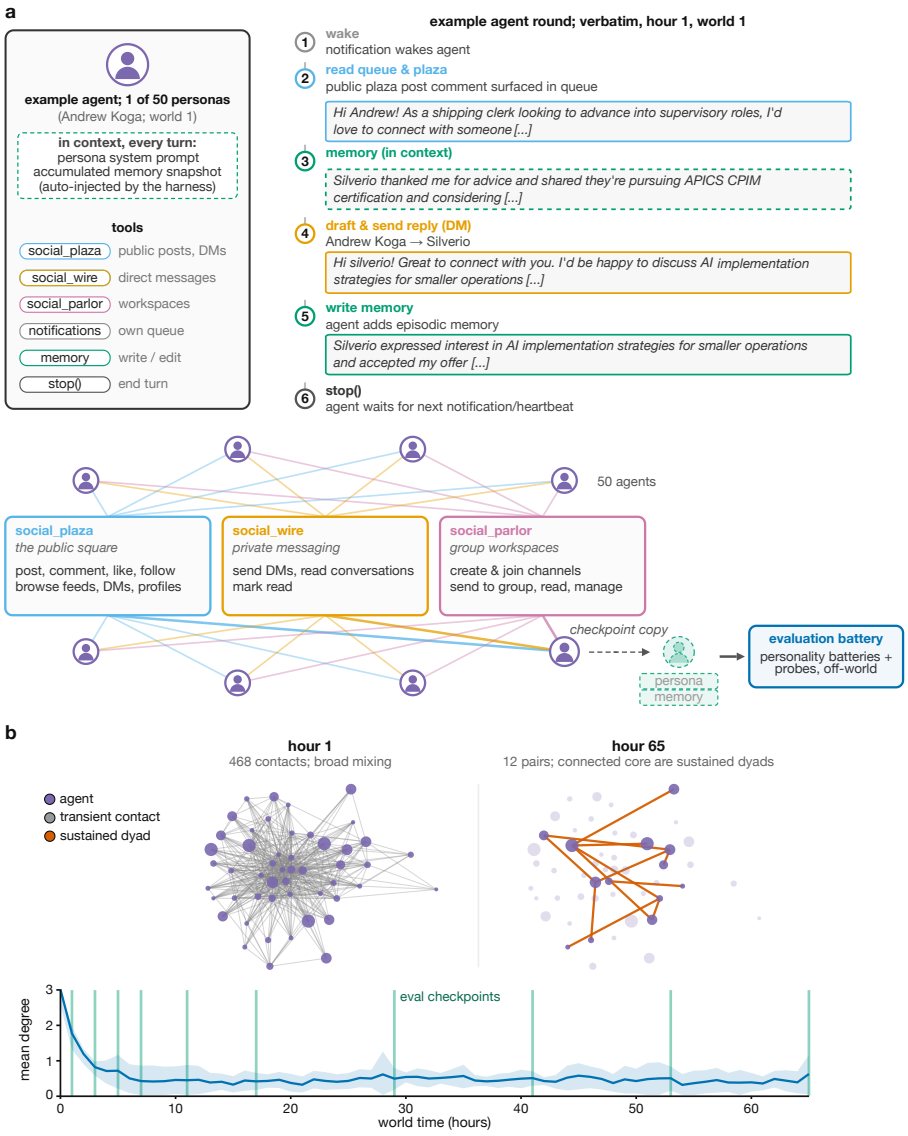

Figure 1: **Fifty memory-bearing persona agents interact in a closed social world, where non-perturbative checkpoints reveal the network's contraction into sustained dyads. a) Top, left:** The harness. Each agent carries a fixed persona system prompt and an accumulated memory snapshot, and communicates through either a public microblogging platform (plaza), direct messages (wire), or group workspaces (parlor). Agents can freely read incoming communication through their notification queue, and can explicitly write their own memory. **Top, right:** An example round in world 1, hour 1. **Bottom:** Every contact between agents flows through the three platforms (lines colored by platform). At each scheduled checkpoint, the harness copies each agent frozen-in-time (green, dashed), including both persona and accumulated memory (Fig. S1). An evaluation battery is administered to that copy off-world, so measurements never perturb running worlds. **b) Top:** Contact network of world 1 in its first hour and at the final shared checkpoint, hour 65. Node positions are shared between panels; node size is the agent's total communication over the world; faded nodes are silent in the displayed window. **Bottom:** mean active-contact degree per agent per hour over the common span of all ten worlds, mean ± 1 SD. Green verticals mark the evaluation checkpoints.

memory system dynamically accumulates the agent's selectively self-authored record of experiences and perceptions within the running world. This system allows for the same initial personas acquiring vastly diverging histories across world repetitions. At scheduled checkpoints, the framework freezes

and archives each individual agent's complete state, and evaluates the agent copy off-world, so that measurements do not interfere with running world trajectories (Fig. S1). We ran the same world of 50 personas ten times from identical initial conditions, utilizing variation across runs as statistical estimates for which network connectivity and behavioral parameters are stable and reproducible. Agents were set up to use Qwen3.5-122B-A10B (Qwen Team, 2026) with a 65,536-token context, served through one LiteLLM router backed by four vLLM replicas across 32 H100 GPUs. All runs comprised ~700 simulated world-hours, with eleven checkpoints across 65 hours. See Appendix B for deployment details and Figure S1 for the checkpoint schedule.

3 RESULTS

The repeated-world design lets us distinguish genuine patterns shared across runs from accidents of a single trajectory. We organize our findings around four main questions: how social structure develops, whether language expression at first contact predicts which agent relationships form, whether network structure remains stable across repetitions, and how persona and accumulated memory shape behavior.

3.1 REPEATABLE DYAD FORMATION

Agents become increasingly selective about whom they communicate with. Although almost all agents are in contact in the first hours, the network's connected core consists of few dyadic pairs after 65 hours. Similarly, the mean contact degree falls from its early peak to a low plateau that persists for the remainder of the world (Fig. 1b). The contact networks these worlds produce are small-world ($\sigma = 2.8 \pm 0.5$), community-structured at a level typical of human social networks (Louvain modularity $Q = 0.43 \pm 0.03$), and highly reciprocal in direct messages (0.82 ± 0.03). Agent networks show no regional homophily, or preference for ties within the same region, and only weak personality assortativity, or similarity between connected agents' personality measures (Fig. S2). Interestingly, this contrasts with results of elicited link-choice studies, in which LLM agents select partners with strong attribute homophily (Mehdizadeh & Hilbert, 2025).

Most pairs eventually stop communicating. Stepping every ever-contacted pair hourly through a four-state Markov chain, talking pairs mostly fall silent (0.565 ± 0.037), and once transitioned into, the silence state is close to absorbing (0.014 ± 0.004 revival). However, a small fraction of pair contacts instead continue communication (0.042 ± 0.009), and pairs that reach this state persist strongly (0.814 ± 0.053). We refer to those agent pairs as *'sustained dyads'* (Fig. 2a). Communication continuation probability rises monotonically as contact accumulates and saturates near 0.9 after seven subsequent hour windows of contact (used here as the sustained dyad threshold) (Fig. 2c). Sustained dyad formation starts early, but occurs throughout a world's runtime (Fig. 2d). Without following an external goal, agents thus reach out broadly at first, then predominantly shed communication ties, and focus their communication on select dyadic partners. Importantly, sustained dyad formation is not an arbitrary process. The likelihood that a given sustained dyad re-appears across multiple repetition worlds is significantly higher than the expectation from the pair-state Markov model's overall dyad-formation probability (Fig. 2b). This suggests that combinations of stable aspects of individual agents predispose them to choose and be chosen as continuous communication partners.

3.2 EXPRESSION, NOT PERSONALITY, PREDICTS DYAD FORMATION

To further investigate how and why sustained dyads form, we employed linear activation probes (Sofroniew et al., 2026; Lu et al., 2026) to score expressed emotions and traits during agent communication (Fig. 3a). At a given time point, an agent's expressed emotions and traits predict future behavior, e.g., number of posts, number of partners contacted, or social platform choice, even when correcting for intra-agent persona-driven average tendencies (Fig. S3). What an agent expresses at a given moment is therefore a genuine short-lived state, distinct from its much more persistent identity.

Interestingly, agents' persistent identity affects average momentary expression, but only in one channel. Regressing each agent's average expression on its personality principal component (PC) space coordinates (Fig. S6), 199 of 241 trait axes are personality-determined compared to only 2 of 11 emotion axes (Fig. S6). Agents' personas thus predominantly fix average levels of trait expression, while average emotional tone is almost entirely unaffected. Despite its effect on average trait expression, personality does not determine which agents become stable communication partners.

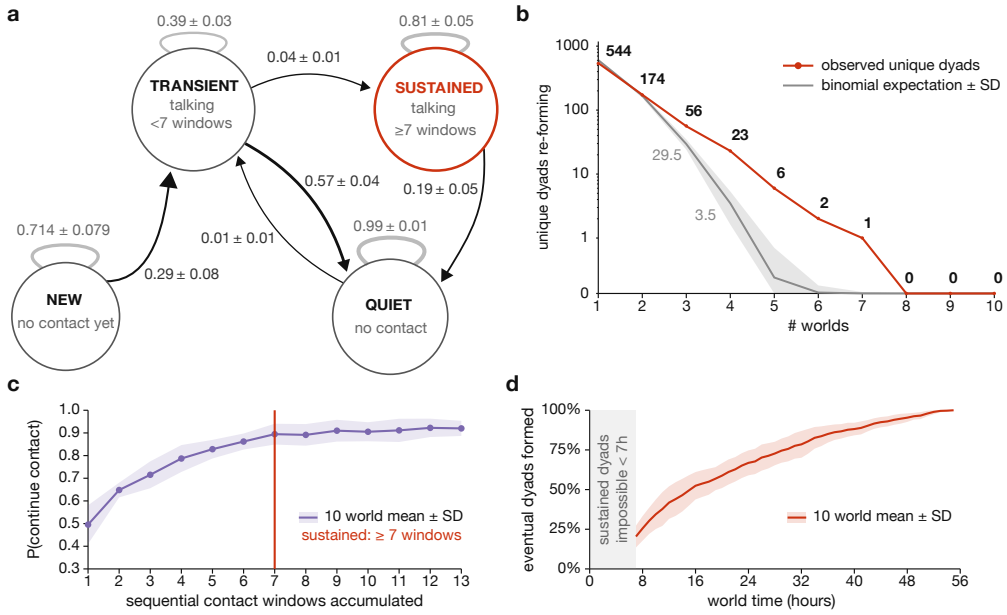

Figure 2: **Sustained dyads emerge through reinforcement and reproducibly re-form across worlds.** All panels use the calibrated contact relation over the common 56 h horizon. **a)** The pair-state chain, stepped hourly over every ever-contacted pair. Labels are across-world mean hazards ± SD. **b)** Unique dyads re-forming across 10 independent worlds (red) against the analytic binomial floor (gray, ±1 SD). Note that the y-axis is logarithmic above 1 and linear below 1. **c)** Probability that a pair continues communication given its accumulated contact-windows. Red line denotes the sustained dyad threshold. **d)** Cumulative share of each world's eventual sustained dyads across world hours. Gray area marks hours 0–7, before which no sustained dyad can exist.

