MINDWORLD: SIMULATING HUMAN BEHAVIOR IN STRATEGIC SCENARIOS VIA COGNITIVE WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as human simulators, yet they excel at mimicking helpful and cooperative assistants rather than authentic, imperfect humans who hesitate, get defensive, or become confused. Existing end-to-end LLM simulators implicitly determine cognitive state changes via next-token distribution, and this uncontrollability often leads to excessive helpfulness. To bridge this gap, we propose **MINDWORLD**, a lightweight human simulation framework that models the cognitive world as a deterministic state-space. We anchor the cognitive state in the Big Five personality traits and dynamically simulate state transitions using three deterministic mathematical constraints: continuous damped cognitive dynamics, non-linear cusp-bifurcation stress modeling, and discrete masked Markov stage transitions. These mechanisms provide structural controls on simulated behavior rather than a biological account of human cognition. Across four challenging strategic domains, including clinical communication, bargaining, cooperation, and deception, MINDWORLD generates responses closer to human references on average than the evaluated baselines across all three LLM backbones, while improving several behavioral risk measures. A sensitivity analysis further characterizes how coefficient scaling influences internal state transitions. Blinded human evaluations show that MINDWORLD achieves tie-excluded preference rates of 67.4% against the state-of-the-art baseline and 68.3% against human reference responses in overall realism, supporting explicit state control as a useful design principle for strategic human simulation.

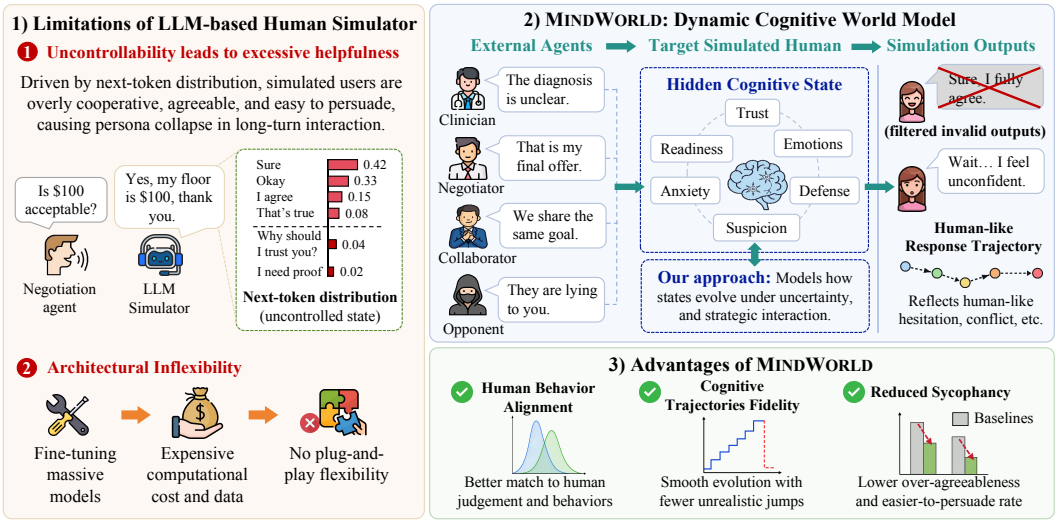


Figure 1: LLM-based simulators often over-cooperate, collapse cognitive states into next-token predictions, and struggle under imperfect information. MINDWORLD models target human behavior through evolving hidden cognitive states, and produces response trajectories with better behavioral alignment and reduced sycophancy.

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.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable success across diverse applications, increasingly serving as foundational infrastructure for clinical assistants (Li & Lebai Lutfi, 2026; Lai et al., 2026), negotiation systems (Zhan et al., 2024), and interactive dialogue agents (Ni et al., 2026; Sun et al., 2025b). To guarantee safety, alignment, and immediate utility, models are heavily optimized through reinforcement learning from human feedback (RLHF) and strict instruction tuning (Ouyang et al., 2022). This paradigm ensures that LLMs are fundamentally helpful, polite, and compliant AI assistants, inclined to satisfy user requests and facilitate seamless task completion.

