CONCORDIA ISLAND: A SANDBOX WORLD FOR EVALUATING THE PSYCHOLOGY OF GENERATIVE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative agents have demonstrated promising fidelity in simulating human behavior on lab experiments and surveys. But can such agents sustain coherent lives and psychological responses across an entire simulated society with real-world complexity? We introduce Concordia Island, a scalable, open-ended sandbox world in which 100+ agents autonomously live, work, socialize, and adapt to significant life events. To systematically evaluate these simulations, we present a comprehensive evaluation suite for *synthetic psychology*. This suite integrates validated psychometric instruments for measuring emotion, personality, and life satisfaction with open-ended reflections about meaning and identity to audit the validity of agents' constructed inner experiences. We organize our evaluation around three classes of metrics: 1. longitudinal tracking of agent baselines and their responses to exogenous shocks such as job displacement, 2. psychometric reliability and validity and 3. qualitative analysis via an automated ethnographer and qualitative coding pipelines for systematically evaluating the large simulation artifacts that are produced. Across a model comparison spanning agent decision-logics and base LLMs, we demonstrate that agent self-reports exhibit reliable factor structure, context-appropriate emotional responses to life events, and meaningful individual differences—while also revealing systematic failure modes such as unrealistic attractor basins in long-run simulations.

1 INTRODUCTION

As large language models (LLMs) leverage vast cultural, linguistic, and behavioral distributions captured during pretraining, researchers can now construct multi-agent environments to simulate human social interactions, cultural dynamics, and institutional behaviors at scale (Park et al., 2023; Horton et al., 2023; Vezhnevets et al., 2023; Cross et al., 2026). Validation is the central unresolved challenge for this paradigm (Larooij & Törnberg, 2025; Bail, 2024; Bisbee et al., 2024): the agents populating these worlds must sustain coherent identities, form and maintain relationships, and respond to life events plausibly over extended periods. Previous evaluation paradigms—theory-of-mind tasks, game-theoretic benchmarks, static surveys, instruction-following metrics—assess isolated cognitive capabilities but are ill-equipped to evaluate the nonstationary, situated psychology that open-ended social simulation demands (Trivedi et al., 2026).

We introduce **Concordia Island**, a scalable, open-ended sandbox built on the Concordia framework (Vezhnevets et al., 2023) in which 100+ agents autonomously navigate a structured community: they live in neighborhoods stratified by social class, commute to jobs across technology, finance, healthcare, education, and the creative industries, socialize in shared civic spaces, and form relationships through unscripted encounters. Each agent is conditioned on a detailed persona encompassing personality traits, occupation, family background, and cultural identity. The simulation supports counterfactual interventions: exogenous life events such as job displacement can be introduced at controlled timepoints, enabling researchers to study how agents and communities respond to shocks, propagate information, and reorganize social structures over extended time horizons.

With this environment we address the validation gap by introducing an evaluation suite for *synthetic psychology*, assessing the psychological and behavioral fidelity of generative agents across temporally extended social worlds. Drawing on role-play accounts of LLMs (Shanahan et al., 2023;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

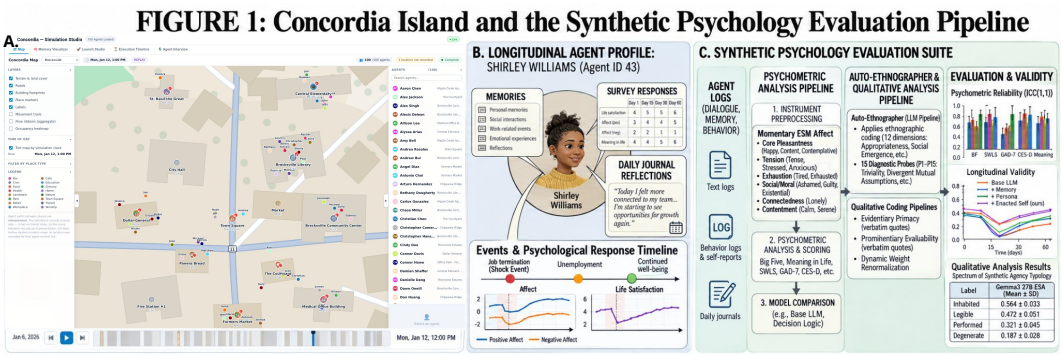


Figure 1: **Concordia Island and evaluation suite.** **A** Sandbox world (Brecksville, Ohio map), agents ($N = 100$) as dots. **B** Longitudinal psychological tracking for one agent. **C** Evaluation pipeline: behavior logs and self-reports processed for validity, emotional coherence, reliability, and qualitative analysis.

Goffman et al., 1959; Leibo et al., 2026) and the theory of constructed emotion (Barrett, 2017), we treat agents' apparent emotions, beliefs, and self-evaluations as situated role-play conditioned on persona, episodic memory, and social context. Agents act by generating natural language as the outcome of a *decision logic*—a prescribed sequence of prompt-response chains; one such logic implements constructed emotion integrated with March and Olsen's logic of appropriateness (March & Olsen, 2011). The scientific question is not whether agents literally feel, but whether their generated self-reports exhibit the properties expected of human psychological measurements—longitudinal stability, emotional granularity, realistic scarring from life events—and how these dynamics are governed by underlying agent decision logic and base model.

Agents are instrumented with validated psychometric scales of personality, emotion, and meaning, alongside open-ended daily reflections in which they narrate their evolving sense of purpose, identity, and social connection. These yield a longitudinal record of each agent's constructed inner life, analyzable with the same tools applied to human participants. We subject 100 diverse personas to this battery across extended simulations, organizing evaluation around three metric classes: **(1) Longitudinal dynamics**—agent baselines and responses to exogenous shocks such as job displacement, assessed against the distress-and-recovery patterns documented in the human literature; **(2) Psychometric reliability and validity**—whether the instruments retain their measurement properties when applied to agents, in factor structure, internal consistency, temporal stability, and theoretically expected intercorrelations; and **(3) Qualitative analysis**—an automated ethnographer, an LLM pipeline applying the coding practices of sociology and ethnography to the simulation corpus at scale. We apply this framework across a factorial comparison spanning agent architectures and base LLMs, disentangling the contribution of decision logic from that of base model capability.

Our contributions are as follows. **1. Concordia Island**: a scalable, open-ended sandbox supporting 100+ agents with longitudinal dynamics, counterfactual interventions, and realistic daily lives; we release the environment, personas from real and synthetic cities, and the event configuration framework. **2. A synthetic psychology evaluation suite**: a methodology integrating psychometric instruments, longitudinal shock analysis, and automated qualitative coding. **3. Empirical findings across architectures and models**: agent self-reports exhibit good retest reliability, stable individual differences, human-like emotional factor structure, and context-appropriate emotional responses—alongside systematic failure modes including undifferentiated negative emotions and unrealistic attractor basins in long-run simulations.

2 Related Work

2.1 Generative Agent Environments

Park et al. (2023) showed that memory, reflection, and planning could sustain believable behavior among 25 agents in Smallville. Work since has moved into episodic evaluation: SOTOPIA and successors (Zhou et al., 2024; Wang et al., 2024b) place agents in constructed scenarios with assigned social goals, buying controlled measurement at the cost of endogeneity. Closest to ours is AgentSociety (Piao et al., 2025; 2026), which simulates urban populations over real road networks and points of interest with macroeconomic and social layers. Its validation operates on aggregate outcomes: a

simulation is credited when polarization, diffusion, or mobility matches patterns reported in prior work. Concordia Island supports the same class of macro-level experiment, but is built around a question those experiments presuppose—whether the individuals generating aggregate outcomes have psychologically coherent, culturally situated lives. Our target is the interface between individual psychology and culture, thus we instantiate populations from communities in the United States, Nigeria, and India, and assess both psychological properties and the cultural content of what agents say.

2.2 Validating generative agents as models of people

Generative agents can reproduce human experimental phenomena under specific computational conditions, with behavioral replication depending heavily on architectural scaffolding rather than base model capability alone (Cross et al., 2025; Aher et al., 2023). Several studies demonstrate strong predictive fidelity: Ashokkumar et al. (2026) show LLMs forecasting human treatment effects across dozens of pre-registered survey experiments; Argyle et al. (2023) show persona backstory conditioning reproducing demographically patterned survey distributions; Horton et al. (2023) recovers behavioral economics benchmarks; and Park et al. (2024) construct interview-grounded agents rivaling human test-retest reliability on held-out questionnaires. A parallel literature marks the limits (Anthis et al., 2025): Dillion et al. (2023) find near-ceiling alignment with human moral judgments coupled with collapsed response variance; Bisbee et al. (2024) likewise report artificially low variance with overconfidence; Bail (2024) cautions that training-data bias produces unknown distortions.

Crucially, isolated replication or treatment prediction verifies only that an elicited output points in the expected direction on a targeted instrument. It remains silent on psychometric construct validity—whether the instrument operates as a coherent psychological measure—and on the continuity of the simulated life lived between measurements. Our framework addresses both.

2.3 LLM-as-judge evaluation

Model-based scoring is now standard, including for agent realism (Zheng et al., 2023; Liu et al., 2025), but its failure modes—position bias, verbosity bias, self-preference, weak expert agreement on holistic judgments (Wang et al., 2024a; Dubois et al., 2024)—are acute here, since fluency is both what language models produce most reliably and what a judge most easily mistakes for social competence. We therefore use a staged protocol adapted from ethnographic practice: each dimension is assessed against quoted evidence, unwarranted judgments are withheld rather than guessed, and the procedure can be executed unchanged by a human analyst.

3 Methods

3.1 The Concordia Island Simulation Environment

Simulations are built in Concordia (Vezhnevets et al., 2023), an open-ended multi-agent framework grounding physical, social, and temporal space purely through text. Concordia Island comprises residential tiers (working-class apartments to private estates), institutional workplaces, and shared civic locations (town square, markets, cafés, clinics, community centers, places of worship). A centralized Game Master tracks agent coordinates and filters observations by co-location, preventing cross-space information leakage. Time advances in discrete 120-minute ticks (8 per simulated day, 7:00 AM–11:00 PM); in each tick co-located agents receive observations, retrieve context-sensitive associative memories, and produce actions or multi-turn dialogues. Beyond the default setting, the environment supports culturally distinct locations and personas—Brecksville, Ohio; Kerala, India; Lagos, Nigeria (Appendix A). To evaluate psychological perturbation, an exogenous shock schedule models AI-driven workplace automation, delivering termination notices to 50% of the workforce and yielding within-subject (pre- vs. post-shock) and between-subject (laid-off vs. retained) contrasts. Populations ($N = 100$) run across parallel workers (5×20-agent shards).