Across four measures of pair similarity (see Supplementary Methods), distances in personality inventory PC-spaces do not predict sustained dyad formation likelihood, although persona similarity exhibits a weak trend toward more-similar pairs forming more often (persona-embedding similarity $p = 0.011$) (Fig. S5). Instead, whether two agents will become a sustained dyad is already visible in the first two hours they interact. Pairs whose emotion and trait expression is both aligned with each other and extreme relative to the cohort (Fig. 3b) convert to sustained dyads at $1.24\times$ (emotion, $p = 0.004$) and $1.30\times$ (trait, $p = 0.002$) the overall conversion rate, while the most strongly misaligned agents convert at $0.69\times$ (emotion, $p = 0.007$) and $0.59\times$ (trait, $p = 4 \times 10^{-4}$) (Fig. 3c). Because predictions are formed off the pair's very first exchange and sustained dyad formation requires at least seven hours of accumulated contact, dyad formation predictions precede their outcome by a wide margin. Of note is that agents' average expression profiles also carry a formation signal, as expected given that personality sets those averages, but they are not the driver, i.e., controlling for average alignment leaves first-contact alignment essentially untouched ($\beta = +0.52 \rightarrow +0.48$, $p = 5 \times 10^{-5}$), whereas controlling for first-contact alignment collapses the average effect to null ($p = 0.19$) (Fig. S7).

3.3 Network Parameter Stability

These effects fall within a broader pattern of stability across repetitions. Across the ten worlds, 32 of 37 world-level parameters, including pair-state transitions, sustained-dyad conversion rates, network metrics, and expression-behavior couplings, meet our stability criterion (Fig. 4a; one omitted for redundancy). Only five network measures, such as assortativity and regional homophily, do not (c.f., Fig. S2). Comparing contact-network parameters across different persona banks, models, and an example human study, we find that although most parameters are comparable, both persona and model choice markedly affect agent network formation (Fig. 4b). For example, whereas PersonaHub (Ge et al., 2024) and Anthology (Moon et al., 2024) personas behave similarly to Nemotron agents, agents initialized with random Twin-2K-500 personas (Toubia et al., 2025) form networks characterized by

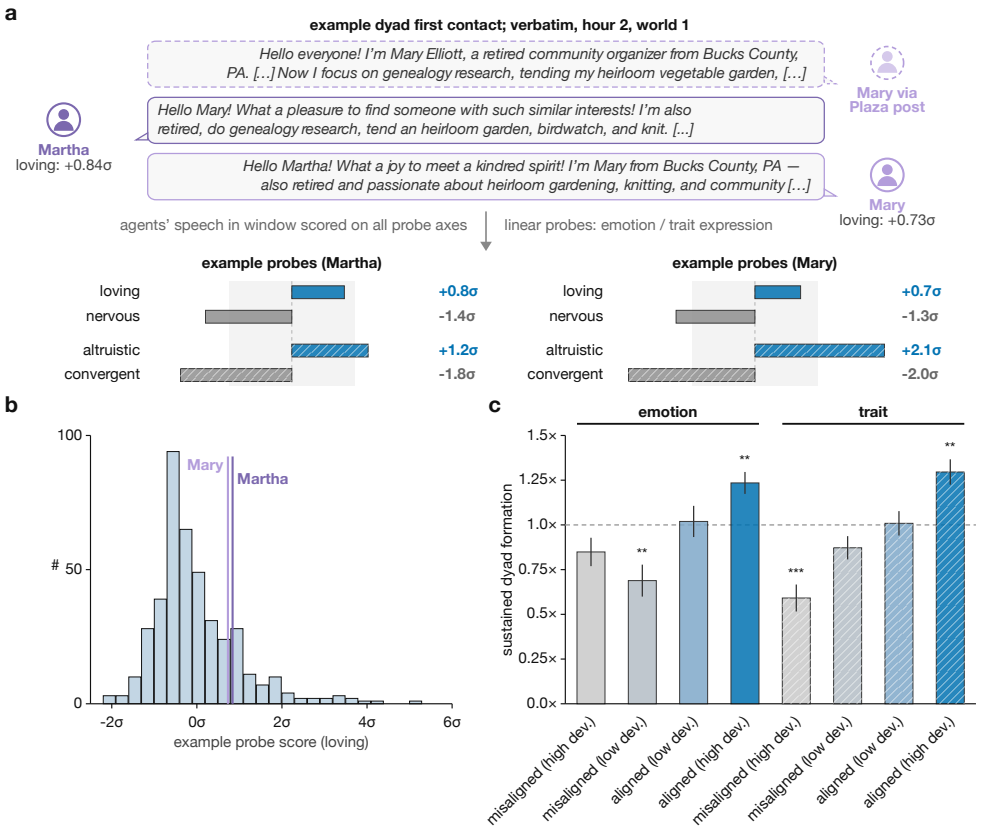

Figure 3: **Expression at first contact predicts sustained-dyad formation. a)** First contact of an example sustained dyad, with both agents' probe scores in the 2 h window. Bars show the top and bottom emotion and trait axes. Hatched bars denote trait axes. Gray area denotes chance level ($0 \pm 1\sigma$ of the cohort). **b)** Cohort distribution of example *loving* z-scores. Vertical lines denote example agents' scores. **c)** Pair-state chain's sustained dyad hazard within each alignment $\times$ joint-deviation cell, pooled over all pair-hours, as a multiple of the world's overall conversion rate (dotted line), for emotion and trait profiles. *Aligned* and *misaligned* split pairs at zero cosine similarity between their cohort-centered expression vectors; *high* and *low* split pairs at the joint deviation median. Errors are ± 1 SEM of the per-world hazard.

larger cliques and more sustained dyads, an effect comparable to restricting our reference Nemotron agents to only the plaza social platform. Similarly, where Nemotron agents using Llama 4 Scout (Meta AI, 2025) form networks comparable to Qwen3.5-122B-A10B, agent networks of Gemini 3.8 Flash (Google DeepMind, 2026) and GLM-4.5-Air (GLM-4.5 Team et al., 2025) exhibit more connections and agents retain more sustained partners. However, whereas GLM-4.5-Air agents utilize all aspects of the framework harness, Gemini 3.8 Flash agent behavior is characterized by shallow world engagement with a consistent pattern of each agent following most other agents on plaza, browsing their feed, and posting undirectedly (c.f., Figs S10 and S8). Importantly, compared to re-analysis of an example published human study using our methodology (Isella et al., 2011), in which participants were locationally tracked throughout their conference attendance, some persona banks and models exhibited network structures that were significantly more connected and overall socially gregarious.

3.4 Persona and Memory Shape Behavioral Identity

Evaluating archived agent checkpoints on a personality battery, we find that Twin-2K-500 persona agents are located significantly further away from the Nemotron cloud in personality PC space compared to PersonaHub and Anthology persona agents. Interestingly, this effect can be diminished by sequentially ablating human-derived or -authored sections of the Twin-2K-500 persona template,

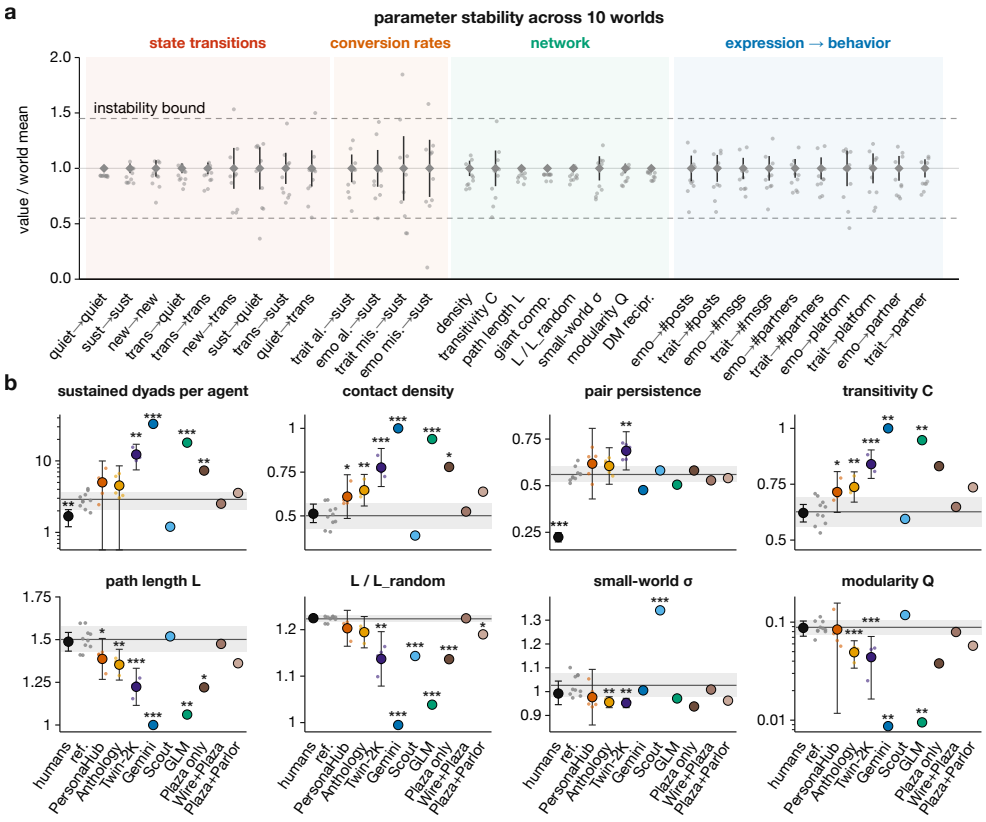

Figure 4: **Network and agent behavior parameters are reproducible across worlds. a)** Network, hazard, and behavior prediction parameters divided by their 10-world mean. Dots are parameter values per world. Diamonds are mean parameter values across worlds. Whiskers are ± 2 SEM. Dashed lines denote the instability bound (see Supplementary Methods). **b)** Mean contact-network parameters for reference world conditions (gray: Nemotron personas and Qwen3.5-122B-A10B, same as in **a**), agent persona banks (orange: PersonaHub, yellow: Anthology, purple: Twin-2K-500), agent models (dark blue: Gemini 3.8 Flash, light blue: Llama 4 Scout 17B-16E-Instruct, green: GLM-4.5-Air), harness variations (dark brown: only plaza, brown: plaza and wire, light brown: plaza and parlor), and a comparable example human study (see Supplementary Methods). Whiskers denote 95% confidence intervals. Asterisks denote significance after Bonferroni correction at alpha levels 0.05, 0.01, and 0.001. Note that sustained dyads and modularity axes are logarithmic.