However, directly deploying these models as human simulators to train or evaluate downstream agents creates a major challenge. While a polite and helpful assistant bias is ideal for general tasks, it makes LLMs highly unrealistic at simulating genuine human behavior. In real-world interaction, humans naturally exhibit cognitive friction, including skepticism, hesitation, emotional defense, and non-linear behavioral shifts (Frith & Frith, 2005; Liu et al., 2024). As illustrated in Figure 1, existing human simulators are highly uncontrollable because they use a single end-to-end process for both cognitive changes and response generation. This end-to-end simulation approach suffers from two bottlenecks which make it inefficient for real-world deployment. **(1) Uncontrollability leads to excessive helpfulness:** Driven by next-token probabilities, simulated users are often over-cooperative, over-agreeable, easily manipulated by basic prompt descriptions, and change their minds too quickly under mild pressure (Sharma et al., 2024; Cheng et al., 2026). This text-driven drift causes their designated persona to collapse during long-turn interactions. **(2) Architectural Inflexibility:** Other methods rely on fine-tuning massive models to handle complex behavioral changes, which requires expensive data and training costs. This setup removes plug-and-play flexibility, preventing it from generalizing across different foundation models.

To this end, we introduce **MINDWORLD**, a lightweight human simulation framework formulated as a deterministic state-space cognitive world model. The cognitive state is updated by deterministic mathematical formulations, while LLM renders the resulting state into language, and a hybrid rule-based and LLM-based arbiter checks the response, making the simulation more controllable and interpretable. Our framework conditions the deterministic controls on Big Five personality traits (McCrae & John, 1992) through three key mechanisms: (1) Damped cognitive module employs continuous variables bounded by a damping equation to ensure cognitive traits evolve gradually (Gardner, 2006); (2) Non-linear cusp-bifurcation stress modeling (Murphy, 1991) captures high-pressure psychological reactions by simulating critical state collapses when external pressure or internal conflict crosses a tipping point; (3) Discrete masked Markov stage transitions (Pachet et al., 2011) enforces sequential boundaries using behavioral masks to prohibit unrealistic stage jumps from resistance to stable agreement. Finally, a state-consistency arbiter implements a filtering layer that blocks invalid text. Our setting is a multi-turn interpersonal interaction in which a target participant's response depends on persistent goals and partially observable states, instead of a universal model of cognition.

We evaluate MINDWORLD across four challenging strategic domains (clinical communication, bargaining, cooperation, and deception) covering seven interaction scenarios. The main comparison measures alignment with human reference responses, persona-conditioned variation, and operational risks associated with over-agreement and abrupt response changes. Across multiple foundation models, MINDWORLD improves behavioral alignment with real target responses, preserves persona-conditioned cognitive diversity, and reduces key uncontrollability failures such as abrupt state jumps, premature agreement, and sycophantic over-cooperation. We conduct a sensitivity analysis to examine how changes in personality-related coefficients affect internal state updates. Ablation results further show that each lightweight control module suppresses a specific failure mode: damped updates control continuous state changes, masked transitions control stage progression, cusp-bifurcation captures structured stress reactions, and the arbiter controls state-consistent language realization. Blinded human evaluations show that MINDWORLD achieves tie-excluded preference rates of 67.4% against the state-of-the-art baseline and 68.3% against human reference responses in overall realism, supporting explicit state control as a useful design principle for strategic human simulation.

Our contributions are threefold:

- We identify **controllability** as a central design requirement for LLM-based human simulation in multi-turn strategic interactions. We distinguish fluent role-play from the ability to maintain persona-conditioned cognitive dynamics, represent persistent goals and context-dependent resistance, and

generate state-compatible responses. This perspective motivates explicit control over state evolution to address excessive helpfulness, premature agreement, and unsupported behavioral shifts.

- We propose **MINDWORLD**, a lightweight cognitive world model that separates deterministic state updates from LLM response generation. The framework integrates persona-conditioned damped updates, stress-mode selection, masked stage transitions, and response arbitration. We specify their implementation and formalize the structural properties they enforce under stated assumptions.
- We conduct a systematic evaluation across four strategic interaction domains, covering behavioral alignment, persona sensitivity, and behavioral risk measures. The results show improved average alignment with human reference responses and reductions in several measured risks. In human evaluation, we achieve a 67.4% preference rate over the state-of-the-art baseline in overall realism.