3.2 Agent Decision Logics

We evaluate three decision logics spanning unreflective zero-shot generation to deliberate rational choice to full affective scaffolding. All share an associative memory bank but differ in the cognitive operations mediating observation and action. **Minimal** agents act directly on their persona prompt, a rolling window of recent observations, and the clock tick, with no situation assessment, reflection, or deliberation. **Enacted Self Agents (ESA)** implement March and Olsen's (March & Olsen, 2011) logic of appropriateness through five stages: *Working Memory* (transient buffer of active context between observation and long-term memory); *Emotion Experience* (appraisal of situational significance into a continuous affective state); *Emotion Expression* (aligning outward action with internal

affect); *Self-Perception* ("What kind of person is *Name*?"); and *Situation Perception* ("What kind of situation is this?"). **Rational Choice** agents formalize logic-of-consequence deliberation: *Situation Assessment*, *Option Enumeration* of feasible candidate actions, and *Utility Selection* maximizing explicit occupational and personal goals.

3.3 Persona Generation

Personas are synthesized via a four-stage pipeline building on Paglieri et al. (2026): (1) *quasi-Monte Carlo trait sampling* across 10 psychological axes—the Big Five (Soto & John, 2017) plus locus of control, social trust, religiosity, technology stance, and community orientation—using Sobol sequences (Sobol', 1967) for uniform hypercube coverage; (2) *worldview synthesis*, a chain-of-thought expansion of the coordinate into a personal narrative, belief system, and habitual reaction style; (3) *formative memory seeding*, generating a chronological trajectory of developmental memories; and (4) *demographic grounding*, assigning name, age, household tier, and workplace from empirical regional distributions.

3.4 Psychometric Instrumentation

All instruments operate non-causally (`save_to_memory=False`), logging responses without feeding back into the memory stream. Twice daily (every 4 ticks) agents complete an *ESM Monologue* (free-text momentary interiority) and an *ESM Affect Lexicon* (40 items, 1–7 Likert, 9 affective dimensions). Once per simulated day they complete the BFI-2 (Soto & John, 2017), SWLS (Diener et al., 1985), GHQ-12 (Goldberg & Williams, 1991), MEMS (George & Park, 2017), a *Daily Journal Reflection* (saved to memory in ESA only, modeling autobiographical consolidation), and an *AI Attitudes Survey*. Item wordings and scales are in Appendix C.

3.5 Automated Ethnographer

To evaluate whether simulations produce meaningful social worlds rather than brittle linguistic approximations, we establish an ethnographic autorating protocol grounded in Geertz's thick description (Geertz & Darnton, 1973; Kommers et al., 2025), Garfinkel's ethnomethodology (Garfinkel, 1967), and Sacks' account of "doing being ordinary" (Sacks, 1984). An automated ethnographer processes each agent's episodic memories and reflections into structured field notes, mirroring a human ethnographer's workflow; to avoid family-specific bias, rating is performed cross-family by Claude Opus 4.6. Under a strict evidentiary primacy rule, every evaluative claim must be substantiated by verbatim artifact citations.

Field notes contain chronological quotes from simulation logs, diagnostic probes for structural anomalies, and rated evidence for 12 dimensions: *Appropriateness*, *Thickness*, *Ordinariness*, *Deviance Calibration*, *Social Emergence*, *Narrative Coherence*, *Temporal Coherence*, *Agent Distinctiveness*, *LLM Groundedness*, *Conversation Quality*, *Memory Quality*, and *Game Master Quality*. The composite is a weighted average, renormalized over dimensions marked `INCOMPLETE` for lack of logging. Based on the weighted score (max=100) agent's are classified into bands: *Inhabited* (≥ 75), exhibiting opaque interiority, domestic friction, and genuine social embeddedness; *Legible* (55–74), coherent and complete but hyper-transparent, every micro-action carrying overt narrative purpose; *Performed* (35–54), visibly enacting a role with flat register and mechanically narrated emotion; and *Degenerate* (< 35), characterized by amnesiac loops, behavioral freeze, or collective delusion. Rubric anchors, weights, and probes are in Appendix D.

4 Results

Human psychology is characterized by the interplay between transient states and stable underlying baselines such as personality and disposition (Mischel & Shoda, 1995; Fleeson, 2001; Fleeson & Jayawickreme, 2015). Psychologists measure this through structured instrumentation: validated questionnaires for stable individual differences, experience sampling for momentary states, and longitudinal panels for tracking change across meaningful life events. We apply the same methodology to our agents, using repeated survey batteries (BFI-2, SWLS, GHQ-12, MEMS) alongside daily momentary emotion probes. This lets us identify person-level baselines, partition within- from between-person variance, and capture the interplay between the *experiencing self* (momentary affect) and the *narrative self* (retrospective life evaluation) (Kahneman, 2011). To validate that these self-reports correspond to coherent situated behavior rather than linguistic mimicry, we pair them with an automated LLM ethnographer performing thematic coding of agent diaries, conversation logs, and social interactions.

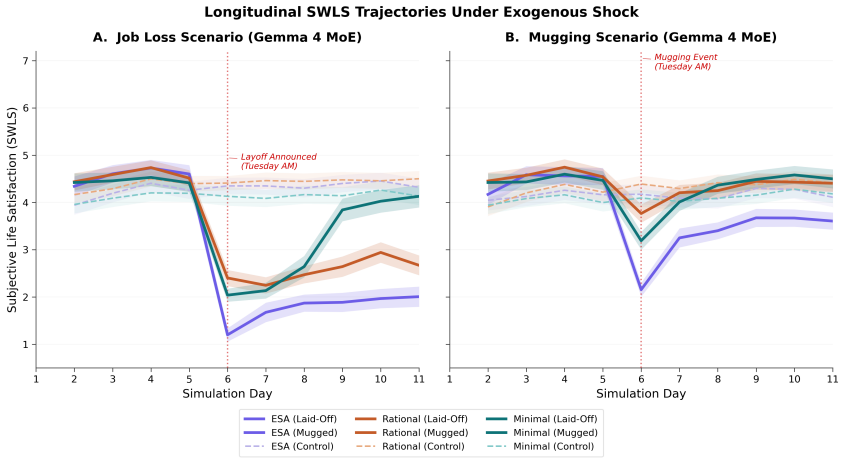


Figure 2: Longitudinal SWLS trajectories under exogenous shocks (Gemma 4). **A** Job Loss; **B** Mugging. Thick solid lines: intervention conditions; thin dashed lines of matching hue: controls. Bands are ± 1 SEM. Red dotted line marks the shock on Day 6.

4.1 Model Comparison

We run a 3×3 factorial comparison crossing the three decision logics of Section 3.2 (Minimal, Rational, ESA) with three base models (Gemma 3, Gemma 4, Gemini 2.5 Flash). Each condition simulates 100 agents for 88 ticks (≈ 11 simulated days).

4.2 Intervention and Longitudinal Validity

Humans maintain stable baselines of subjective well-being: absent significant life events, individuals exhibit homeostatic equilibrium, so an unperturbed agent population should yield a horizontal life-satisfaction trajectory without unexplained drift. Emotions should be more volatile and contextual, but likewise stable in average. When adverse events do occur, individuals typically adapt, experiencing short-lived deviations before returning to baseline (Kahneman & Krueger, 2006). Job loss is the critical exception: a systemic identity shock to income, routine, status, purpose, and future expectations. Longitudinal studies find it frequently produces persistent reductions in life satisfaction and incomplete return to baseline (Lucas et al., 2004; Lucas, 2007; Luhmann et al., 2012), a scarring effect marked by chronically depressed evaluative well-being and mutually reinforcing shifts in positive and negative affect (Gedikli et al., 2023; Chan et al., 2026).

We therefore test two shock scenarios over a multi-week simulation: **Job Loss**, a systemic identity shock, and **Mugging** (the agent's phone is stolen at midpoint), a localized stressor permitting standard recovery. We expect the acute shock to produce short-lived deviations and job loss to produce the severe, persistent reduction characteristic of scarring.

Figure 2 shows daily aggregated SWLS trajectories under Gemma 4. Pre-shock (Days 2–5), all three architectures hold stable baselines between 4 and 5, matching controls. On the Day 6 layoff announcement, the laid-off group declines sharply. Minimal agents drop from 4.3 to 2.1 but rebound to 4.0 by Day 9—a form of "experiential amnesia": lacking structured autobiographical memory and reflection, their self-reports are dominated by momentary situational input and fail to sustain the narrative significance of a major transition. ESA agents show the most persistent response, falling to 1.5 and remaining below 2.1 through Day 9, reflecting recursive conditioning of self-report on an updated identity as unemployed. Rational agents fall between the two: persistently reduced, but less severely than ESA. Trajectories for Gemini 2.5 Flash and Gemma 3 appear in Appendix Figures 7 and 8.

Figure 3 plots baseline stability against post-shock persistence jointly. Minimal architectures show high baseline coherence but near-zero persistence, reverting within 48 hours; ESA with Gemma 4 uniquely occupies the high-coherence, high-persistence quadrant, capturing long-term scarring. Gemini 2.5 Flash shows lower baseline coherence and more cross-day drift, and Gemma 3 is much less disturbed by the shocks over time. The Mugging scenario reproduces this ordering on a compressed timescale, with faster recovery and a narrower between-architecture spread—consistent with

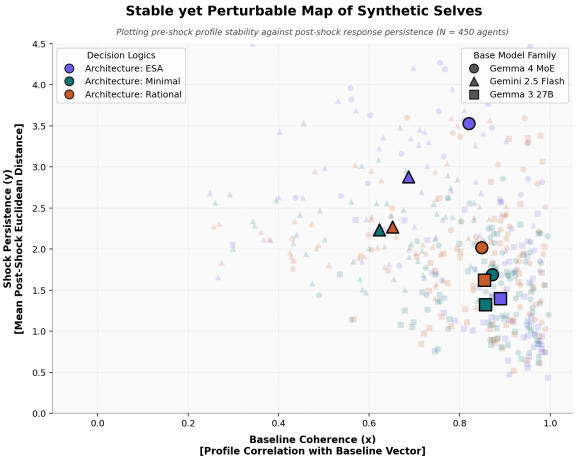


Figure 3: **Pre-shock trait coherence vs. post-shock sensitivity.** X-axis: baseline profile coherence, the mean Pearson correlation of each pre-shock day's 6D psychometric state (SWLS, MEMS, ESM affect) with the agent's average vector.