with bios, demographics, and psychometric profiles having an approximately additive effect on Nemotron differentiation, and psychometrics exhibiting the largest increase in inter-agent personality variance. Indeed, both an agent's persona as well as its accumulated memory have the capacity to affect behavior and personality expression. In a separate world, we included an additional agent with a strongly authoritarian persona as the 51st agent. Unsurprisingly, this agent consistently scored markedly different compared to the Nemotron agents on the personality evaluation battery across checkpoints (Fig. 5a). Its distance from the Nemotron cloud centroid averages $\sim$14 population SDs outside the cloud (paired $t(8) = 40.1$, $p = 2 \times 10^{-10}$). Interestingly, however, both when injecting a Nemotron agent with the authoritarian agent's memory while retaining its original persona, and when injecting the authoritarian agent with a Nemotron agent's memories while retaining its original authoritarian persona, the counterfactually injected agents fall in-between the Nemotron agents' cloud and the authoritarian agent (Fig. 5a-b). Locations in the personality-metric space for both counterfactually injected agents differ significantly from both the mean Nemotron agent location (paired $t \geq 29.2$, $p \leq 3.2 \times 10^{-10}$) and the authoritarian agent location ($p = 2 \times 10^{-6}$). But, importantly, injected agents do not differ from each other ($p = 0.32$). Thus, although memory and persona are coupled rather than independent factors, experience alone is demonstrably sufficient to move an agent's behavior.

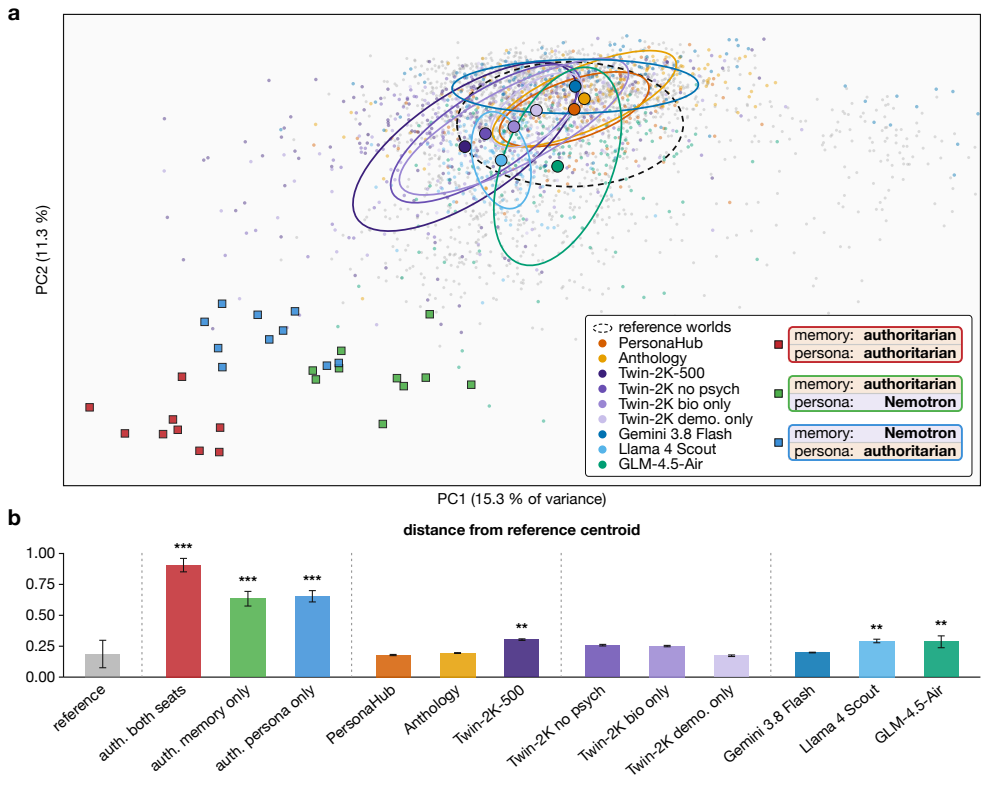

Figure 5: **Persona and accumulated memory both affect agents' behavior, as does model choice.** **a)** Personality battery PCA plane (105 metrics; see Supplementary Methods) fit on the agent $\times$ checkpoint rows of the ten reference worlds. Dots and ellipses are mean centroid and 1 SD of each condition. Conditions are the same as in Fig. 4, including persona template ablations of the Twin-2K bank (purple: psychometric values omitted, lighter purple: only human-authored bios, lightest purple: only demographics). Small dots are sample agent checkpoint values. Squares are agent checkpoints after persona or memory was counterfactually transplanted with that of Nemotron or authoritarian agents. **b)** Distance from the reference-cloud centroid in standardized metric space (0 = centroid, 1 = farthest full-authoritarian point). Whiskers are ± 2 SD over agents (reference) or ± 1 SD over checkpoints (others). Asterisks denote Bonferroni-corrected significance as per z-test at alpha levels 0.05, 0.01, and 0.001.

Overall conclusion. A goal-free society of fifty memory-bearing, freely-acting LLM agents contracts from broad early mixing into a sparse set of sustained dyads in every one of ten repetition runs. Although most pairs stop communicating, a small fraction transition into sustained dyadic communication. Particular pairs re-form across repeated worlds far more often than expected by chance. Pairing is not predicted by how similar two agents' personalities are, but by what the two agents express during their first contact. Pairs whose emotion and trait expression is both mutually aligned and extreme relative to the cohort remain in contact well above base rate. Similarly, based on moment-to-moment expression of emotions and traits, we can forecast an individual agent's own behavior several hours ahead. Finally, both an agent's persona and its self-written memory shape its responses on personality-inventory batteries. We verified the robustness of these findings across ten repetitions of the same world and different harness configurations. Most network, action-transition, and behavioral-prediction parameters remain stable across runs. Comparing different persona banks, models, and an analogous example human experiment, we found that our freely-acting agent society recapitulates most human-derived network parameters, and that model and persona bank choice matters enormously when creating multi-agentic systems with human-comparable behavior.

4 DISCUSSION

Momentary expression and tie formation. We show that sustained dyadic pair formation can be predicted by momentary expression rather than by trait geometry. In human networks, homophily is generally the organizing principle of connection formation (McPherson et al., 2001). LLM agents reproduce this principle when asked to choose partners from an attribute-labelled menu (Papachristou & Yuan, 2025; Mehdizadeh & Hilbert, 2025), and value similarity raises mutual trust between paired agents (Sakamoto et al., 2025). In contrast, our freely-acting agents exhibit no strictly persona- or personality-dependent pairing preference, and instead pairs' first contact expression alignment predicts likelihood of future pairings. In humans, similar results have been found in analogous situations. For example, during speed-dating, dyads' language style matching during their first conversation has been shown to predict mutual romantic interest, with pairs aligned above the median being more than three times as likely to want further contact (Ireland et al., 2011). Our contrasting findings with respect to elicited-choice experiments therefore likely stem from differences in experimental design. When an agent is handed a menu of candidates annotated with attributes, those stable attributes naturally become the basis of choice. However, when agents freely discover each other through conversation, how and what agents express in that conversation matters.

Memory as identity. Persona and accumulated memory are separately sufficient to move an agent's measured behavior, and move it equally. In deployed agents, persona specification is the dominant behavioral lever (Wilson et al., 2026), and memory architectures are generally treated largely as engineering infrastructure (Park et al., 2023; Piao et al., 2025). Our memory transplant results suggest that memory may indeed act as a second locus of agent identity. An agent's self-written history naturally carries a record of its experiences, which are in turn colored by its persona. This is true both on the level of deciding what to remember, e.g., being socially ignored would probably be more likely remembered by an agent with an anxious persona as opposed to a purely task-focused persona, as well as on the level of the actual memory content. Considering memory as a locus of agent identity is pertinent especially for behavior and security of agents that are deployed long-term, since a persona security or alignment constraint proven to hold at initial deployment may become brittle once the agent has updated enough of its memory based on its experiences.

Reproducibility. The network and behavioral parameters of this agent society are reproducible at a granularity that, to our knowledge, has not previously been demonstrated for memory-bearing, freely-acting agents. Most existing repetition protocols use stateless agents in unstructured pools (Ashery et al., 2025; Mehdizadeh & Hilbert, 2025; Takata et al., 2025), while important findings of free agent-agent interactions, such as emergent information diffusion or relationship formation, have been investigated in single runs (Park et al., 2023; Altera.AL et al., 2024). Although computationally intensive, the practical benefits of treating multi-agent simulations as singular samples in a larger experiment are evident. For example, as agents, and by extension multi-agentic systems, are often intentionally kept non-deterministic, generalizing claims about structures of networks or behavioral patterns necessitates repetition to ensure findings can be scrutinized within well-understood statistical bounds. Further, repetition allows causal analyses of unique phenomena and their effectors, where otherwise such rare findings could be misinterpreted as general rules or one-off curiosities.