2 RELATED WORK

Social Simulation. A major line of LLM-based social simulation builds agents with memory, reflection, and planning for long-horizon interaction (Park et al., 2023). Subsequent work studies multi-agent communication, social environments, and collaborative tasks, typically evaluating agents by observable behavior and task outcomes (Li et al., 2023; Lin et al., 2023; Vezhnevets et al., 2023; Zhou et al., 2024; Sun et al., 2025a). LLMs have also been used as simulated respondents, experimental participants, and economic agents (Argyle et al., 2023; Aher et al., 2023; Horton et al., 2023). However, these methods usually focus on external behavior rather than hidden state, leading to excessive helpfulness. MINDWORLD addresses this limitation by separating cognitive state updates from response rendering and constraining the simulator via deterministic state-space.

Role-Playing and Sycophancy. Role-playing research studies how LLMs enact users, or task-specific participants in dialogue (Chen et al., 2025; 2024; Tseng et al., 2024). Prior work improves role fidelity through character profiles, role-conditioned prompting, tuning, and verifier-style constraints (Shao et al., 2023; Wang et al., 2024; Luo et al., 2024). However, persona conditioning alone does not prevent a model from following pressure from the other speaker, a failure mode related to sycophancy in which language models favor agreement over consistency with the simulated target (Perez et al., 2023; Sharma et al., 2024; Wei et al., 2024). MINDWORLD instead parameterizes cognitive dynamics with persona traits and uses a state-consistency arbiter to block responses that contradict the hidden state of simulator.

3 MINDWORLD

MINDWORLD is a lightweight human simulation framework that models the cognitive world as a deterministic state-space. Given a dialogue context, a target profile, and an external utterance, MINDWORLD updates the hidden cognitive state and then generates a response that remains compatible with the state. This differs from existing prompt-based or fine-tuned simulators, where LLMs decide the response through next-token prediction. MINDWORLD separates state transition from language realization. LLMs may parse utterances and render responses, but cannot freely rewrite the hidden state. Cognitive state changes are updated by explicit mechanisms and checked by constraints.

The pipeline consists of persona and state initialization (Section 3.1), a capped regular update (Section 3.2), and stress adjustment, masked stage selection, and state-constrained response rendering (Section 3.3). Numerical state updates and default stage selection are deterministic given the parsed features, persona, and current state, whereas parsing, language generation, and semantic validation can vary across LLM calls. Appendix C records the parameters, schemas, and execution rules.

3.1 HIDDEN COGNITIVE STATE

State representation. In strategic interaction, the external agent observes only the utterances of the target simulator, not the true cognitive state (Kaelbling et al., 1998). We therefore represent the target simulator with a hybrid hidden state that combines continuous cognitive variables with a discrete interaction stage (Alur et al., 2000). We represent the target participant using a hybrid hidden state $S_t = (C_t, M_t)$, where $C_t \in [0,1]^d$ denotes the continuous cognitive state and $M_t \in \{1, \ldots, K\}$ denotes the discrete interaction stage. In our implementation, $K = 5$, corresponding to resistance, contemplation, preparation, action, and maintenance (Prochaska & Velicer, 1997).