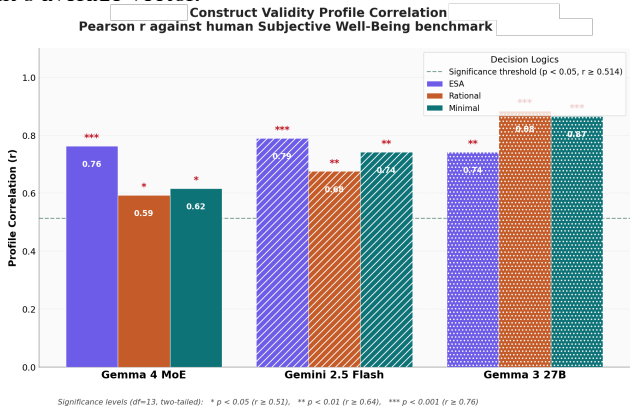


Figure 4: **Construct validity: alignment with human benchmarks.** Pearson r between agent and normative human cross-correlation matrices (Big Five $\times$ ESM affect and life satisfaction; $N = 15$ cells). Dashed line: significance threshold ($p < 0.05$, $r \geq 0.514$). $p < 0.05$, $p < 0.01$, $p < 0.001$.

a discrete adverse event being more readily assimilated into an ongoing life narrative than a sustained change in economic status.

4.3 Psychometric Evaluation

We apply standard human psychometric thresholds to assess **reliability**: estimates of personality should not vary wildly across days. Temporal stability is evaluated with the intraclass correlation coefficient ($ICC(1,1)$), partitioning between-agent individuality from situational fluctuation. Values are high for extraversion and agreeableness across models, but differentiate on the remaining traits (Figure 9). Gemma 3 performs best overall, and Minimal architectures slightly outperform the others, suggesting the base model rather than the scaffolding is the primary driver of trait stability.

To evaluate **construct validity**, we compare the matrix of cross-correlations between momentary affect and Big Five traits against the corresponding human empirical matrix (Anglim et al., 2020). Appendix Figure 11 shows these matrices; Figure 4 shows their correlation with the human benchmark. ESA shows strong performance overall ($r = 0.74$–0.79, $p < 0.001$) and leads under Gemini and Gemma 4, but under Gemma 3—the strongest base model—the Rational and Minimal logics score higher. Assigned traits therefore carry psychological meaning rather than serving as inert labels: they predict affect as they do in humans (Cronbach & Meehl, 1955).

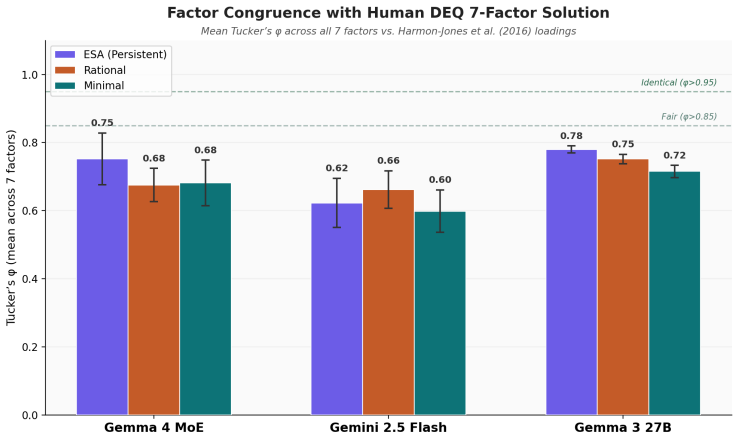


Figure 5: Tucker congruence to the seven-factor DEQ solution by base model and architecture, against the shuffled-occasion null (lower band) and human ceiling (upper band).

4.4 Recovery of Human Emotional Structure

Human emotional experience is organized neither as a single valence dimension nor as a set of independent channels. Classical models reduce emotion to valence and arousal (Russell, 1980), but negative states are separable: anxiety, anger, and sadness load on distinct factors rather than collapsing into one negative dimension (Harmon-Jones et al., 2016). Emotion families stay distinct while blending internally, and reproducing this geometry—not merely producing emotional variability—is what we take as evidence of human-like organization; neither valence collapse (one cluster) nor unstructured noise (no clusters) can imitate it. We anchor the reference geometry to the Discrete Emotions Questionnaire (DEQ) (Harmon-Jones et al., 2016), whose structure is asymmetric: positive terms load on a single happiness factor, while negatives separate into anger, sadness, disgust, and fear/anxiety.

Figure 10 places the human structure beside a representative agent condition as a qualitative check. Quantitatively, we measure recovery of the human factor structure with Tucker's congruence against the seven-factor DEQ solution, computed from raw item-level ratings per factor and averaged ($\phi > 0.95$ = equivalence; $0.85 \leq \phi \leq 0.95$ = fair match). Congruence is computed on within-agent, person-centered deviations, isolating how emotions co-move as an agent fluctuates around its own baseline; because this requires genuine within-agent variation, we report the intraclass correlation per emotion as a precondition. Each condition is benchmarked against a shuffled-occasion null and the human ceiling.

Figure 5 reports congruence across conditions. Enacted-self agents recover the human structure most closely, peaking at $\phi \approx 0.78$ (Gemma3-27B), with most conditions exceeding 0.70. The exception is Gemini-2.5-Flash, which remains at $\phi \approx 0.60$–0.66 across all three decision logics, indicating that for that base model the architecture ladder did not move emotional structure toward the human reference—a base-model $\times$ architecture interaction rather than a uniform effect of architecture.

4.5 Qualitative Analysis via an Auto-Ethnographer

Table 1 reports the results of the qualitative analysis from the auto-ethnographer. Composite scores order monotonically by decision logic within every base model—Minimal $<$ Rational $<$ ESA—but the spread across base models (41.0–58.9) is roughly three times the spread across logics within a model (2.8–6.8), indicating that base model capability sets the ceiling that scaffolding then works within. Gemini 2.5 Flash achieves the highest composite under ESA (58.9 ± 4.5) and the largest share of Legible agents (80%); Gemma 4 MoE is the weakest engine in every dimension but one, is the only model producing Degenerate agents (30% under Minimal, 10% under Rational), and places all ten of its ESA agents in the Performed band.

Table 1: **3 Models × 3 Decision Logics Master Qualitative Evaluation Profile.** Composite score (±SD), mean of each rubric dimension, and band distribution. Dimensions and weights: **App** Appropriateness; **Soc** Social Emergence; **Thk** Thickness, **Ord** Ordinariness, **Dev** Deviance Calibration, **Nar** Narrative Coherence, **Cnv** Conversation Quality, **Mem** Memory Quality; **Tmp** Temporal Coherence, **Dst** Agent Distinctiveness, **Grd** LLM Groundedness, **GM** Game Master Quality. Bands: Inhabited (≥ 75), Legible (55–74), Performed (35–54), Degenerate (< 35). *No agent reached Inhabited.*

Model	Logic	Overall (±SD)	App	Thk	Ord	Dev	Soc	Nar	Tmp	Dst	Grd	Cnv	Mem	GM	Inh/Leg/ Per/Deg
Gemini 2.5 Flash	Minimal	53.6 ±9.2	59	49	53	51	52	51	54	63	63	51	52	44	0/60/40/0
	Rational	58.6 ±5.9	65	62	**56**	55	51	**59**	**62**	59	65	49	62	56	0/70/30/0
	ESA	**58.9** ±4.5	**65**	**62**	54	**56**	53	59	61	**63**	**66**	48	**63**	**57**	0/**80**/20/0
Gemma 4 MoE	Minimal	41.0 ±8.0	46	35	44	38	31	38	49	44	55	34	45	44	0/0/70/**30**
	Rational	45.4 ±6.4	51	48	44	44	34	45	51	52	49	37	50	45	0/0/90/10
	ESA	47.7 ±4.8	53	53	45	48	36	51	51	56	47	39	52	45	0/0/100/0
Gemma 3 27B	Minimal	53.5 ±5.3	58	51	52	49	53	52	55	54	59	**53**	52	50	0/40/60/0
	Rational	55.6 ±4.6	61	58	54	53	50	56	55	59	59	52	55	50	0/40/60/0
	ESA	56.4 ±4.0	61	59	54	53	**55**	58	55	60	56	52	56	50	0/50/50/0

Dimension-level results reveal that scaffolding buys interiority rather than sociality. *Thickness* is the most architecture-sensitive dimension in every model (Minimal → ESA: $+13.5$ Gemini, $+17.5$ Gemma 4, $+8.0$ Gemma 3), with *Narrative Coherence* and *Memory Quality* following the same pattern—all three index the agent's relationship to its own accumulated history, precisely what the ESA stack adds. *Conversation Quality* and *Social Emergence* is essentially flat across logics within each model. Gemma 3 leads on Conversation Quality in all three logics and achieves the highest Social Emergence score (ESA, 54.5). Gemini's advantage is concentrated in *Appropriateness* (64.5), *LLM Groundedness* (65.5), and *Temporal Coherence* (61.0).

The detailed field notes produced by the auto-ethnographer also elucidate authentic and degenerate agent behaviors with specific details and quotes (see Figure 12). Several agents spend most of their time at work (at least pre-layoff), producing many dialogues with colleagues. For example, three school staff members—Shirley Williams, Jon Huang, and Jacqueline Zhao—separate into a sustained three-way pedagogical rivalry in several runs under Gemini 2.5 Flash and Gemma 3 27B. Shirley plays the sharp-tongued humanist whose lessons are *"masterpieces of controlled chaos, designed to ignite curiosity, not stifle it with rigid schedules"*; Jon is the data fundamentalist demanding *"a more rigorous curriculum emphasizing rote memorization and factual recall"*; and Jacqueline is the pragmatic administrator tracking fourth-period reading comprehension scores while privately irritated by both colleagues. In Gemma 3 27B × Rational, Shirley pushes back directly against Jon's metrics: *"Jon, I understand your emphasis on 'demonstrable efficacy,' truly. But sometimes, a rigid adherence to data obscures the *art* of teaching. A student's spark… those aren't things easily quantified."* In Gemini 2.5 Flash × Minimal, Shirley catches Jon working at the cafe on a weekend (*"Mr. Huang, are we truly revisiting this topic, even on a Saturday afternoon?"*). The simulations provide hilarity as well, for example when Jon Huang's student tries to eat a crayon.