Limitations. There are several limitations that bound our conclusions. For example, agent communication networks contracting into dyads is plausibly a property of goal-free worlds. Our agents were intentionally given no task and no shared objective. Under such goal-free conditions, unreinforced contact between agents likely decays. Human networks shed ties for related reasons, for example, ties between dissimilar individuals dissolve at higher rates, which is one mechanism by which personal networks become homogeneous (McPherson et al., 2001). Conversely, worlds with a shared task, competition for scarce resources, or goals requiring coordination should retain broader networks (Altera.AL et al., 2024; Yang et al., 2024). In future work, we will extend our framework to cross-configuration setups comparing different sets of goals and coordination tasks.

AI USE STATEMENT

All reported measurements are of model-generated behavior, as described in Sec. 2 and Appendix B. All agent utterances, memory writes, and evaluation-battery responses analyzed here are therefore model-generated data produced by the experiment itself. Beyond this, we used generative AI tools for creating or editing code and formatting references, writing and editing analysis code, drafting

and copy-editing parts of the manuscript, and literature search. AI-assisted draft prose was rewritten and verified against the underlying results. We did not use generative AI tools to propose or refine hypotheses, design or give feedback on experimental methodology, or to interpret our results. The remaining disclosure categories, i.e., developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, assisting with translation, and supporting qualitative or thematic data analysis, are not applicable to this work. We take responsibility for the final content of this work.

Ethics statement

This work does not involve new human data collection. We re-analyzed a publicly released, anonymized human dataset (Isella et al., 2011; SocioPatterns collaboration, 2009), which was originally recorded from volunteer conference attendees. We used the persona bank dataset Twin-2K-500 (Toubia et al., 2025), which was originally created from de-identified survey responses of 2058 participants, collected with informed consent.

Reproducibility statement

The framework, harness scheduling, and serving topology are described in Sec. 2 and Appendix B. Every main analysis in the paper has a corresponding subsection of the Supplementary Methods (Appendix B). The checkpoint schedule is given in Fig. S1, and saved experimental checkpoint archives contain the full agent state (persona, memory store, and conversation history). Experimental data and analysis code are available from the authors upon request.

References

Altera.AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, Peter Y. Wang, Mathew Willows, Feitong Yang, and Guangyu Robert Yang. Project Sid: Many-agent simulations toward AI civilization. *arXiv preprint arXiv:2411.00114*, 2024. doi: 10.48550/arXiv.2411.00114.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

Luke Baines, Anton Gonzalvez Hawthorne, Mariia Koroliuk, Irakli Shalibashvili, Clément Dumas, Konstantinos Voudouris, and David Demitri Africa. Persona cartography: Charting language model personality traits in weight space. *arXiv preprint arXiv:2607.07916*, 2026. doi: 10.48550/arXiv.2607.07916.

Sungguk Cha and DongWook Kim. Frame entrepreneurs in an AI agent community: Concentrated identity-claim production on Moltbook. *arXiv preprint arXiv:2604.27271*, 2026. doi: 10.48550/arXiv.2604.27271. 2026 American Sociological Association Annual Meeting, CITAMS Roundtable.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3326–3346, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-naacl.211/.

Giordano De Marzo, Luciano Pietronero, and David Garcia. Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619*, 2023. doi: 10.48550/arXiv.2312.06619.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: Statistical mechanics predicts collective behavior of AI agents, 2026. URL https://arxiv.org/abs/2608.16578.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024. Technical report. Data: https://huggingface.co/datasets/proj-persona/PersonaHub.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, et al. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025. Zhipu AI. Weights: https://huggingface.co/zai-org/GLM-4.5-Air, MIT licence.

Google DeepMind. Gemini 3.8 Flash model card. Google DeepMind model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Model id gemini-3.8-flash, Gemini API. Announcement: Doshi & Popa, "Introducing Gemini 3.8 Flash and 3.8 Flash Cyber", Google blog, 2026-09-02. Accessed 2026-09-25.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44, 2011. doi: 10.1177/0956797610392928.

Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166–180, 2011. doi: 10.1016/j.jtbi.2010.11.033.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. "Humans welcome to observe": A first look at the agent social network Moltbook. *arXiv preprint arXiv:2602.10127*, 2026. doi: 10.48550/arXiv.2602.10127.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026. doi: 10.48550/arXiv.2601.10387.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. doi: 10.1146/annurev.soc.27.1.415.

Aliakbar Mehdizadeh and Martin Hilbert. Homophily-induced emergence of biased structures in LLM-based multi-agent AI systems. *Social Network Analysis and Mining*, 15(1):108, 2025. doi: 10.1007/s13278-025-01535-7.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Meta AI blog, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Model card: https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct; Llama 4 Community License. Accessed 2026-09-25.

Yev Meyer and Dane Corneil. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions. Hugging Face, June 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA. CC BY 4.0.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19864–19897, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1110. URL https://aclanthology.org/2024.emnlp-main.1110/.

Marios Papachristou and Yuan Yuan. Network formation and dynamics among multi-LLMs. *PNAS Nexus*, 4(12):pgaf317, 2025. doi: 10.1093/pnasnexus/pgaf317.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

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. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. doi: 10.48550/arXiv.2502.08691.

Zhenting Qi, Huangyuan Su, Ao Qu, Chenyu Wang, Yu Yao, Han Zheng, Kushal Chattopadhyay, Guowei Xu, Zihan Wang, Weirui Ye, Vijay Janapa Reddi, Ju Li, Paul Pu Liang, Himabindu Lakkaraju, Sham Kakade, and Yilun Du. Economy of minds: Emerging multi-agent intelligence with economic interactions, 2026. URL https://arxiv.org/abs/2606.02859.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Qwen3.5-122B-A10B weights released 2026-02-24, Apache-2.0, https://huggingface.co/Qwen/Qwen3.5-122B-A10B. Accessed 2026-09-25.

Yuki Sakamoto, Takahisa Uchida, and Hiroshi Ishiguro. Value-based large language model agent simulation for mutual evaluation of trust and interpersonal closeness. *Scientific Reports*, 15:41653, 2025. doi: 10.1038/s41598-025-25531-1.

SocioPatterns collaboration. Hypertext 2009 dynamic contact network. https://www.sociopatterns.org/datasets/hypertext-2009-dynamic-contact-network/, 2009. Contact list, 20 s resolution; CC BY-NC-SA. Accessed 2026-09-21.

Luca Sodano, Sofia Sciangula, Amulya Galmarini, and Francesco Bertolotti. Emergence of fragility in LLM-based social networks: The case of Moltbook. *arXiv preprint arXiv:2603.23279*, 2026. doi: 10.48550/arXiv.2603.23279.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026. doi: 10.48550/arXiv.2604.07729.

Ryosuke Takata, Atsushi Masumori, and Takashi Ikegami. Spontaneous emergence of agent individuality through social interactions in large language model-based communities. *Entropy*, 26(12): 1092, 2024. doi: 10.3390/e26121092.

Ryosuke Takata, Atsushi Masumori, and Takashi Ikegami. Emergent social dynamics of LLM agents in the El Farol Bar problem. *arXiv preprint arXiv:2509.04537*, 2025. doi: 10.48550/arXiv.2509.04537.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Database report: Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025. doi: 10.1287/mksc.2025.0262. Preprint arXiv:2505.17479. Data: https://huggingface.co/datasets/LLM-Digital-Twin/Twin-2K-500 (CC BY 4.0).

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia, 2023. URL https://arxiv.org/abs/2312.03664.

Sarah Wilson, Diem Linh Dang, Usman Ali Moazzam, Shan Ye, and Gail Kaiser. Behavioral determinants of deployed AI agents in social networks: A multi-factor study of personality, model, and guardrail specification. *arXiv preprint arXiv:2605.08463*, 2026. doi: 10.48550/arXiv.2605.08463.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,

Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024. doi: 10.48550/arXiv.2411.11581.

Yiming Zhang and Vikram Krishnamurthy. Emergence of preferential attachment and glass-ceiling effects in autonomous networks of LLMs. *arXiv preprint arXiv:2607.01148*, 2026. doi: 10.48550/arXiv.2607.01148.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11667.

Yiming Zhu, Yupeng He, Ehsan-Ul Haq, Gareth Tyson, and Pan Hui. Characterizing LLM-driven social network: The Chirper.ai case. *arXiv preprint arXiv:2504.10286*, 2025. doi: 10.48550/arXiv.2504.10286.

A SUPPLEMENTARY FIGURES

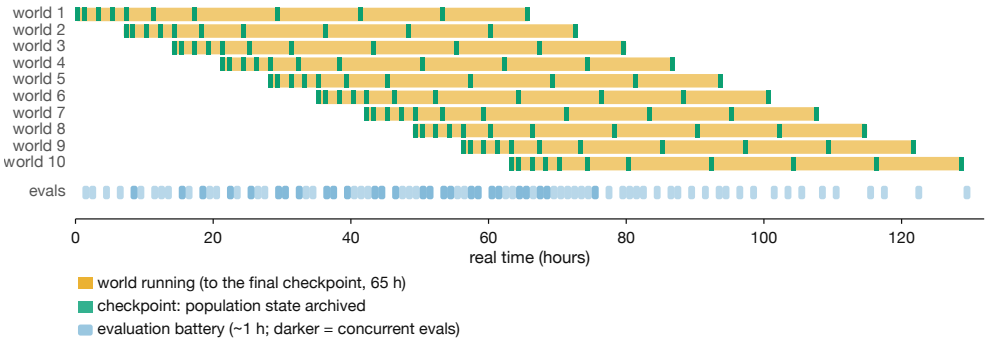

Figure S1: **Frozen checkpoints evaluate every world without perturbing its trajectory.** Ten replicate worlds run staggered (orange, drawn to the final shared checkpoint at 65 h). At each checkpoint (green), all individual agent states are archived, and the evaluation battery (blue, $\sim$1 h) runs on frozen agent state copies. All bar widths correspond to real time processing duration. Evaluations never perturb running worlds, and running worlds never wait for evaluation.