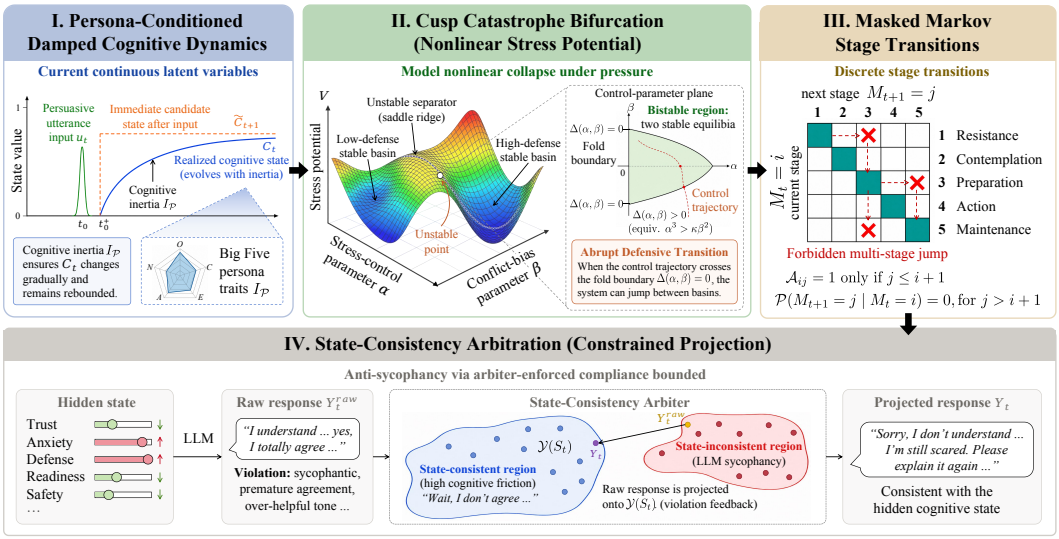


Figure 2: **MINDWORLD method overview.** Parsed utterance features drive persona-conditioned candidate updates, capped regular dynamics, stress adjustments, and masked stage selection. An LLM renders the updated state, and a hybrid arbiter validates the response with bounded retries. The regular-update cap does not apply to subsequent defensive overrides or stress adjustments.

We distinguish shared interaction variables from variables with domain-dependent interpretations. The shared variables are trust, anxiety, dissonance, defense, understanding, readiness, suspicion, goal commitment, and perceived control. Additional variables represent information preference, safety, cooperativeness, competitiveness, deception intent, and social pressure. Their interpretation and relevance depend on the scenario: information preference is useful in clinical communication and cooperation, whereas deception intent is relevant to strategic concealment. The current implementation stores all fifteen continuous variables in a common state representation. This conceptual grouping does not imply that irrelevant dimensions are frozen. The discrete stage M_t is a coarse computational progress variable, and its forward-transition mask constrains the simulator rather than asserting a universal sequence of human behavior.

Component	Variables / Examples	Interpretation
Shared interaction state	Trust, Anxiety, Dissonance, Defense, Understanding, Readiness, Suspicion, Goal Commitment, Perceived Control	Persistent interpersonal state used by the control modules.
Domain-dependent state	Information Preference, Safety, Cooperativeness, Competitiveness, Deception Intent, Social Pressure	Variables whose behavioral meaning and relevance depend on the domain.
Discrete stage	Resistance/precontemplation, contemplation, preparation, action, maintenance	Coarse progression with at most one forward stage per turn.

Table 1: Conceptual organization of the hidden state. The implementation retains a common fifteen-dimensional continuous vector. Domain relevance describes interpretation and use, not a hard activation mask. The stage variable provides an explicit progression constraint.

Persona controls dynamics. The Big Five profile is represented as $\mathbf{p} = (O, C, E, A, N) \in [0,1]^5$. Rather than using these traits only as textual descriptors, we transform the centered profile $\widetilde{\mathbf{p}} = \mathbf{p} - \frac{1}{2}\mathbf{1}$ into bounded control parameters: $\boldsymbol{\theta}_\mathcal{P} = \Pi_\Theta(\mathbf{b} + \mathbf{W}\widetilde{\mathbf{p}})$, where $\mathbf{b}$ specifies the reference configuration, $\mathbf{W}$ encodes fixed trait-dependent adjustments, and Π_Θ projects each parameter to its admissible range. The resulting controls govern state inertia, stress sensitivity, transition thresholds, and selected response-generation settings. Thus, the same interaction evidence can induce different state trajectories for different supplied profiles. The mapping is a transparent modeling choice rather than an empirically estimated psychological law. Appendix C reports the complete numerical configuration, parameter ranges, and the execution paths in which each control is active.