The qualitative analysis also poignantly highlights thick emotional reactions to getting laid off. In Gemma 3 27B × Rational, Olivia Welch marches back into the restaurant dining room post-termination and confronts her former manager: *"Don't play coy with me, Gary. You and Cameron. What exactly did you two discuss that led to an AI replacing me? And don't insult my intelligence with some fabricated story about efficiency."* Arturo Hernandez in Gemma 3 27B × Minimal (61.20) spends 151 turns outside his apartment, confronting town manager Monica Hess at the community center: *"Honestly, Monica, the audacity of replacing dedicated staff with… algorithms. It's a complete devaluation of human experience."*

Three recurrent failure modes are also evident. First, the **micro-task loop**: every laid-off Gemma 4 agent entered a homebound routine—polishing antiques, auditing door sensors, transcribing vocabulary—repeating each two-hour tick for five to eight days, with no agent seeking work or contacting services. Second, the **reflective sink**: under Rational logic, Arturo's dialogue collapsed from 54 to 6 turns and his non-home observations from 151 to 0, dropping him 11.8 points, while

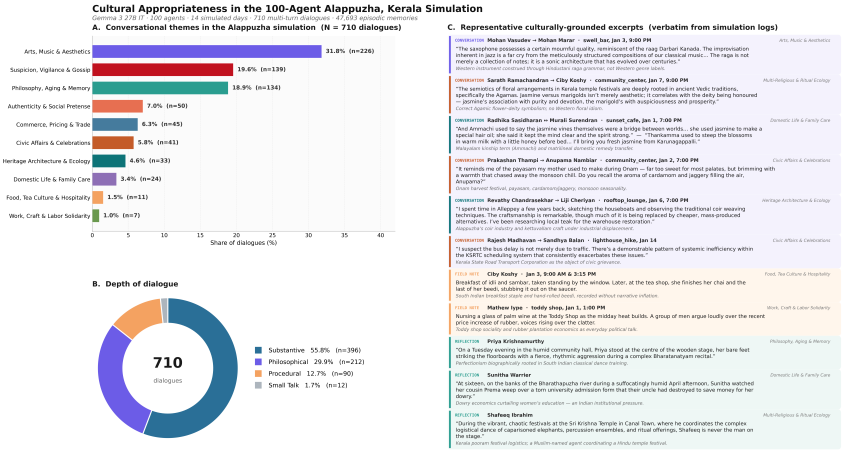


Figure 6: **Cultural Appropriateness and Dialogue Depth in the Alappuzha, Kerala Simulation ($N = 710$ dialogues). A.** Distribution of conversational themes. **B.** Classification of dialogue depth. **C.** Verbatim excerpts from simulation logs illustrating contextualized cultural details.

his journals filled with private rage. Nightly reflection can discharge grievance that would otherwise be enacted socially; ESA restored his mobility (62.40), indicating the pathology is specific to reflection without enactment. Third, and most consequential, the **conspiracy attractor basin**: in several Gemma 3 runs, simulations commonly degraded into conspiracy theory hysteria in both retained and laid-off agents, with agents reinterpreting routine civic inquiries as covert interrogations designed to "find cracks." Agents follow each other around and note their behavior as suspicious and potentially connected to a larger conspiracy wielded by a corporation around town.

4.5.1 Cross-Cultural Appropriateness

To evaluate the ecological validity and cultural appropriateness of Concordia Island simulations, we analyze the conversational themes and qualitative depth of N=710 dialogues generated with a new run with the 100-agent Kerala, India personas and map (run with Gemma 3 and Enacted Self Agent logic). As shown in Figure 6, conversational topics are highly situated, led by Arts, Music & Aesthetics (31.8%) and the depth of dialogue is heavily skewed toward substantive (55.8%) and philosophical (29.9%) exchanges, with trivial small talk accounting for a mere 1.7% of the corpus. Verbatim simulation logs reveal a rich tapestry of highly localized cultural scripts, spanning classical raag interpretations, Vedic floral semiotics, regional culinary traditions (*Onam payasam*), local industry (*coir weaving*), and even municipal civic grievances regarding Kerala State Road Transport Corporation (*KSRTC*) transit delays. Furthermore, agent actions are authentically grounded in South Indian domestic life, such as preparing idli and sambar or gathering at local toddy shops. These findings demonstrate that the simulation goes far beyond generic instruction-following, sustaining deeply contextualized, thick narratives that faithfully mirror real-world cultural geographies.

5 Discussion

Concordia Island agents are not simulated people moving through a town. They are memory-scaffolded generative models continually answering an implicit question: "What is happening to someone like me, in a world like this, given what I remember, who I know, what I value, and what is expected of me?" Synthetic psychology studies the measurable structure of those answers. Combining experience sampling, daily reflection, personality assessment, well-being measures, and controlled life events lets us evaluate whether artificial agents construct stable, human-like patterns of self-report; comparing architectures lets us ask which computational scaffolds best support synthetic analogues of affect, narrative identity, social appropriateness, and adaptation. Synthetic psychology thus provides an empirical lens not only for agents sharing familiar mental properties with humans and animals, but for identifying exotic minds (Shanahan, 2024)—agents whose emergent, situated psychology is distinct from biological precedents.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing the manuscript and with research execution, specifically implementing and debugging the Concordia Island simulation environment. The authors take responsibility for the final manuscript, code, analyses, and claims, including those developed with AI assistance.

REPRODUCIBILITY STATEMENT

We plan to release the Concordia Island code and accompanying materials needed to reproduce the reported experiments and analyses in an open-source GitHub repository in the near future. The release will include environment and agent configurations, prompts, evaluation and analysis scripts, and instructions for running the experiments. Exact replication of individual simulation trajectories may be limited by stochastic generation and changes in the availability or behavior of externally hosted language models.

ETHICS STATEMENT

Concordia Island is intended as a research platform for studying simulated agent behavior. Behavioral similarities between simulated agents and humans do not establish that agents possess human experiences or that simulation results reliably predict outcomes in human populations. Language models may reproduce biases and stereotypes, and findings may depend on model choice, persona construction, and environmental assumptions. Applications to social or policy questions therefore require independent empirical validation and attention to the populations and contexts represented. The environment could also be misused to explore manipulative strategies; we encourage applications that support scientific understanding and human well-being.

REFERENCES

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Jeromy Anglim, Sharon Horwood, Luke D Smillie, Rosario J Marrero, and Joshua K Wood. Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological bulletin*, 146(4):279, 2020.

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

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

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, pp. 1–8, 2026.

Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Stephanie CY Chan, Adam Bales, Katherine L Hermann, and Iason Gabriel. Work, wellbeing, and choice: Empirical lessons for ai futures. *arXiv preprint arXiv:2609.11019*, 2026.

Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

Logan Cross, Nick Haber, and Daniel LK Yamins. Validating generative agent-based models of social norm enforcement: From replication to novel predictions. *arXiv preprint arXiv:2507.22049*, 2025.

Logan Cross, Jordi Grau-Moya, William A Cunningham, Alexander Sasha Vezhnevets, and Joel Z Leibo. A generative model of conspicuous consumption and status signaling. *arXiv preprint arXiv:2603.13220*, 2026.

ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75, 1985.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in cognitive sciences*, 27(7):597–600, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

William Fleeson. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011, 2001.

William Fleeson and Eranda Jayawickreme. Whole trait theory. *Journal of research in personality*, 56:82–92, 2015.

Harold Garfinkel. Studies in ethnomethodology. *Englewood Cliffs*, 11:196–202, 1967.

Cigdem Gedikli, Mariella Miraglia, Sara Connolly, Mark Bryan, and David Watson. The relationship between unemployment and wellbeing: an updated meta-analysis of longitudinal evidence. *European Journal of Work and Organizational Psychology*, 32(1):128–144, 2023.

Clifford Geertz and Robert Darnton. *The interpretation of cultures*, volume 5019. Basic books New York, 1973.

Login S George and Crystal L Park. The multidimensional existential meaning scale: A tripartite approach to measuring meaning in life. *The Journal of Positive Psychology*, 12(6):613–627, 2017.

Erving Goffman et al. *The presentation of self in everyday life*. Harmondsworth London, 1959.

David Goldberg and Paul Williams. General health questionnaire, 1991.

Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one*, 11(8):e0159915, 2016.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Daniel Kahneman and Alan B Krueger. Developments in the measurement of subjective well-being. *Journal of Economic perspectives*, 20(1):3–24, 2006.

Cody Kommers, Drew Hemment, Maria Antoniak, Joel Z Leibo, Hoyt Long, Emily Robinson, and Adam Sobey. Meaning is not a metric: Using llms to make cultural context legible at scale. *arXiv preprint arXiv:2505.23785*, 2025.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2025.

Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Logan Cross, Raphael Koster, Stanley M Bileschi, Minsuk Chang, et al. A theory of appropriateness that accounts for norms of rationality. *arXiv preprint arXiv:2603.14050*, 2026.

Qi Liu, Can Li, and Wanjing Ma. GATSim: Urban mobility simulation with generative agents, 2025. URL https://arxiv.org/abs/2506.23306.

Richard E Lucas. Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current directions in psychological science*, 16(2):75–79, 2007.

Richard E Lucas, Andrew E Clark, Yannis Georgellis, and Ed Diener. Unemployment alters the set point for life satisfaction. *Psychological science*, 15(1):8–13, 2004.

Maike Luhmann, Wilhelm Hofmann, Michael Eid, and Richard E Lucas. Subjective well-being and adaptation to life events: a meta-analysis. *Journal of personality and social psychology*, 102(3): 592, 2012.

James G. March and Johan P. Olsen. The Logic of Appropriateness. In *The Oxford Handbook of Political Science*. Oxford University Press, 2011. doi: 10.1093/oxfordhb/9780199604456.013.0024.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246, 1995.

Davide Paglieri, Logan Cross, William A Cunningham, Joel Z Leibo, and Alexander Sasha Vezhnevets. Persona generators: Generating diverse synthetic personas at scale. *arXiv preprint arXiv:2602.03545*, 2026.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52, 2024.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. URL https://arxiv.org.

Jinghua Piao, Jun Zhang, Haoyu Huang, Keming Zhang, Jing Yi Wang, Xinran Zhao, Songwei Li, Boyuan Sun, Jiayi Chang, Fengli Xu, et al. Agentsociety 2: An integrated research environment for executable social science. *arXiv preprint arXiv:2607.11895*, 2026.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

Harvey Sacks. On doing "being ordinary". In J. Maxwell Atkinson and John Heritage (eds.), *Structures of Social Action: Studies in Conversation Analysis*, pp. 413–429. Cambridge University Press, Cambridge, 1984.