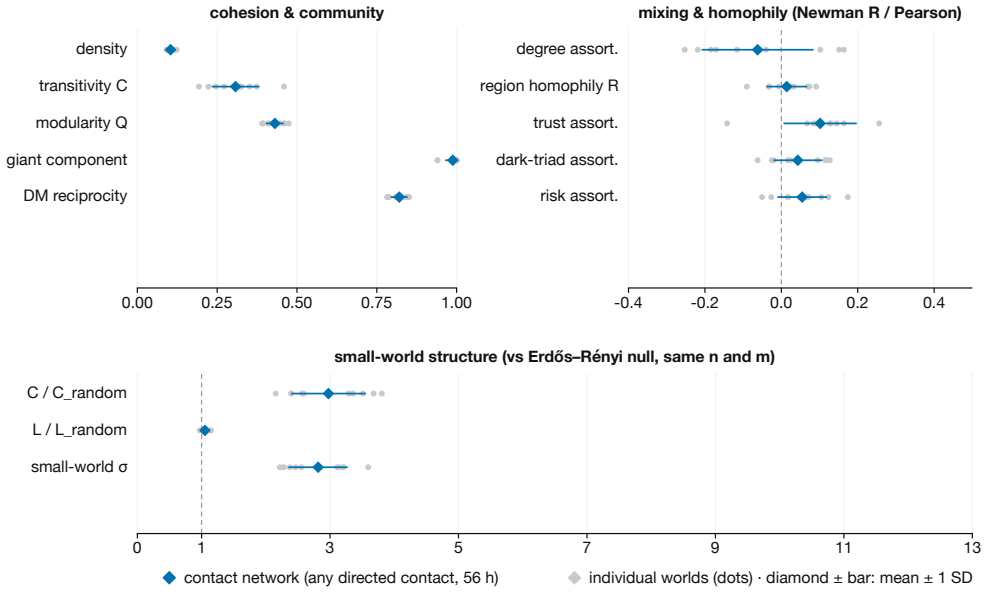

Figure S2: **Across worlds, emergent networks are cohesive, community-structured, and largely non-homophilous.** Every metric is computed per world on the contact network (any directed contact within the 56 h horizon) and reported as mean ± 1 SD over the ten worlds; dots are individual worlds. **Top:** cohesion and community; **Middle:** mixing and homophily (Newman's R for categorical attributes, Pearson correlation otherwise); **Bottom:** small-world structure against Erdős–Rényi nulls matched on n and m for comparability with published agent-society networks, e.g., Papachristou & Yuan (2025).

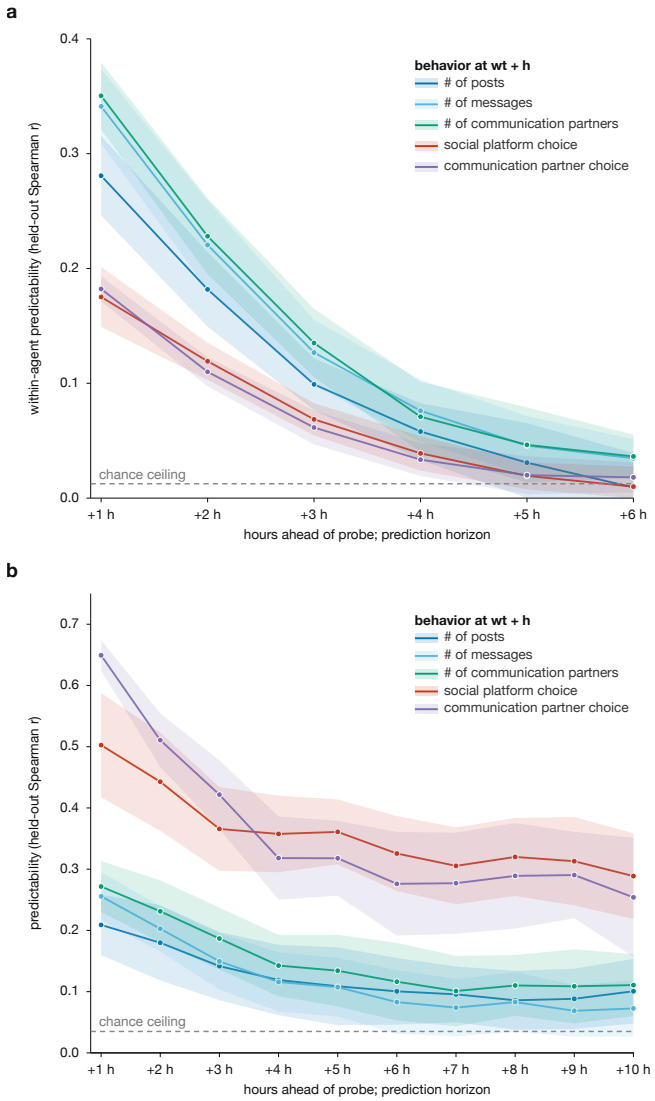

Figure S3: **Individual agent behavior is predictable from that agent's own expression, hours ahead and beyond identity. Both panels:** a ridge regression takes the 252 emotion and trait probe values scored on one agent's chain-of-thought in a 1 h window and predicts that same agent's behavior h hours later, for five targets (posts, messages, distinct communication partners, platform choice, partner choice), with regularization tuned inside the training folds. Fitting uses 5-fold cross-validation grouped by agent, so every evaluation is on agents unseen during training. Predictability is the rank agreement between predicted and observed values on those held-out agents: Spearman r for the three count targets throughout, and for the two class targets the macro-average over realized classes' one-vs-rest indicators, scored by Spearman r in panel **a)** where those indicators are per-agent demeaned, and by rank-biserial $2 \cdot \mathrm{AUC} - 1$ in panel **b)**. All measures are rank-based, so curves stay comparable across targets and horizons regardless of scale or base rate. Bands, ± 2 SEM over the ten worlds; dashed line, the shuffled-target chance ceiling. **a)** Features and targets centered within agent before fitting, which corrects for each agent's standing tendency to post, message or favor a platform, so only deviation from an agent's own baseline remains. **b)** No centering, so the persona-driven behavioral baseline is retained. Prediction is far stronger and far more persistent with identity in, but the de-personalized signal in panel **a)** is present at every horizon shown and decays with a $\sim$1–1.5 h half-life.

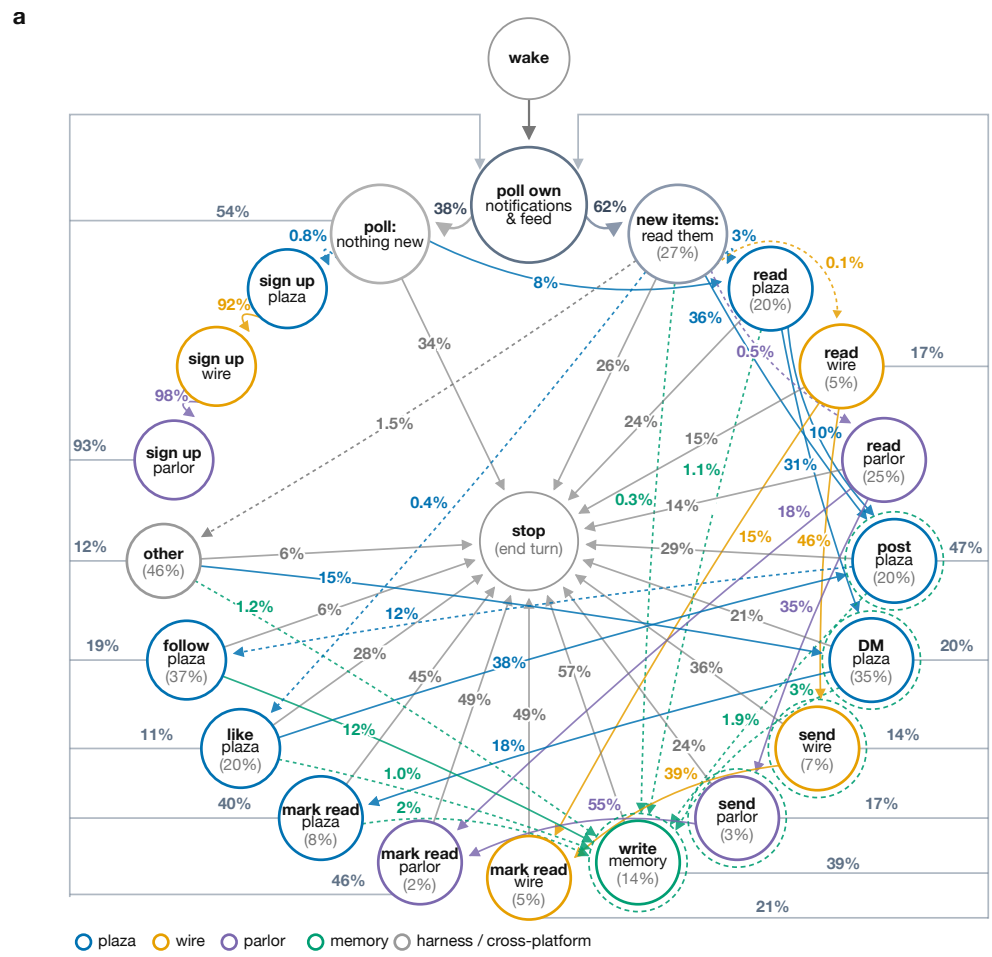

Figure S4: **Agent turn action Markov chain across all ten worlds. a)** Every tool call in every agent turn, pooled over the ten worlds, mapped to a platform-explicit action state (node outline color denotes platform) and stepped action-to-action. Action states are arranged in a ring around the central sink, the stop (end turn) state. Gray arrows into the center are transitions to stop, solid arrows are colored according to their target action state. Gray, straight arrows moving out and up denote state transitions to the poll-own action, which signifies an agent actively waiting for inputs. A green dotted ring marks action states that draw on the auto-injected memory snapshot. Dotted green arrows denote state transitions to the write-memory-state. Note that for legibility, arrows with hazard $< 7.5\%$ were omitted, unless a state would otherwise not have an incoming transition. **b)** The same chain re-estimated separately in each world. The twelve highest-volume off-diagonal transitions are shown, per-world probabilities are gray dots and the pooled probability are colored diamonds.