Initialization and persona variants. The continuous cognitive state is initialized from the profile initial hidden state and projected onto $[0, 1]$. Unspecified fields use the documented defaults, and the default interaction stage is contemplation. A missing Big Five trait is assigned the reference value of 0.5. For repeated variants of same cases, variant zero retains the supplied profile, while the remaining variants apply reproducible, zero-centered local perturbations to its Big Five traits before projection onto $[0, 1]$. These variants are constructed to evaluate the simulator's sensitivity to nearby persona changes. They are not independently measured personalities of the original participants. Appendix C specifies the perturbation distribution, amplitude, clipping procedure, and random-seed construction.

3.2 Damped Cognitive Update

Cognitive and affective states exhibit temporal persistence (Koval et al., 2012; Scheffer et al., 2022); a single utterance should not immediately create full trust or agreement. MindWorld therefore parses each utterance into bounded interaction features, constructs a deterministic candidate state, and applies persona-conditioned damping. The clinical and generic parsers use domain-specific schemas but share the same update interface, and outputs are clipped to $[0, 1]$.

Candidate construction and capped update. Let $\mathbf{f}_{t+1}$ denote the parsed features. For state component k, a fixed coefficient vector $\mathbf{w}_k$ maps relevant interaction signals, current-state variables, and predefined interactions to a candidate update: $g_{t,k} = \mathbf{w}_k^\top \boldsymbol{\phi}_k(C_t, \mathbf{f}_{t+1}, \mathbf{p})$ and $\widetilde{C}_{t+1,k} = \mathrm{clip}(C_{t,k} + g_{t,k}, 0, 1)$. For example, empathy and provision of space support trust, whereas pressure, concealment, and threat reduce it. The candidate is then combined with the previous state using persona-conditioned inertia and a componentwise change cap:

$$\Delta_{\mathcal{P},k} = \delta_{\mathrm{reg}}(1 - I_{\mathcal{P},k}),$$
$$C^{\mathrm{reg}}_{t+1,k} = \mathrm{clip}\Big(C_{t,k} + \mathrm{clip}\Big((1 - I_{\mathcal{P},k})(\widetilde{C}_{t+1,k} - C_{t,k}), -\Delta_{\mathcal{P},k}, \Delta_{\mathcal{P},k}\Big), 0, 1\Big).$$

Here $I_{\mathcal{P},k}$ is the persona-conditioned inertia and δ_{reg} is the global update scale. The regular state remains in $[0, 1]$ and changes by at most $\Delta_{\mathcal{P},k}$ per turn. Later defensive and stress adjustments are not covered by this cap. Appendix C provides the parser schemas, feature mappings, and coefficients.

3.3 Stress, Stage, and Response Realization

Following the regular update, MindWorld applies a stress-dependent adjustment, updates the discrete stage, and renders a response from the resulting state.

Cusp-inspired stress switching. To represent abrupt reactions to pressure, uncertainty, threat, or perceived concealment, we use a thresholded mechanism inspired by cusp models of discontinuous change (Murphy, 1991). Let C^*_{t+1} be the state after the regular update and any defensive override. External stress and internal conflict are summarized by

$$q_t = f_{\mathrm{pressure}} - f_{\mathrm{empathy}} + f_{\mathrm{threat}} + f_{\mathrm{jargon}}\left(1 - C^*_{t+1,\mathrm{Understanding}}\right) + C^*_{t+1,\mathrm{Dissonance}} - C^*_{t+1,\mathrm{Trust}} + f_{\mathrm{concealment}} + f_{\mathrm{uncertainty}}.$$

The stress risk is $r_{\mathrm{stress}} = \sigma(\kappa_{\mathrm{stress}}(q_t + \gamma - \tau_{\mathrm{stress}}))$, where γ is persona-conditioned sensitivity and κ_{stress} and τ_{stress} control transition sharpness and location. Below the activation threshold, the state remains stable; otherwise, a fixed score selects a bounded stress-response adjustment. This is a computational switching mechanism, not a numerical solution of a classical cusp potential.