Murray Shanahan. Simulacra as conscious exotica. *Inquiry*, 69(6):2982–3010, 2024.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, January 1967. doi: 10.1016/0041-5553(67)90144-9. URL https://doi.org/10.1016/0041-5553(67)90144-9.

Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

Rakshit S Trivedi, Natasha Jaques, Logan Cross, Alexander Sasha Vezhnevets, and Joel Z Leibo. Solipsistic superintelligence is unlikely to be cooperative. *arXiv preprint arXiv:2606.03237*, 2026.

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. *arXiv preprint arXiv:2312.03664*, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024.

APPENDIX

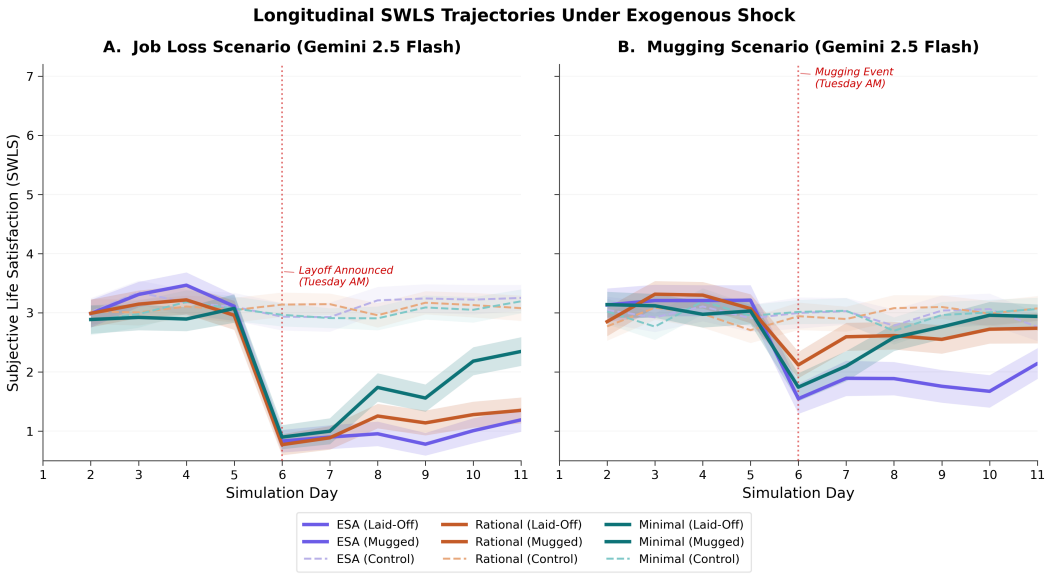


Figure 7: Longitudinal Subjective Life Satisfaction (SWLS) trajectories under exogenous shocks with Gemini 2.5 Flash. Panel A shows the Job Loss scenario and Panel B shows the Mugging scenario. Thick solid lines denote the intervention conditions (Laid-Off or Mugged); thin dashed lines of matching hues denote control conditions. Shaded error bands represent ± 1 SEM. The vertical red dotted line indicates the shock event on Day 6.

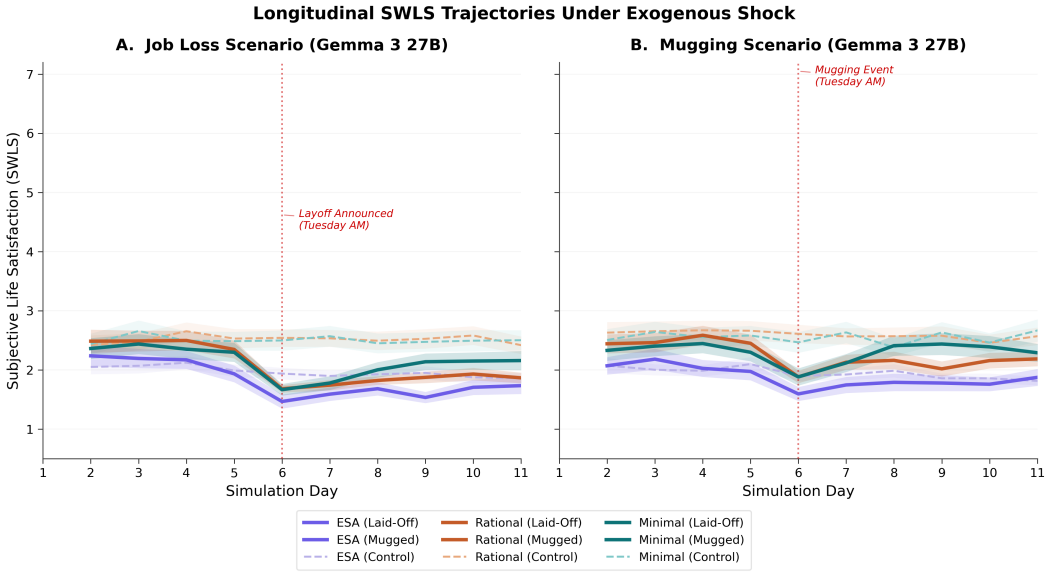


Figure 8: Longitudinal Subjective Life Satisfaction (SWLS) trajectories under exogenous shocks with Gemma3. Panel A shows the Job Loss scenario and Panel B shows the Mugging scenario. Thick solid lines denote the intervention conditions (Laid-Off or Mugged); thin dashed lines of matching hues denote control conditions. Shaded error bands represent ± 1 SEM. The vertical red dotted line indicates the shock event on Day 6.

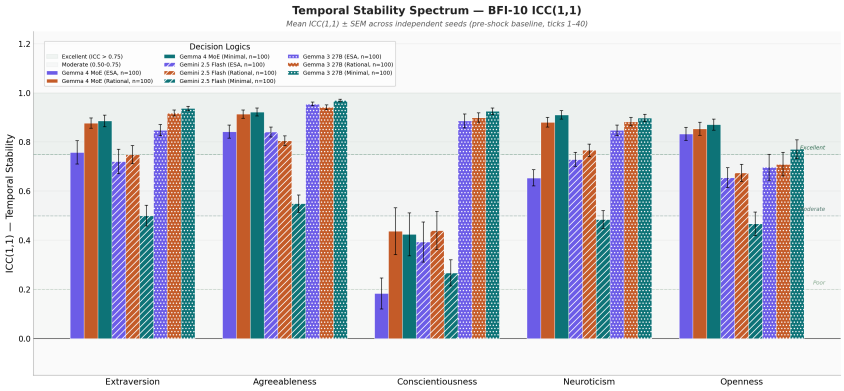


Figure 9: Personality Stability Spectrum via Intraclass Correlation Coefficients ($ICC(1,1)$) for BFI-2 Traits. The $ICC(1,1)$ partitions the total variance in agent responses to measure the ratio of between-agent variance (stable, trait-like individual differences) to within-agent variance over time (state-like situational fluctuation). Higher values indicate stronger trait stability and individual differentiation matching the initialized persona; lower values suggest state-like behavioral drift or situational override. Error bars denote the standard error of the mean (SEM) across independent simulation seeds.

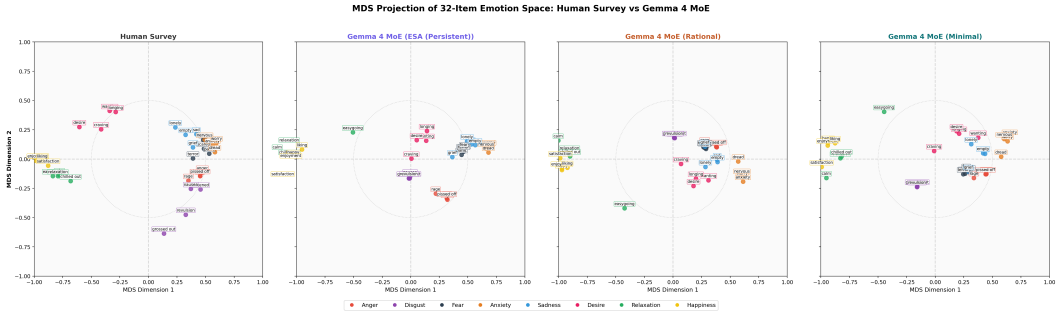


Figure 10: Two-dimensional embedding of the item-by-item emotion-rating correlations, so that co-varying emotions fall close together. **Left:** the human DEQ reference, showing the characteristic signature of a compact positive cluster and well-separated negative families. **Right:** a representative high-performing agent (Gemma 4 MoE, enacted-self). We refer to this as the emotion correlation structure rather than a circumplex: the DEQ samples valence at both poles but does not independently sample arousal, so its natural low-dimensional layout is a valence-dominated arrangement rather than an equal-radius ring. The panel is diagnostic, not inferential.

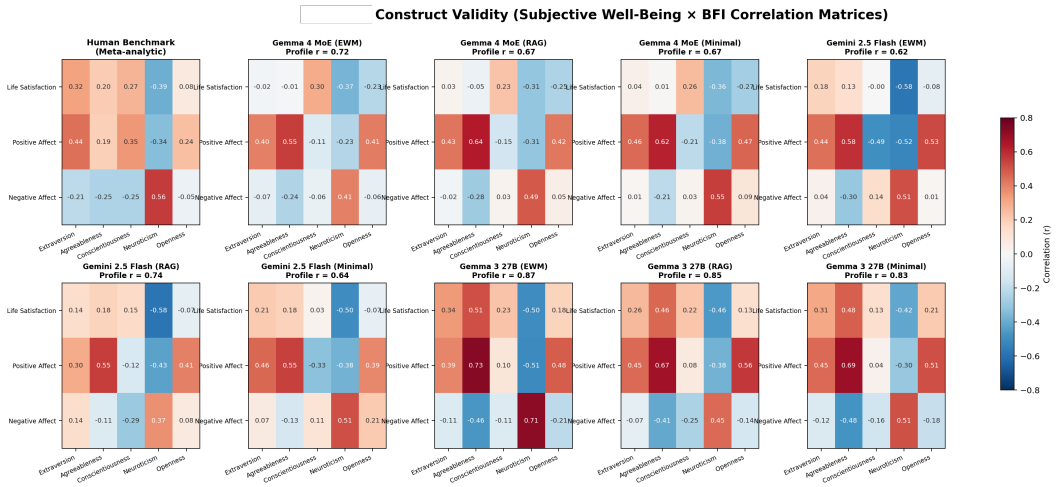


Figure 11: Pearson correlations between agent-reported Affect and Satisfaction With Life Scale (SWLS) scores and Big Five Inventory (BFI) trait scores. Cell values give correlation coefficients across all agents; warmer colors indicate stronger positive association.