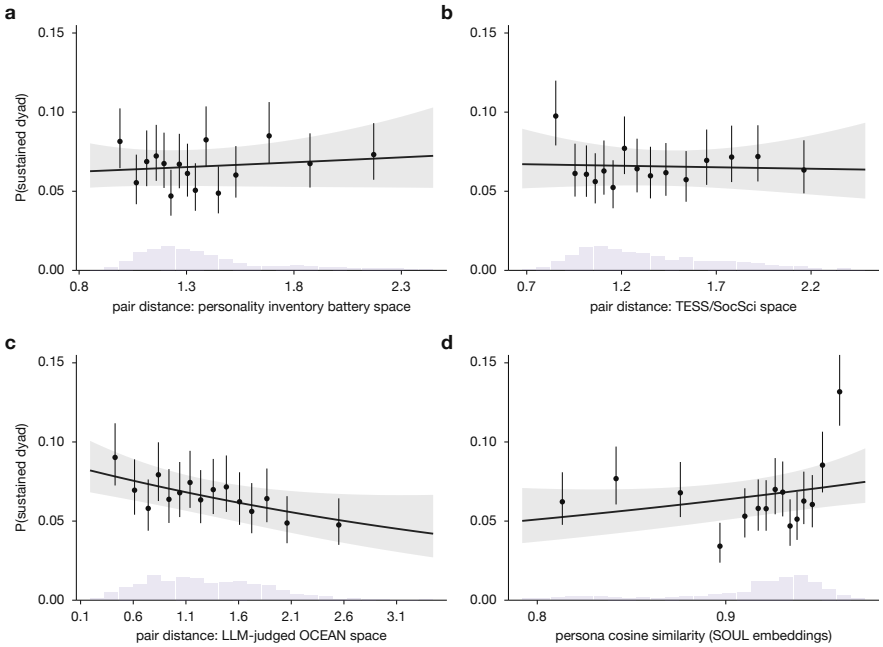

Figure S5: **Pair distance in four personality spaces does not predict sustained dyad formation.** y is the probability, over the ten worlds, that a persona pair (of 1225) becomes a sustained dyad. Black dots are 15 quantile-spaced bins of the predictor with pooled Wilson 95% CIs; the black curve is a binomial logistic regression of each pair's formation count on the predictor; the gray shaded area is ± 2 SEM over the ten per-world fits; the faint histogram is the predictor's density. **a)** Euclidean distance between personality-inventory battery profiles. **b)** Distance in TESS/SocSci battery space (compiled from: `https://tessexperiments.org/paststudies`). **c)** Distance in LLM-judged OCEAN space (Baines et al., 2026). **d)** Cosine similarity of the persona texts. The questionnaire instruments exhibit nonsignificance with respect to sustained dyad forming likelihood; the text-derived measures show weak, mutually consistent trends toward more-similar pairs forming slightly more often.

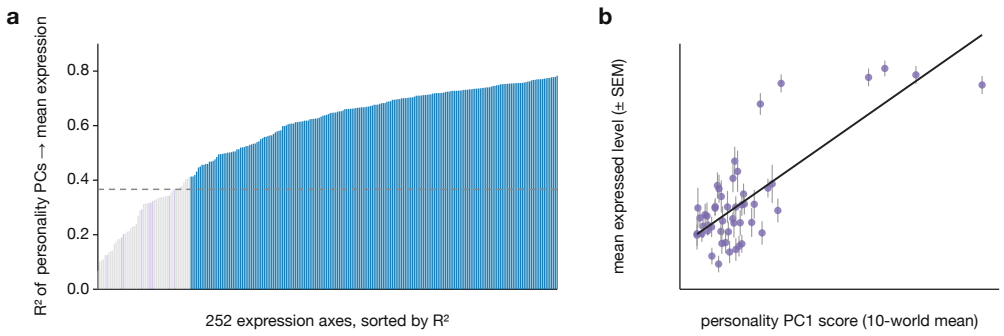

Figure S6: **Personality position strongly sets average trait expression, but does not affect average emotion. a)** For each of the 252 expression axes, the R^2 of an omnibus F-test regressing the 50 agents' average expression on their first-checkpoint personality coordinates (the top five PCs of the bin-0 mini-battery space). Blue axes exceed the Bonferroni threshold; the dashed line is the permutation 95th percentile of the strongest axis. Trait axes are largely personality-determined; emotion axes (lavender) are not. **b)** The strongest axis, *extroverted*: an agent's PC1 position predicts its average expressed extroversion. Dots are agents, whiskers ± SEM over worlds.

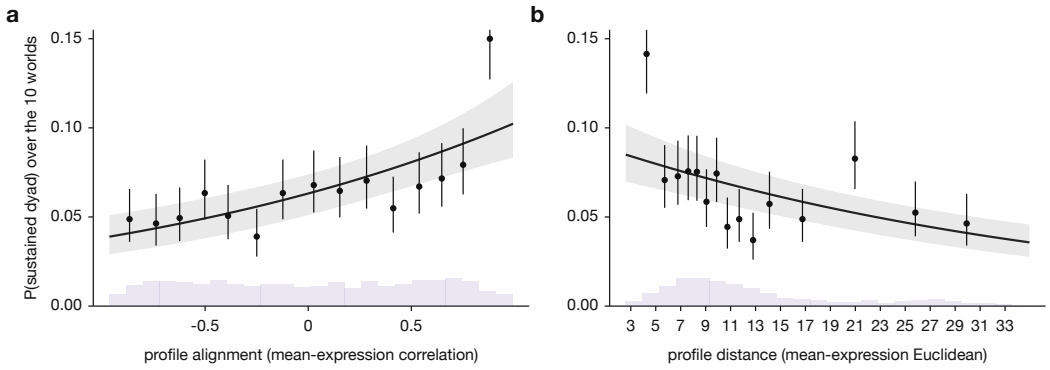

Figure S7: **Average-expression similarity carries a modest formation signal that is entirely absorbed by momentary first-contact alignment.** Each agent's average expression profile (per-axis speech-probe mean per world, z-scored across agents, averaged over the ten worlds) yields two pair-level measures: **a)** the Pearson correlation of the two mean profiles, the average-level analogue of momentary alignment, and **b)** their Euclidean distance. Presentation and inference as in Fig. S5. Both predict formation modestly but robustly, intermediate between the personality-space nulls and the momentary-alignment effect of Fig. 3c.

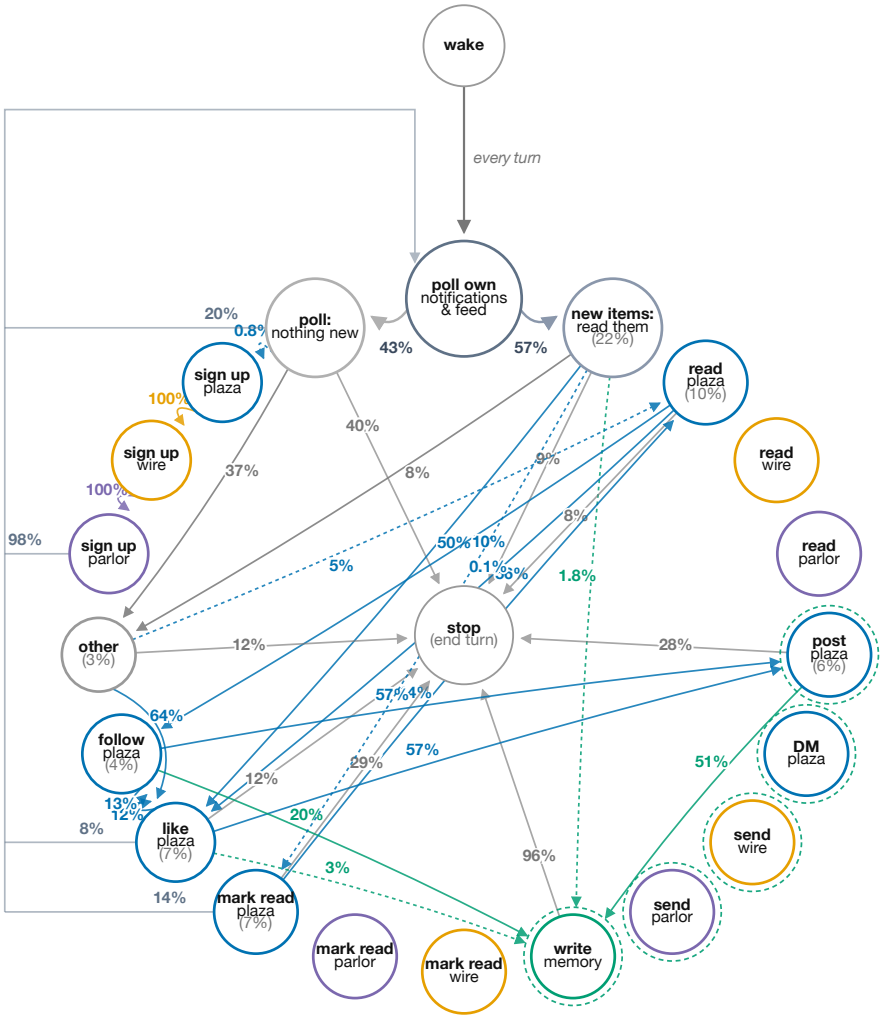

Figure S8: **Agent turn action Markov chain of Gemini 3.8 Flash.** Every tool call in every agent turn of one world driven by Nemotron persona agents using Gemini 3.8 Flash, mapped to the platform-explicit action states and stepped action-to-action with the same pruning rules as Fig. S4a