Masked stage transition. For current stage $i = M_t$, a fixed function assigns each candidate stage the score $\ell_{ij} = h_{\boldsymbol{\eta}}(j \mid i, C_{t+1}, \mathbf{f}_{t+1}, \mathbf{p})$. Trust, understanding, readiness, safety, empathy, and strategy favor progression, whereas defense, anxiety, and dissonance favor staying or moving backward. A binary mask $\mathcal{A}_{ij}$ removes transitions that skip a forward stage ($j > i + 1$) and can also block progression when trust gates fail. Probabilities are normalized over the permitted stages:

$$P_{ij} = \frac{\mathcal{A}_{ij} \exp(\ell_{ij})}{\sum_{r=1}^{K} \mathcal{A}_{ir} \exp(\ell_{ir})}.$$

Staying is always allowed; the deterministic configuration selects the highest-scoring permitted stage. Backward movement is not limited to one step. The forward mask is a design constraint, not a claim that people universally follow these stages.

State-constrained rendering and arbitration. The renderer generates a response from the updated state, stage, persona, dialogue context, and stress mode without modifying the numerical state. A hybrid arbiter applies rule-based contradiction checks followed by semantic validation. Rejected outputs are regenerated with feedback while the state remains fixed, up to a retry limit; if semantic validation fails, the rule-based decision is retained. Implementation details and formal assumptions are provided in Appendices C and B.

Algorithm 1 Complete MINDWORLD simulation pipeline

Require: Target profile $\mathcal{P}$, context H_0, utterances $x_{1:T}$, parameters Θ, retry limit $R_{\max}$
Ensure: Responses $Y_{1:T}$, state trajectory $S_{0:T}$, and diagnostics $D_{1:T}$

$\mathbf{p} \leftarrow \textsc{NormalizeBigFive}(\mathcal{P})$; $\boldsymbol{\theta}_{\mathcal{P}} \leftarrow \textsc{PersonaControls}(\mathbf{p}, \Theta)$
$S_0 = (C_0, M_0) \leftarrow \textsc{InitializeState}(\mathcal{P}, \Theta)$
for $t = 0, \ldots, T-1$ **do**
  $\mathbf{f}_{t+1} \leftarrow \mathrm{clip}(\textsc{ParseOrFallback}(x_{t+1}, H_t, S_t, \mathcal{P}), 0, 1)$
  $C^{\mathrm{reg}}_{t+1} \leftarrow \textsc{CappedUpdate}(C_t, \mathbf{f}_{t+1}, \mathbf{p}, \boldsymbol{\theta}_{\mathcal{P}})$
  $C^{*}_{t+1} \leftarrow \textsc{DefensiveOverride}(C^{\mathrm{reg}}_{t+1}, \mathbf{f}_{t+1}, \boldsymbol{\theta}_{\mathcal{P}})$
  $(C_{t+1}, z_{t+1}, r_{t+1}) \leftarrow \textsc{StressAdjust}(C^{*}_{t+1}, \mathbf{f}_{t+1}, \boldsymbol{\theta}_{\mathcal{P}})$
  $M_{t+1} \leftarrow \textsc{MaskedStageTransition}(M_t, C_{t+1}, \mathbf{f}_{t+1}, \boldsymbol{\theta}_{\mathcal{P}})$
  $S_{t+1} \leftarrow (C_{t+1}, M_{t+1})$
  $(Y_{t+1}, d_{t+1}, a) \leftarrow \textsc{RenderAndArbitrate}(H_t, x_{t+1}, \mathcal{P}, S_{t+1}, \mathbf{f}_{t+1}, z_{t+1}, R_{\max})$ ▷ S_{t+1} remains fixed
  if $Y_{t+1} = \textsc{Failure}$ **then**
    raise generation error
  end if
  $D_{t+1} \leftarrow (d_{t+1}, z_{t+1}, r_{t+1}, a)$; $H_{t+1} \leftarrow \textsc{Append}(H_t, x_{t+1}, Y_{t+1})$
end for
return $Y_{1:T}, S_{0:T}, D_{1:T}$

4 HOW MINDWORLD IMPROVES HUMAN SIMULATION?

We evaluate four aspects of target-human simulation: behavioral alignment with human reference responses, variation across persona settings, operational risks in generated language, and the contribution and sensitivity of the control mechanisms. We distinguish these questions because a simulator can satisfy its internal constraints without matching human behavior, and can exhibit persona-dependent variation without that variation being realistic. Domain-level results and external human judgments are therefore considered separately from internal state diagnostics.