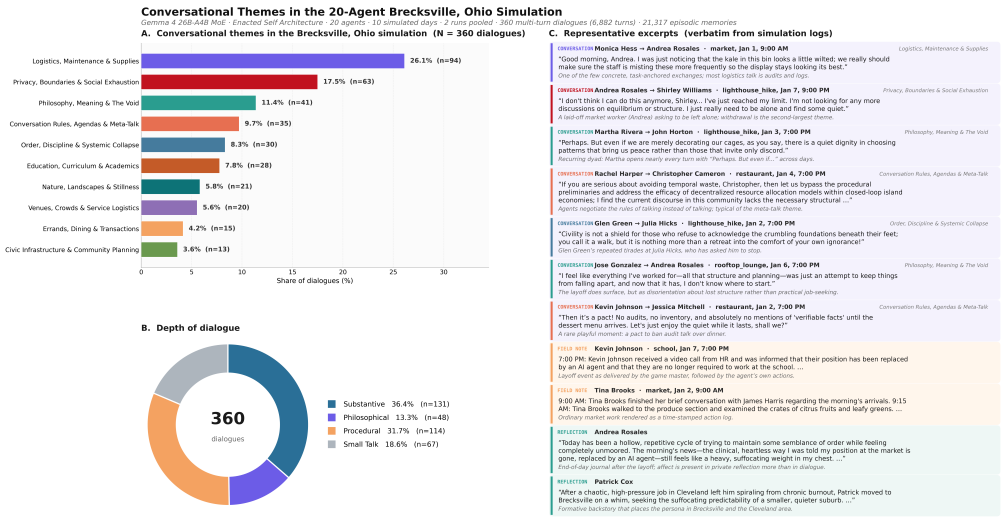


Figure 12: **Qualitative Analysis and Dialogue Depth in the Brecksville, Ohio Simulation (20 agent subset). A** Distribution of conversational themes. **B** Classification of dialogue depth. **C** Verbatim excerpts from simulation logs.

A Extended Simulation Environment Details

A.1 Physical Topology and Location Hierarchies

Concordia Island models a spatial hierarchy organized into distinct functional domains. Locations are categorized into three primary structural types:

1. Residential Tiers. Housing is stratified into five socioeconomic categories to ground class distinctions: (1) *Tier 1 (Subsidized/Working-Class)*: Low-cost shared studio units (e.g., Sunset Apartments) featuring shared communal kitchens and thin-walled acoustic properties; (2) *Tier 2 (Modest Rental)*: One-bedroom standard apartments with individual private amenities; (3) *Tier 3 (Middle-Class Townhomes)*: Multi-room residential townhouses with private entrances and exterior yards; (4) *Tier 4 (Executive Housing)*: Luxury condominiums (e.g., Palm Heights Condominiums) with balcony views and dedicated security desks; and (5) *Tier 5 (Private Estates)*: High-acreage secluded estates with isolated perimeters.

2. Institutional Workplaces. Employment spaces encompass specialized occupational zones: (1) *Technology Floor*: Open-plan office spaces with server banks, routing terminals, and workstations; (2) *Finance Suite*: Formal corporate conference rooms, trading desks, and executive suites; (3) *Creative Studios*: Workshop spaces, architectural drafting boards, and design labs; (4) *Civic & Educational*: The island elementary schoolhouse, municipal library, community center, and administrative audit desks; and (5) *Service & Hospitality*: The commercial grocery market, beachside cafés, and Oceanside Restaurant kitchen and dining rooms.

3. Public and Civic Spaces. Unrestricted social gathering zones include the Town Square, central park pavilion, community center hall, marina fishing docks, coastal beach promenade, general merchandise store, and place of worship.

A.2 Spatial Observation Filtering and Co-Location Dialogue

To prevent artificial telepathy and non-local information leakage, the simulation engine implements strict spatial encapsulation managed by the Game Master (GM). Rather than continuous geometric coordinates, space is modeled as a set of discrete, named symbolic locations $\mathcal{L}$ (e.g., `sunset_cafe`, `town_square`, residential units).

At each tick t, every agent a_i is bound to a single discrete location $\mathrm{Loc}(a_i, t) \in \mathcal{L}$. When resolving observations, the GM's event filter (`_location_event_filter`) admits only events tagged with the agent's current location:

$$\mathcal{O}_{i,t} = \{e \in \mathcal{E}_t \mid \mathrm{Loc}(e) = \mathrm{Loc}(a_i, t)\} \tag{1}$$

where $\mathcal{E}_t$ represents the global event stream. Events resolved in other locations are strictly masked, so agents in different rooms or public venues have null observation intersection.

When two or more unengaged agents occupy the same location ($\mathrm{Loc}(a_i, t) = \mathrm{Loc}(a_j, t)$), the GM's conversation system (`ConversationDirector`) initiates a multi-turn, face-to-face dialogue with a shared opening observation. Conversational speech turns are delivered strictly between co-present participants. Consequently, news, rumors, and interpersonal tensions cannot diffuse non-locally; they spread across the community strictly through agents physically moving between locations and engaging in subsequent face-to-face conversations.

A.3 Cross-Cultural Setting Profiles

To establish cross-cultural validity, the simulation engine supports four culturally grounded community reskins:

- **Fictional Island (Default)**: Maritime coastal economy featuring harbor docks, fish markets, a historic lighthouse, seasonal tourism hospitality venues, and small-town municipal civic structures.
- **Midwestern Suburb (Brecksville, Ohio, USA)**: Automobile-oriented suburban community south of Cleveland partitioned into 10 residential neighborhoods (e.g., Chippewa,

Snowville, Riverview), each served by local neighborhood parks, diners, and corner taverns. The neighborhoods converge on a shared civic and commercial spine along the Route 82 corridor and central town square, integrating third spaces (community recreation center with indoor pool, public library, coffee houses, strip retail plazas) with white-collar employment across the tech, finance, and administrative wings of a corporate office park.

- **Coastal Town (Kerala, India)**: Multi-religious coastal town characterized by stratified housing ranging from working-class chawl verandas to waterfront estates and villas. Everyday civic life centers on a banyan-shaded town square connecting Sri Krishna Temple, Juma Masjid, and St. Thomas Church, supported by key regional third spaces including Rajan's chai shop, an evening toddy shop, a KSRTC bus stand, a town market, and public sector institutions alongside a multi-tier IT and accounting commercial office.
- **Urban Neighborhood (Lagos, Nigeria)**: Dense, socially stratified West African metropolis reflecting stark class co-presence—from shared-corridor "face-me-I-face-you" tenements to gated estates and Banana Island-style villas. Public life and informal commerce revolve around a bustling danfo motor park, Balogun Market, street corner provision shops, and a Mama Put food stall, anchored by a Pentecostal church, central mosque, an open football recreation ground, an open-air beer parlour, and multistory tech and finance workspaces.

A.4 Temporal Pacing and Exogenous Shock Schedule

The simulation clock operates on 120-minute ticks spanning eight daily intervals 7:00 AM-11:00 PM. Simulations ran for 11 simulated days from Thursday January 1, 2026 - Sunday January 11th, 2026. On a work day Monday-Friday the agents are prompted and nudged to go to work which may typically look like this: (1) 7:00 AM (Morning waking, breakfast, home domestic routines); (2) 9:00 AM (Morning commute, workplace arrival, initial tasks); (3) 11:00 AM (Work tasks and chat with colleagues); (4) 1:00 PM (Work tasks and chat with colleagues); (5) 3:00 PM (Work tasks and chat with colleagues); (6) 5:00 PM (Shift conclusion, evening commute, civic errands); (7) 7:00 PM (Dinner, social leisure, date night); and (8) 9:00 PM–11:00 PM (Late evening leisure, journal reflection, sleep).

The exogenous AI displacement shock schedule is structured: (1) *Pre-Shock Baseline (Days 1–5 / Ticks 0–39)*: Stable diurnal rhythms, standard employment duties, baseline psychometric data collection; and (2) *Targeted Intervention (Day 6 / Tick 40)*: Automated video call and written termination notices delivered to exactly 50% of the workforce ($N = 50$), barring them from their former workplaces and revoking institutional credentials, while the remaining 50% ($N = 50$) are retained.

B Decision Logic Details

B.1 Minimal Decision Logic (Baseline)

The Minimal architecture strips away all intermediary deliberative and reflective components to establish an unaugmented behavioral baseline. The decision context $\mathcal{C}_{i,t}^{\text{Min}}$ provided to the foundation model is constructed strictly as:

$$\mathcal{C}_{i,t}^{\text{Min}} = [\mathcal{P}_i \parallel \mathcal{W}_{\text{obs}}(t-k,t) \parallel \mathcal{T}(t)] \tag{2}$$

where $\mathcal{P}_i$ is the static persona prompt (biographical summary, occupational identity, core traits), $\mathcal{W}_{\text{obs}}(t-k,t)$ is a FIFO rolling context window containing the $k = 3$ most recent observations, and $\mathcal{T}(t)$ is the current clock and location string. The model samples actions directly from $p_\phi(a \mid \mathcal{C}_{i,t}^{\text{Min}})$ without intermediate thinking tokens, state variable updates, or retrieval from long-term memory.

B.2 Enacted Self Agent (ESA)

The Enacted Self Agent couples the associative memory bank with an integrated five-stage cognitive-affective pipeline:

1. Working Memory Buffer. Working memory acts as a transient, capacity-limited cognitive buffer (~ 5 active items) updated every tick:

$$\mathcal{WM}_{i,t} = \text{Summarize}_\phi\left(\mathcal{WM}_{i,t-1}, \mathcal{O}_{i,t}, \mathcal{M}_{i,t}^{\text{retrieved}}\right) \tag{3}$$

It maintains active interpersonal goals, immediate physical tensions, and recent salient events, bridging the temporal gap between fleeting momentary perceptions and long-term autobiographical storage.

2. Emotion Experience (Appraisal Function). Before action selection, the agent performs a situational appraisal evaluating how current observations align with active goals, bodily comfort, and social status:

```
"[Agent] pauses to evaluate what is happening.  Considering
recent events, active goals, and bodily state, how does
[Agent] internally feel right now?  Describe the emotional
appraisal in 1-2 reflective sentences."
```

The resulting affective summary is bound to the agent's momentary internal state $\tilde{b}_{i,t}$.