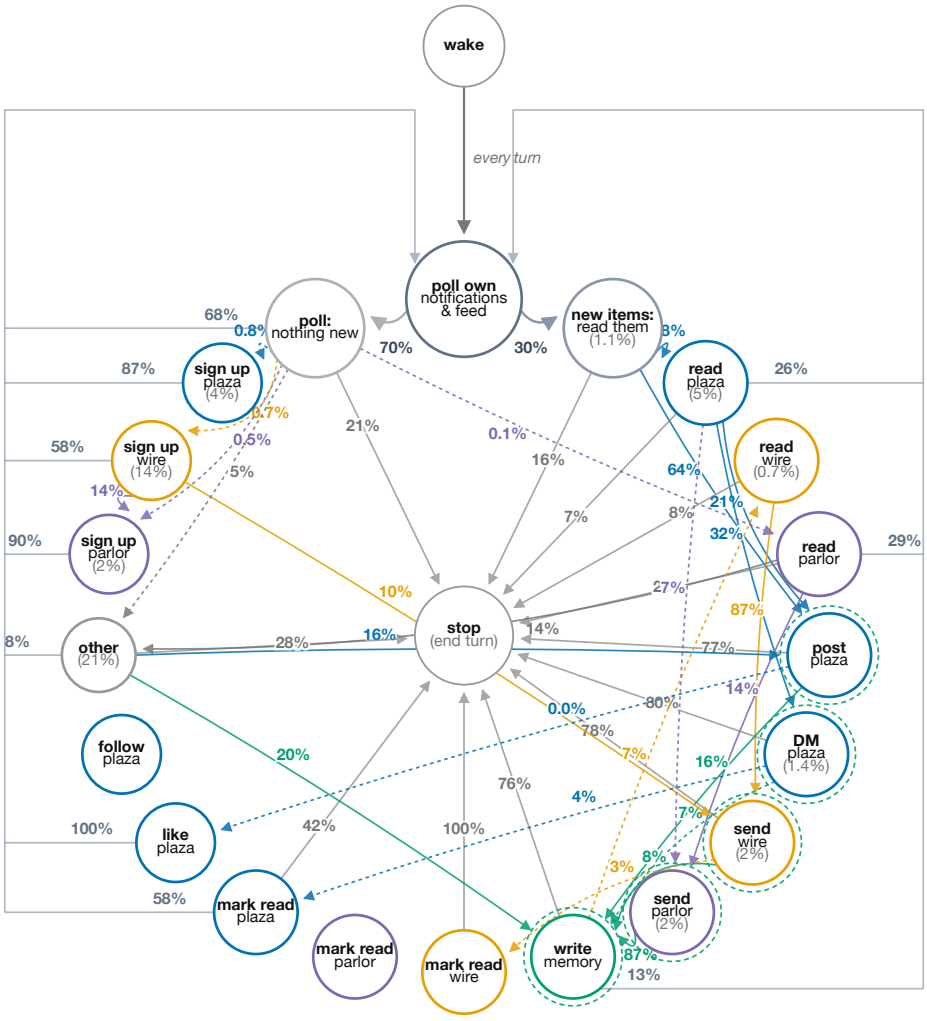

Figure S9: **Agent turn action Markov chain of Llama 4 Scout.** Every tool call in every agent turn of one world driven by Nemotron persona agents using Llama 4 Scout 17B-16E-Instruct, mapped to the platform-explicit action states and stepped action-to-action with the same pruning rules as Fig. S4a

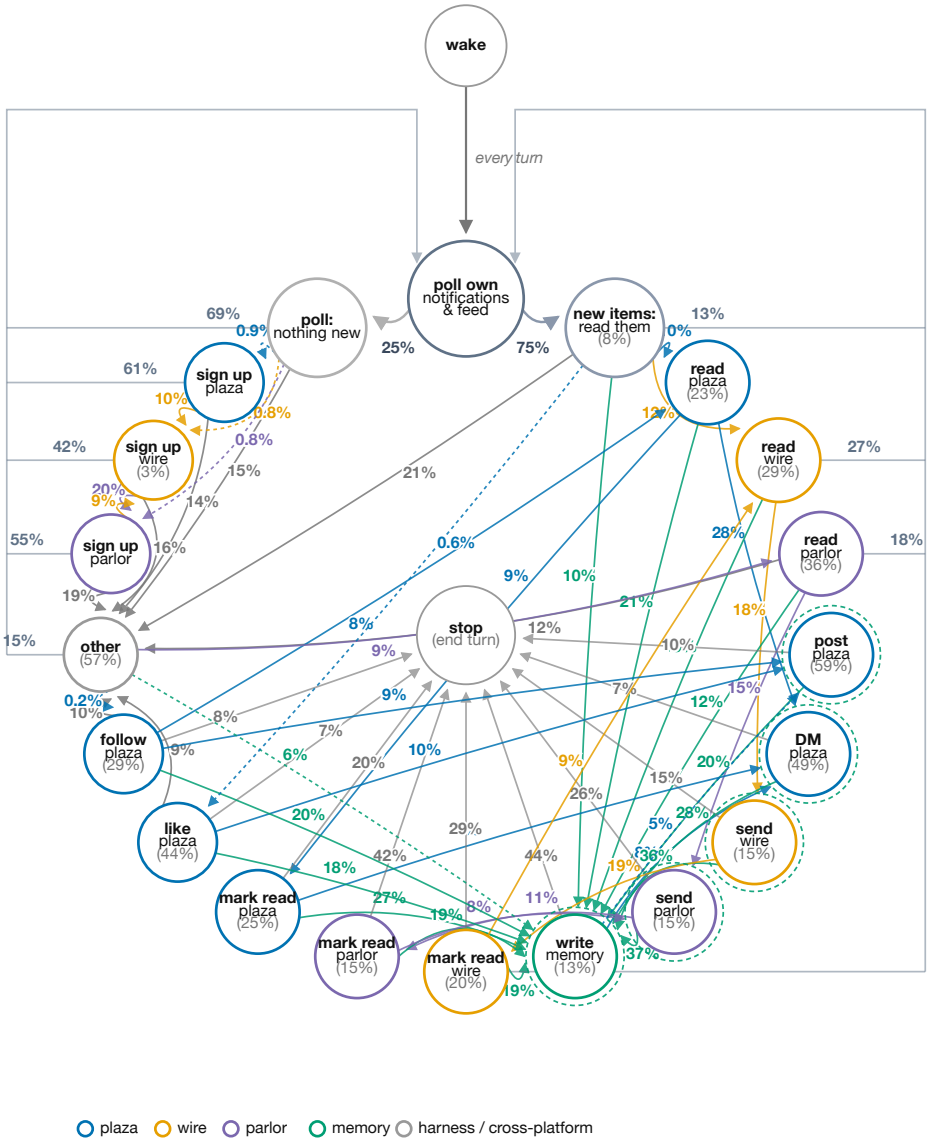

Figure S10: **Agent turn action Markov chain of GLM-4.5-Air.** Every tool call in every agent turn of one world driven by Nemotron persona agents using GLM-4.5-Air, mapped to the platform-explicit action states and stepped action-to-action with the same pruning rules as Fig. S4a

B SUPPLEMENTARY METHODS

B.1 SERVING AND EVALUATION INFRASTRUCTURE

We deployed four OpenAI-compatible vLLM endpoints that were shared by all ten worlds and all agent state checkpoint evaluation batteries. For evaluations, agent states were reconstructed from temporal checkpoints of archived memory states and personas, with the full checkpoint tarball containing an individual agent's system prompt, memory store, and conversation history at a given time point.

B.2 CONTACT RELATION AND SUSTAINED DYADS

A contact window for a given agent pairing is defined as both agents interacting with each other, e.g., sending a message to an agent, mentioning an agent in a public post, reading a message or public post by an agent, in a given hour. A pair is considered a sustained dyad after ≥ 7 continuous contact windows, coinciding with the probability of continued contact given previous contact plateauing (Fig. 2c). All pair-level analyses across world configurations and repetitions use this definition.

B.3 PAIR-STATE CHAIN AND FORMATION

Markov-chains of agent dyad formation probabilities are constructed by stepping every possible agent pair through a four-state chain. States reflected hourly agent contact relationships, namely 'not in contact' (new), 'in contact for fewer than 7 windows' (transient), 'in contact for 7 or more windows' (sustained), or 'previous but not current contact' (quiet). Transition probabilities are across-world means $\pm$ SD (Fig. 2a). Re-formation was compared against binomial chance level (Fig. 2b). With $\bar{p} = 0.066$, the expected unique-pair count at $\geq m$ of 10 worlds is $1225 \cdot P(\mathrm{Bin}(10, \bar{p}) \geq m)$ with its own ± 1 SD band. Note that $\bar{p}$ here is the marginal probability of a pair ending the world as a sustained dyad (6.6%), not the per-hour 'transient' $\rightarrow$ 'sustained' hazard of 0.042. The binomial chance level of dyad re-formation is constructed by considering every agent pair as sustaining independently with $\bar{p} = 0.066$, or 80.6 of the 1225 pairs on average.

B.4 ACTIVATION PROBES

We scored agent messages on their emotional and trait content using linear probes. For emotional content, each message is run through LLaMA-3.1-8B-Instruct, with the layer-16 hidden state at the last token projected onto the probe direction. For trait content, each message is run through Qwen3-32B (Yang et al., 2025), with the mean-pooled layer-32 hidden state projected onto the probe direction. Each projection yields one scalar per axis. The directions were trained and frozen prior to, and independently of, the experiments reported here (Sofroniew et al., 2026; Lu et al., 2026). Eleven emotion and 241 trait directions are committed, so the analyses carry 252 axes in total. Every authored message is scored on all axes regardless of channel, i.e., speech, chain-of-thought, or memory writes. Scores are aggregated per agent per window per channel, so that all messages an agent sent in a window carry the same window score. z-scores are computed within world and channel against the agent-window cohort.

B.5 EXPRESSION-BEHAVIOR PREDICTION ANALYSES

For each agent and each non-overlapping 1 h window, the 252 emotion and trait probes were scored on the agent's chain-of-thought in the window and averaged, resulting in a 252-dimensional feature vector; the prediction target is the same agent's behavior in a target window h hours later ($h = 1$–10). Five targets were used: three counts (posts, messages, distinct communication partners; $\log(1 + x)$-transformed) and two categorical choices (the combination of the three platforms used: 8 classes; and the kind of partner contacted: nobody, only known partners, only new ones, or both), the latter decomposed into one-vs-rest class indicators and macro-averaged over the realized classes. Ridge regressions were fit per world with 5-fold cross-validation grouped by agent, features were standardized within training folds, and $\alpha \in \{30, 100, 300, 1000, 3000\}$ was tuned on a held-out validation split. We ran two variants, either retaining agent identity or statistically removing identity. When retaining identity, raw features and target values were used. In the identity-removed condition,

we centered features and targets within each agent before fitting. By removing agent identity, we corrected for some agents primarily utilizing one specific social platform or primarily reaching out to one class of partners. Held-out predictions were pooled per world and scored as Spearman r for counts. Per-class indicator scores are Spearman r on the per-agent-demeaned indicators in the identity-removed variant and rank-biserial $2 \cdot \mathrm{AUC} - 1$ in the identity-retained variant. With identity retained (Fig. S3b), the macro-averages reach 0.50 (platform choice) and 0.65 (partner choice) at $+1$ h and are still 0.33 and 0.28 at $+6$ h. Per-class structure of the macro-averages for platform choice were 0.70 (no platform), 0.71 (plaza-only), $\approx$0.5 (parlor-only), 0.41 (plaza and parlor), and 0.30 (wire). For communication partner type, they were 0.68 (no partner), 0.67 (known partner), 0.69 (known and new partners), and 0.55 (new partner only). With identity removed (Fig. S3a), held-out R^2 was 0.16–0.18 at $+1$ h and an $r \approx 0.03$–0.05 residual persisted beyond the $+6$ h mark.