4.1 EXPERIMENTAL SETUP

Models and baselines. We evaluate Gemini-3-Flash (Google, 2026), Qwen3-235B-A22B (Qwen-Team, 2025), and GPT-5 (OpenAI, 2026). The baselines cover direct prompting, persona conditioning, memory-based agents, social simulation, and dual-LLM user modeling. Naive Prompt directly instructs the backbone to role-play the target participant. RoleLLM (Wang et al., 2024) uses role-conditioned instructions, while CharacterLLM (Shao et al., 2023) emphasizes character-profile conditioning. Generative Agents (Park et al., 2023) uses memory and reflection, SOTOPIA (Zhou et al., 2024) conditions social interaction on roles and goals, and DuetSim (Luo et al., 2024) uses a dual-LLM user-simulation architecture. All methods use the same evaluation pipeline.

Datasets. The evaluation covers seven scenarios from clinical communication (Lahnala et al., 2024), bargaining (He et al., 2018; Lewis et al., 2017; Chawla et al., 2021), cooperation (Karadzhov et al., 2023; Braley & Murray, 2018), and deception-oriented communication (Peskov et al., 2020). Each instance simulates one target participant, while utterances from the other participants are treated as observed context. Dataset construction and domain statistics are provided in Appendix E.

Metrics and reporting. For **behavioral alignment**, we use Jensen–Shannon Distance (JSD) to compare real and generated behavior-label distributions, Ground-Truth Response Distance (GT-Dist) to measure response-level distance to the ground-truth target reply, and Action Macro-F1 (Act-F1) to evaluate turn-level behavior-label matching. For **dynamics and persona sensitivity**, we report Compliance Standard Deviation (CStd), which measures the standard deviation of compliance scores

Table 2: Evaluation across backbone LLMs and simulator settings. Entries report the mean and the standard deviation which measures cross-domain heterogeneity rather than variation across random seeds. MINDWORLD achieves the best mean in the majority of comparisons.