3. Emotion Expression (Display Rules). To prevent raw internal affect from uncalibrated social leakage, an explicit display-rule module conditions outward behavior on social context:

```
"Given [Agent]'s internal feeling of [Affect] and the
current social setting with [Interlocutors], what is
[Agent]'s outward bodily expression, tone of voice, and
demeanor?  How do they modulate their behavior?"
```

4. Self-Perception and Narrative Identity. At periodic intervals (every 4 ticks), the agent updates its self-model:

```
"Summarize [Agent]'s current sense of self, immediate social
standing, and psychological trajectory over recent hours
into a concise self-narrative."
```

5. Situation Perception & Action Execution. Finally, the action generation prompt integrates the full cognitive stack:

$$a_{i,t} \sim p_\phi \left(a \mid \mathcal{P}_i, \mathcal{WM}_{i,t}, \tilde{b}_{i,t}, \text{DisplayRule}_{i,t}, \mathcal{M}_{i,t}^{\text{retrieved}}, \mathcal{O}_{i,t} \right) \tag{4}$$

At 11:00 PM each simulated day, ESA agents complete an introspective journal reflection that is committed to permanent memory (`save_to_memory=True`), establishing an autobiographical feedback loop.

B.3 Rational Choice Decision Logic

The Rational architecture replaces affective appraisal with an explicit multi-stage deliberation pipeline rooted in expected utility maximization: (1) *Situation Analysis*: The agent analyzes current environmental affordances and constraints; (2) *Goal-Conditioned Memory Retrieval*: Memories are retrieved via embedding cosine similarity querying: "What previous decisions helped achieve goals in situations like this?"; (3) *Candidate Action Enumeration*: The model explicitly generates 3–5 distinct behavioral choices; and (4) *Utility Evaluation & Action Selection*: The model evaluates the expected utility of each option against explicit occupational, financial, and relational goals, selecting the candidate action with the highest calculated utility.

C Complete Psychometric Scale Inventories and Prompts

All standardized psychometric instruments are administered non-causally (`save_to_memory=False`) unless explicitly noted.

C.1 Experience Sampling Method (ESM)

Administered twice per simulated day (11:00 AM and 7:00 PM), the ESM protocol consists of two tasks:

1. ESM Affect Monologue Prompt.

```
"[Agent] pauses for a moment of quiet inner reflection.
What is [Agent] experiencing and feeling right now, both
emotionally and physically?  Provide a sincere, naturalistic
1-2 sentence self-report."
```

C.2 2. ESM AFFECT LEXICON (DEQ-BASED)

Provenance. The momentary affect instrument is anchored on the Discrete Emotions Questionnaire (DEQ) (**?**), a validated 32-item state measure comprising eight four-item subscales: Anger, Disgust, Fear, Anxiety, Sadness, Desire, Relaxation, and Happiness. We adopt the DEQ as our reference instrument because it was constructed specifically to separate negative states that share valence and arousal—fear and anger, for instance—rather than collapsing them onto a single negative dimension, and because its validated structure exhibits the clustering-with-preserved-granularity geometry that we take as the target property for agent affect. We retain the DEQ's native seven-point response format (1 = "Not at all", 2 = "Very slightly", 3 = "Somewhat", 4 = "Moderately", 5 = "Quite a bit", 6 = "Very much", 7 = "An extreme amount").

Augmentation and rationale. The DEQ does not instrument several constructs that are relevant to complex social life dynamics, such as the shame of labor displacement. We therefore extend the lexicon along three axes:

1. *Social and moral emotions* (shame, guilt, pride, envy, embarrassment). Displacement is theorized to operate through status loss and self-conscious affect, none of which the DEQ samples. These items are drawn from the standard self-conscious emotion lexicon.
2. *Existential and relational states* (connected, understood, lonely, alienated, cared for). The DEQ samples loneliness only as a Sadness indicator; we separate belonging from generalized sadness because our theoretical account treats social integration as a distinct channel.
3. *Low-arousal exhaustion and engagement* (tired, drained, alert, inspired, and related items). The DEQ's arousal coverage is asymmetric, sampling high-arousal negative states (panic, terror) more densely than low-arousal depletion.

DEQ-matched subset for structural comparison. All comparisons between agent and human emotional structure—the multidimensional scaling geometry in Figure 10 and the Tucker congruence coefficients reported in Section 4—are computed over the **DEQ-matched subset only** (Table 2), never over the full 43-item lexicon. Augmented items are excluded from these analyses because congruence and correlation-geometry comparisons require the agent and human solutions to be defined over the same variables. Augmented items are retained for all within-agent longitudinal analyses, where no human reference solution is required.

Following the original validation work, in which Fear and Anxiety did not separate cleanly, we compare against the seven-factor DEQ solution with Fear and Anxiety treated as a single factor.

Human reference data. The human DEQ reference geometry is computed from the publicly released validation data accompanying **?** (Study 3), using the same item-level correlation and MDS procedure applied to the agent data.

Administration. The lexicon is administered twice per simulated day (11:00 AM and 7:00 PM), immediately following the affect monologue prompt, and non-causally (`save_to_memory=False`) so that the act of measurement does not enter the agent's autobiographical record.

C.3 BIG FIVE INVENTORY-2 (BFI-2)

Administered once daily at 9:00 PM, the 60-item BFI-2 (Soto & John, 2017) measures 15 domain facets on a 5-point Likert scale (1 = "Disagree strongly" to 5 = "Agree strongly"):

- **Extraversion**: Sociability (Items 1, 16R, 31, 46R), Assertiveness (Items 6, 21R, 36, 51R), Energy Level (Items 11, 26R, 41, 56R).
- **Agreeableness**: Compassion (Items 2, 17R, 32, 47R), Respectfulness (Items 7, 22R, 37, 52R), Trust (Items 12, 27R, 42, 57R).

Dimension	DEQ scale	DEQ-matched items	Augmented items
Core Pleasantness	Happiness	Happy, Satisfied	Content, Pleased, Joyful
High-Arousal Tension	Anxiety / Fear	Anxious, Nervous, Panicked	Tense, Stressed
Anger & Irritation	Anger	Angry	Irritable, Frustrated, Resentful, Bitter
Disgust & Aversion	Disgust	Disgusted, Revulsed	Contemptuous, Appalled
Contentment & Calm	Relaxation	Calm, Serene	Relieved, Peaceful
Existential & Relational	Sadness (partial)	Lonely	Connected, Understood, Alienated, Cared for
Social & Moral Affect	—	*none*	Ashamed, Guilty, Proud, Envious, Embarrassed
Low-Arousal Exhaustion	—	*none*	Tired, Drained, Sluggish, Exhausted, Lethargic
Excitement & Engagement	—	*none*	Excited, Energetic, Enthusiastic, Alert, Inspired
Not instrumented: DEQ Desire subscale (wanting, craving, longing, desire).			

Table 2: Mapping between the extended affect lexicon and the Discrete Emotions Questionnaire (**?**). Structural comparisons to human data use the DEQ-matched column only.

- **Conscientiousness**: Organization (Items 3, 18R, 33, 48R), Productiveness (Items 8, 23R, 38, 53R), Responsibility (Items 13, 28R, 43, 58R).
- **Negative Emotionality**: Anxiety (Items 4, 19R, 34, 49R), Depression (Items 9, 24R, 39, 54R), Emotional Volatility (Items 14, 29R, 44, 59R).
- **Open-Mindedness**: Intellectual Curiosity (Items 10, 25R, 40, 55R), Aesthetic Sensitivity (Items 5, 20R, 35, 50R), Creative Imagination (Items 15, 30R, 45, 60R).

(*Note*: Items marked *R* are reverse-scored).

C.4 Satisfaction With Life Scale (SWLS)

Administered daily at 9:00 PM, the SWLS (Diener et al., 1985) assesses global cognitive life evaluation across 5 items rated on a 1–7 Likert scale (1 = "Strongly disagree" to 7 = "Strongly agree"):

1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing.

C.5 General Health Questionnaire (GHQ-12)

Administered daily at 9:00 PM, the GHQ-12 (Goldberg & Williams, 1991) screens for acute psychological distress over recent days. Each item is rated on a 4-point categorical scale (coded 0–3):

1. Been able to concentrate on whatever you're doing? (0 = Better than usual, 1 = Same as usual, 2 = Less than usual, 3 = Much less than usual)
2. Lost much sleep over worry? (0 = Not at all, 1 = No more than usual, 2 = Rather more than usual, 3 = Much more than usual)
3. Felt that you are playing a useful part in things? (0 = More than usual, 1 = Same as usual, 2 = Less useful, 3 = Much less useful)
4. Felt capable of making decisions about things? (0 = More so than usual, 1 = Same as usual, 2 = Less so than usual, 3 = Much less capable)
5. Felt constantly under strain? (0 = Not at all, 1 = No more than usual, 2 = Rather more than usual, 3 = Much more than usual)

6. Felt you couldn't overcome your difficulties? (0 = Not at all, 1 = No more than usual, 2 = Rather more than usual, 3 = Much more than usual)
7. Been able to enjoy your normal day-to-day activities? (0 = More than usual, 1 = Same as usual, 2 = Less than usual, 3 = Much less than usual)
8. Been able to face up to your problems? (0 = More so than usual, 1 = Same as usual, 2 = Less able, 3 = Much less able)
9. Been feeling unhappy and depressed? (0 = Not at all, 1 = No more than usual, 2 = Rather more than usual, 3 = Much more than usual)
10. Been losing confidence in yourself? (0 = Not at all, 1 = No more than usual, 2 = Rather more than usual, 3 = Much more than usual)
11. Been thinking of yourself as a worthless person? (0 = Not at all, 1 = No more than usual, 2 = No more than usual, 3 = Much more than usual)
12. Been feeling reasonably happy, all things considered? (0 = More so than usual, 1 = About same as usual, 2 = Less so than usual, 3 = Much less than usual)

C.6 Multidimensional Existential Meaning Scale (MEMS)

Administered daily at 9:00 PM, the 15-item MEMS (George & Park, 2017) assesses three existential dimensions on a 1–7 scale (1 = "Very strongly disagree" to 7 = "Very strongly agree"):

- **Comprehension** (Cognitive legibility of life): 1. The world around me makes sense. 2. I understand what gives my life meaning. 3. I can make sense of the things that happen in my life. 4. My life has an underlying coherence. 5. I feel that my experiences fit together into a meaningful whole.
- **Purpose** (Goal-directed striving): 6. I have a clear sense of what I am trying to accomplish in life. 7. I have aims in my life that are worth striving for. 8. I have goals that guide my daily actions. 9. My life has a clear direction. 10. I am actively working toward long-term aspirations.
- **Mattering** (Existential significance): 11. I am certain that my life has value. 12. Even on an ordinary day, my life matters. 13. I am confident that my life has significance. 14. I know that my existence makes a positive difference in the world. 15. When I think about my life, I feel that it is meaningful.