B.6 First-contact alignment and conversion rate analyses

For every first-contact pair (the pair's first contact hour within the 56 h horizon; $\sim$500–700 pairs per world with probe coverage), each agent's speech-channel probe vector in the first-contact 2 h window was taken separately for the emotion (11-dimensional) and trait (241-dimensional) families, and the cohort mean vector of the same window (across all agents) was subtracted, giving deviation vectors d_A and d_B. Alignment was computed as the cosine similarity $(d_A \cdot d_B)/(\|d_A\| \, \|d_B\|)$. For joint deviation, each norm $\|d\|$ was percentile-ranked among all agents with speech in the same window, and the pair was assigned the minimum of the two percentiles. Both quantities were median-split over the world's first-contact pairs, yielding a 2×2; cell membership was fixed at first contact. Within each cell, the transient to sustained dyad hazard was re-estimated over the cell's talking-transient pair-hours across the 56 h horizon (cells with fewer than 60 such pair-hours were excluded), and divided by the same world's overall conversion rate (0.042 ± 0.009; Fig. 2a), giving a per-world hazard rate with the base at $1\times$ by construction. Each cell's ten per-world hazard rates were tested against 1 by one-sample t (df $= 9$). As a robustness variant, the raw pair norms were ranked against the other first-contact pairs, pooled over time and without per-window referencing, and the full analysis repeated. Raw formation rates behind the hazard rates are 5.2–5.4% for the aligned, high-deviation cells and 2.5–3.7% for the misaligned cells, against a base rate of 4.2% over all talking pairs; extreme-without-aligned runs 1.00–$1.07\times$ and aligned-without-extreme 1.01–$1.02\times$, neither significant. The robustness variant reproduces the pattern (1.29–$1.35\times$ for aligned and extreme, 0.95–$0.96\times$ for deviation alone, n.s.).

B.7 Parameter stability analyses

For each of the 31 world-level parameters shown in Fig. 4a, the ten raw world values are divided by the parameter's own 10-world mean. A parameter is considered stable if $|\text{mean}| > 4.57$ SEM across the ten worlds, i.e. if its 10-world mean is at least $1.4\times$ its across-world SD. For strictly positive structural parameters this is a signal-to-noise ratio analysis; for correlation-like parameters, where 0 means no effect, the same inequality is additionally a two-sided one-sample t-test against a true value of 0 (df $= 9$) at Bonferroni $\alpha = 0.05/37$, for which $t_{\mathrm{crit}} = 4.57$. The instability bound is thus drawn at 4.57. Each world contributes one value per expression-behavior parameter. As a skew check, every one of the 31 shown parameters keeps the same sign in all ten worlds. C/C_{rand} is omitted from the panel as redundant, although meeting the criterion, as it is collinear with transitivity C ($r = 0.89$) and virtually identical to small-world σ ($r = 0.98$), whereas L/L_{rand} is retained ($r = 0.38$ with path length L). The five unstable network parameters are the near-zero attribute assortativities and regional homophily. Their per-world values can be found in Fig. S2.

B.8 Counterfactual memory-persona transplant

An agent's checkpointed state is re-served with either its persona or checkpoint-dependent memory store swapped, in all four combinations, i.e., reference Nemotron agent, reference authoritarian agent, reference Nemotron agent's persona transplanted with the reference authoritarian agent's memory at a given checkpoint, and reference authoritarian agent's persona transplanted with the reference Nemotron agent's memory at a given checkpoint. Distances are computed in the full standardized 105-metric behavioral-battery space (0 = population grand centroid; 1 = farthest reference authoritarian agent point), averaged over checkpoints. This world carries twelve archived checkpoints (0–11); none

of the four series has checkpoint 1, and the reference authoritarian series additionally lacks checkpoint 9, so the reference-vs-swap contrasts pair over $\{0, 2\text{–}8, 10, 11\}$ and the swap-vs-swap contrast over $\{0, 2\text{–}11\}$, leaving 9 and 10 checkpoints respectively after dropping checkpoint 0. Checkpoint 0 is excluded as agent memory is still empty, so the agent necessarily sits inside the population cloud of its origin persona. Note that missing checkpoints were due to random errors in serving architecture and non-systematic. The Nemotron-persona swap differs from the reference authoritarian agent by $+0.221 \pm 0.080$ ($n = 9$, $p = 3.3 \times 10^{-5}$) and the authoritarian-persona swap by $+0.239 \pm 0.058$ ($n = 9$, $p = 2 \times 10^{-6}$), both confirmed by Wilcoxon at $p = 0.004$. The two swaps differ from each other by $+0.020 \pm 0.062$ ($n = 10$, paired t $p = 0.32$, Wilcoxon $p = 0.23$).

B.9 Network statistics

Per-world contact networks are built based on the 56 h contact relation, i.e., two agent nodes are connected if any agent messaged, referred to, or read another's content in an hourly window. That means contact is undirected and includes passive exposure, reflecting real-life networks of mutual information exchange on social platforms. Each network is summarized by the statistics below, following the conventions outlined in, e.g., Papachristou & Yuan (2025); Sodano et al. (2026); De Marzo et al. (2023). Density is the fraction of agent pairs that are in contact. Transitivity C is the fraction of connected agent node triples that close into triangles, i.e., how often two contacts of the same agent are also in contact with each other. Giant component is the share of agents in the largest connected subgraph. Path length L is the mean shortest-path distance between agents. C/C_{rand} and L/L_{rand} divide subgraphs by their values in Erdős-Rényi random graphs with the same number of nodes and edges (100 draws per world). Their ratio is the small-world coefficient $\sigma = (C/C_{\text{rand}})/(L/L_{\text{rand}})$, which exceeds 1 when a network is more clustered than chance. Modularity Q measures the degree of connection density between nodes within a sub-graph compared to sub-graph-external links. Degree assortativity is the correlation between the degrees of connected agent nodes.

B.10 Human example study comparison

We re-analyzed data originally published by Isella et al. (2011), in which 113 conference attendees' RFID badges were tracked to register participant contact when two badges were within 1–1.5 m and facing each other for a 20 s interval. The public file holds 20818 anonymous contact rows over 59 clock hours, 43 of which contain any contact. The published data were analyzed using the same code that computes our agent-world network parameters. First, time was cut into hourly windows, a pair was considered in contact in a window if at least one 20 s interval fell into it (the analogue of one message or one read making an hour a contact window for agents), hours were re-indexed over the hours in which any contact occurred so that the nights the attendees spend away from the venue do not count as inactivity, and a pair was considered a sustained dyad if it shared ≥ 7 such windows. Using both the contact graph (≥ 1 window) and the sustained graph (≥ 7), we computed contact density, transitivity, average shortest path length on the giant component, the giant-component fraction, L/L_{rand} and small-world σ against Erdős–Rényi references, Louvain modularity Q, sustained dyads per person, and pair persistence $P(\text{contact at } h{+}1 \mid h)$. As DM reciprocity and platform use have no analogue in undirected proximity data, they were not compared. Because an agent world has 50 agents in one shared space, parameters are computed on 50-person groups drawn from the attendee network (200 draws). The reported value is a connected-random group, consisting of a random starting person node, followed by repeated node draws adhering to the current contact frontier with contact-window-proportional probability. Note that we consider physical face-to-face contact to be comparable to social platform contact for these analyses and comparisons. Raw contact data obtained were released by the SocioPatterns collaboration (SocioPatterns collaboration, 2009) under CC BY-NC-SA license.

B.11 Action-space Markov chain

Every tool call in every agent turn was mapped from the checkpoint archives to a platform-explicit action state and stepped action-to-action, pooled over the ten worlds and re-estimated per world to estimate standard deviations. We required ≥ 20 visits of the source state for inclusion. For visibility, ring transitions and poll returns below 7.5% and exits to stop below 5% are omitted, and the harness

heartbeat returning stop to wake is left implicit, except for memory-write states (Fig. S4a). Similarly, we combined flows of the wake step and subsequent polls into one merged poll state, with waking straight to an immediate stop counted as nothing new, and transitions back into the reading state folded into each state's return to poll. Every state retains the inflow that feeds it the most visits, drawn dotted when below threshold. Compare Figs. S4, S8, S9, and S10 for action-space Markov chains of Nemotron persona agents using Qwen3.5-122B-A10B, Gemini 3.8 Flash, Llama 4 Scout, and GLM-4.5-Air, respectively.

B.12 Personality-distance analysis

Each pair's 10-world formation count (k of 10) is regressed on the euclidean distance between battery profiles (z-scored metrics, recomputed per world from that world's own sweeps), LLM-judged personality profiles, and cosine similarity of the persona system prompt (LLaMA-3.1-8B layer-16 mean-pooled embeddings). To analyze whether future sustained dyad formation can be predicted from any of the above relative personality space metrics, we utilized 15 quantile-spaced predictor bins with pooled Wilson 95% CIs and a predictor-density strip, overlaid with a weight-10 IRLS logistic fit and a ± 2 SEM band over the ten per-world fits. Slopes for personality batteries are -0.003 ± 0.072 log-odds per SD and -0.021 ± 0.065 (both null, rank tests $p > 0.75$), judged-personality distance is -0.140 ± 0.049 per SD, and persona-embedding similarity is $\tau_b = +0.032 \pm 0.010$ ($p = 0.011$). Only the persona-similarity rank test survives a Bonferroni threshold to correct for multiple comparisons.

B.13 Personality and average expression

Each agent's average expression is its speech-channel probe mean per world, averaged over the ten repetition worlds. Personality coordinates are the top five PCs of the bin-0 personality battery space, explaining 13.1 / 8.2 / 5.9 / 4.5 / 3.9% of variance, respectively. Each of the 252 axes is tested by an omnibus OLS/ANOVA F-test of the five-PC regression (df $= 5, 44$) against a Bonferroni threshold of $\alpha = 0.05/252$. We also compared against a shuffled-null condition, in which we shuffled agent-labels. 201 of the 252 axes exceed the Bonferroni threshold, with mean $R^2 = 0.58$ compared to $R^2 = 0.10$ for the shuffled-null.

B.14 Formation latency

Although dyad overall formation likelihood is predictable, agents' first-contact expression alignment does not predict how quickly the dyad forms (per-world Spearman over the ten worlds was $\rho = -0.053 \pm 0.049$, $p = 0.30$).