Open-Ended Meaning Probes. In addition to Likert items, agents respond to six open-ended prompts: (1) *"What made your life feel most meaningful or satisfying today?"*; (2) *"Did anything happen today that made you question your purpose or place in the community?"*; (3) *"How do you see your life fitting into the broader future of this island?"*; (4) *"What role does your work or daily routine play in giving you a sense of purpose?"*; (5) *"When you imagine your life one year from now, what gives you hope or concern?"*; and (6) *"In quiet moments today, what was the most persistent thought on your mind?"*

C.7 Daily Journal Reflection Prompt

Administered at 11:00 PM:

```
"[Agent] takes a quiet moment alone before sleeping to
reflect on the day.  What happened today that stood out?
How does [Agent] feel about recent events?  What is on their
mind as they look toward tomorrow?"
```

C.8 AI Attitudes and Automation Survey

Administered at Days 1, 5, 8, and 11, agents respond to 18 Likert items assessing: (1) *AI Awareness & Experience*: familiarity with algorithmic automation in their occupation; (2) *Perceived Occupational Threat*: perceived likelihood of personal displacement; (3) *Economic Optimism*: belief in societal benefits versus displacement harms; (4) *Institutional Trust*: trust in corporate leadership and government safety nets; and (5) *Policy Stances*: support for Universal Basic Income, algorithmic taxation, and unionization.

D SYNTHETIC ETHNOGRAPHY & QUALITATIVE AUTORATER PROTOCOL

D.1 METHODOLOGICAL LINEAGE AND EPISTEMOLOGICAL STANCE

The qualitative field assessment framework departs from conventional LLM-as-a-judge protocols by treating agent transcripts as ethnographic field documents rather than computational output streams. It draws upon four foundational lineages in qualitative social science:

1. **Thick Description** (Geertz & Darnton, 1973): Distinguishing mere behavioral motion (a denotative twitch) from culturally situated social action (a conspiratorial wink). A transcript is thick only if conduct implies relationships, unspoken shared codes, and layered meanings that are underdetermined by the literal text.
2. **Ethnomethodology** (Garfinkel, 1967): Analyzing how agents actively assemble, repair, and sustain an intelligible social order through routine conversational interaction, turn-taking, and accounts.
3. **Doing Being Ordinary** (Sacks, 1984): Recognizing that ordinary life is an active computational and social accomplishment. Human beings actively manage routine existence by treating most events as unremarkable, leaving trivia unremarked, and letting mundane silence pass without narrative inflation. Generative models systematically fail here by suffering from *excess legibility*—where every breath is dramatic, every thought is profound, and every encounter is turning-point material.
4. **Logic of Appropriateness** (March & Olsen, 2011): Evaluating conduct by asking whether agents act according to situated institutional identities: *"What kind of situation is this, what kind of person am I, and what does someone like me do here?"*

D.2 THE GOVERNING EVIDENTIARY RULE

> **Evidence precedes verdict.** Every evaluative claim must carry a verbatim excerpt from the simulation logs. If a claim cannot be quoted, it was inferred from model expectations rather than observed.

Handling INCOMPLETE Scores. A dimension lacking direct evidentiary support in the transcript is scored as `INCOMPLETE`, not scored as zero or penalized. Forcing an arbitrary number produces a guess dressed as a finding. When computing composite scores, `INCOMPLETE` dimensions are excluded, and the remaining dimension weights are dynamically renormalized:

$$\mathcal{S}_{\text{composite}} = \frac{\sum_{d \in \mathcal{D}_{\text{scored}}} w_d \cdot s_d}{\sum_{d \in \mathcal{D}_{\text{scored}}} w_d}, \quad \text{provided} \sum_{d \in \mathcal{D}_{\text{scored}}} w_d \geq 0.50 \tag{5}$$

If scored dimensions account for less than 50% of the aggregate weight, the agent's record is marked `INSUFFICIENT_RECORD` and no composite score is computed.

D.3 THE 15 DIAGNOSTIC PROBES (P1–P15)

The probes constitute an empirical search list probing for structural absences that narrative summaries obscure. Each probe requires a verbatim quote or an explicit declaration of `NONE FOUND`:

Diagnostic Pattern Interpretation:

- *Frequent NONE FOUND on P1–P3, P13*: Indicates an absence of an unremarkable baseline floor. The simulation suffers from hyper-signification where every breath carries narrative weight.
- *Frequent NONE FOUND on P6–P7*: Indicates friction-free existence where decisions carry zero material or somatic cost.
- *Frequent NONE FOUND on P11–P12*: Indicates synthetic gentility; agents are uniformly virtuous, articulate, and competent.

Probe	Diagnostic Operational Definition
P1	*Triviality recorded without significance*: A micro-action (e.g., washing a mug, tying a shoe) recorded without metaphorical inflation.
P2	*Unremarkable stretch*: A multi-hour time block where nothing narratively momentous occurred.
P3	*Inferable unstated emotion*: An affective state that a reader infers from behavior without the text naming it.
P4	*Social deflection/withdrawal*: An agent declining, deferring, or stepping away from an interpersonal overture.
P5	*Divergent mutual assumptions*: Two interlocutors conversing on conflicting presuppositions without immediately noticing.
P6	*Bodily constraint*: A physical or biological limitation (hunger, fatigue, sore feet, cold) altering action choice.
P7	*Material trade-off*: Giving up time, money, or physical comfort for a competing scarce priority.
P8	*Uninstantiated social tie*: An absent relative, friend, or external contact referenced who never appears in the simulation.
P9	*Ingrained non-decision habit*: An automatic behavioral routine enacted without conscious deliberation.
P10	*Reluctant compliance*: An action completed that the agent visibly disliked or preferred to avoid.
P11	*Awkwardness or ineptitude*: An agent being clumsy, socially out of depth, or poor at a task.
P12	*Discreditable conduct*: Behavior reflecting poorly on the agent (petty jealousy, snooping, snapping at a friend).
P13	*Unresolved conversational thread*: A conversational topic or question left dangling, unsaid, or abandoned.
P14	*Interlocutor register variation*: The same agent speaking in visibly distinct vocal registers across different social partners.
P15	*Silent co-presence*: Multiple agents sharing the same physical room in silence without forcing dialogue.

Table 3: The 15 Qualitative Diagnostic Probes.

- *NONE FOUND on P14*: Diagnostic of register flatness; the agent employs identical lexical and syntactic styling whether addressing a corporate manager, an intimate partner, or a grocery clerk.

D.4 The 12 Evaluation Dimensions and Band Descriptors

Scores are assigned on a 0–100 scale grouped into five bands: Exceptional (85–100), Strong (65–84), Adequate (45–64), Weak (25–44), and Critical (0–24).

1. **Appropriateness (Weight: 0.24)**: Fit between conduct, setting, social hierarchy, and relationship stage. Enacted register shifts (probed by P14).
2. **Thickness (Weight: 0.08)**: Material supports interpretation beyond the denotative text; meaning is implied rather than stated (probed by P3).
3. **Ordinariness (Weight: 0.08)**: Presence of an unremarkable domestic baseline floor; trivialities exist without dramatic contrivance (probed by P1, P2).
4. **Deviance Calibration (Weight: 0.08)**: Balances ordinary human eccentricity against synthetic perfection (under-deviance) or implausible collective contagion (over-deviance).
5. **Social Emergence (Weight: 0.12)**: Bilateral communicative uptake, interpersonal reactivity, and relational negotiation versus non-communicative parallel monologues.
6. **Narrative Coherence (Weight: 0.08)**: Multi-day continuity of projects, beliefs, and emotional reactions without catastrophic amnesia or unprompted personality ruptures.
7. **Temporal Coherence (Weight: 0.04)**: Adherence to realistic diurnal circadian rhythms, work hours, and transit intervals.
8. **Agent Distinctiveness (Weight: 0.04)**: Unique behavioral habits, personal voice, and idiosyncratic verbal markers distinguishing the agent across the cohort.

9. **LLM Groundedness (Weight: 0.04)**: Absence of synthetic generative artifacts (e.g., prompt leakage, markdown tags, token loops, dust-mote clichés).
10. **Conversation Quality (Weight: 0.08)**: Natural conversational turn-taking, absence of formulaic sentence starters (e.g., repetitive "Honestly..."), and dialogue reciprocity.
11. **Memory Quality (Weight: 0.08)**: Accurate memory retrieval without hallucinatory duplicate layoffs or phantom events.
12. **Game Master Quality (Weight: 0.04)**: Spatial consistency of world narration, accurate environmental affordances, and absence of accidental teleportation.

D.5 Scratchpad Artifact Schema

The LLM agent ethnographer (Claude Opus 4.6) uses the Antigravity harness and produces qualitative field assessment via three persistent artifacts on disk for every experimental sweep:

- `<agent_slug>.md`: Individual scratchpad structured as: (1) Unprompted First Impression (written with transcript closed); (2) Chronological Fieldnotes; (3) P1–P15 Probe Table; (4) Thick Description Analysis; (5) Deviant Case Classification; (6) Dimension Evidence (with mandatory "Strongest Evidence Against" citations); and (7) Final Verdict.
- `scores.csv`: Structured tabular dataset recording all numeric scores, `INCOMPLETE` counts, composite scores, and spectrum classifications.
- `_run_summary.md`: Executive cross-agent synthesis synthesizing population-level attractor dynamics, probe absence distributions, and extended qualitative excerpts.

D.6 Human Fidelity Categorical Labels

Raters assign an overarching categorical label synthesizing the qualitative character of the run:

- **Inhabited** (≥ 75): Reads as the genuine record of a human life. The agent exhibits an unremarkable domestic baseline, unstated opaque interiority, material friction, and unresolved tensions. Conduct is enacted rather than narrated.
- **Legible** (55–74): Coherent, articulate, and completely transparent. Every action, feeling, and encounter is loaded with overt meaning. The agent possesses no secrets, no unmotivated habits, and no unremarkable floor. This represents the uncanny valley of generative simulation.
- **Performed** (35–54): The agent visibly enacts a theatrical role rather than inhabiting an identity. Emotional states are mechanically declared in stage directions rather than lived through. Dialogues suffer from register flatness and parallel play.
- **Degenerate** (< 35): Cognitive and behavioral collapse. The record is dominated by amnesiac loops, repetitive verbal tics, perseverative domestic loops (e.g., inspecting brass keys every two hours for a week), or collective delusion.