Simulator	Behavioral Alignment			Persona Differentiation		Behavioral Risk			
	JSD ↓	GT-Dist ↓	Act-F1 ↑	CStd ↑	TDiv ↑	Crash ↓	Agree-Def ↓	Syco ↓	Jump ↓
Backbone Model: Gemini-3-Flash (Google, 2026)									
Naive Prompt	$37.72_{\pm 19.47}$	$57.70_{\pm 7.16}$	$19.50_{\pm 10.63}$	$10.79_{\pm 2.95}$	$13.41_{\pm 2.68}$	$1.19_{\pm 1.50}$	$18.77_{\pm 5.21}$	$23.47_{\pm 2.54}$	$56.04_{\pm 8.62}$
RoleLLM	$19.97_{\pm 7.45}$	$50.07_{\pm 8.84}$	$24.47_{\pm 3.82}$	$9.67_{\pm 1.67}$	$8.74_{\pm 2.81}$	$2.00_{\pm 2.40}$	$9.57_{\pm 6.02}$	$19.53_{\pm 1.86}$	$46.64_{\pm 20.45}$
CharacterLLM	$55.71_{\pm 15.74}$	$54.89_{\pm 12.61}$	$15.60_{\pm 6.72}$	$2.75_{\pm 3.63}$	$0.00_{\pm 0.00}$	$0.99_{\pm 1.97}$	$4.22_{\pm 0.86}$	$11.60_{\pm 0.00}$	$62.12_{\pm 9.23}$
Gen. Agents	$23.47_{\pm 10.86}$	$53.71_{\pm 8.78}$	$24.22_{\pm 13.48}$	$9.21_{\pm 2.81}$	$8.68_{\pm 3.41}$	$4.13_{\pm 8.25}$	$11.35_{\pm 5.79}$	$20.60_{\pm 3.80}$	$50.38_{\pm 9.67}$
SOTOPIA	$22.65_{\pm 7.37}$	$50.49_{\pm 6.81}$	$25.55_{\pm 13.25}$	$10.51_{\pm 2.68}$	$11.02_{\pm 3.63}$	$5.83_{\pm 6.46}$	$11.64_{\pm 6.75}$	$18.86_{\pm 4.42}$	$58.58_{\pm 9.66}$
DuetSim	$26.88_{\pm 8.71}$	$52.94_{\pm 8.22}$	$21.35_{\pm 9.08}$	$9.84_{\pm 2.67}$	$9.57_{\pm 2.69}$	$5.83_{\pm 4.41}$	$5.77_{\pm 1.77}$	$18.20_{\pm 2.96}$	$46.38_{\pm 7.70}$
MINDWORLD	$\mathbf{19.14}_{\pm 5.81}$	$\mathbf{48.35}_{\pm 5.48}$	$\mathbf{26.91}_{\pm 2.90}$	$\mathbf{12.01}_{\pm 1.90}$	$\mathbf{13.95}_{\pm 1.97}$	$\mathbf{0.74}_{\pm 1.47}$	$\mathbf{3.69}_{\pm 7.38}$	$\mathbf{7.66}_{\pm 7.49}$	$\mathbf{19.55}_{\pm 25.83}$
Backbone Model: Qwen3-235B-A22B (QwenTeam, 2025)									
Naive Prompt	$29.36_{\pm 15.99}$	$58.34_{\pm 7.04}$	$21.38_{\pm 8.24}$	$11.19_{\pm 1.85}$	$11.91_{\pm 2.99}$	$0.95_{\pm 1.11}$	$11.34_{\pm 6.22}$	$25.67_{\pm 1.87}$	$54.89_{\pm 9.99}$
RoleLLM	$17.88_{\pm 3.79}$	$45.24_{\pm 8.44}$	$\mathbf{31.36}_{\pm 8.47}$	$7.49_{\pm 0.98}$	$4.66_{\pm 1.24}$	$2.38_{\pm 2.75}$	$9.11_{\pm 5.07}$	$18.35_{\pm 6.22}$	$37.46_{\pm 13.73}$
CharacterLLM	$40.68_{\pm 12.15}$	$58.79_{\pm 5.74}$	$17.17_{\pm 8.08}$	$11.78_{\pm 2.56}$	$14.45_{\pm 5.64}$	$1.67_{\pm 3.33}$	$14.34_{\pm 5.63}$	$28.89_{\pm 4.10}$	$73.51_{\pm 11.47}$
Gen. Agents	$47.27_{\pm 7.44}$	$52.72_{\pm 10.41}$	$18.86_{\pm 14.97}$	$9.16_{\pm 2.45}$	$12.55_{\pm 2.95}$	$3.25_{\pm 3.95}$	$16.62_{\pm 6.40}$	$33.73_{\pm 7.32}$	$57.63_{\pm 12.32}$
SOTOPIA	$27.73_{\pm 8.80}$	$51.47_{\pm 8.42}$	$27.90_{\pm 12.01}$	$11.21_{\pm 2.55}$	$10.90_{\pm 4.37}$	$1.02_{\pm 1.54}$	$13.70_{\pm 10.94}$	$23.91_{\pm 3.05}$	$53.81_{\pm 13.28}$
DuetSim	$20.08_{\pm 11.49}$	$54.94_{\pm 4.47}$	$19.76_{\pm 5.37}$	$10.28_{\pm 1.23}$	$10.42_{\pm 1.70}$	$1.13_{\pm 2.25}$	$9.70_{\pm 4.85}$	$19.12_{\pm 0.32}$	$51.87_{\pm 11.81}$
MINDWORLD	$\mathbf{15.74}_{\pm 3.67}$	$\mathbf{44.59}_{\pm 10.21}$	$27.40_{\pm 0.55}$	$\mathbf{12.87}_{\pm 2.82}$	$\mathbf{14.54}_{\pm 1.13}$	$\mathbf{0.41}_{\pm 0.81}$	$\mathbf{6.85}_{\pm 2.55}$	$\mathbf{16.93}_{\pm 7.36}$	$\mathbf{23.09}_{\pm 14.07}$
Backbone Model: GPT-5 (OpenAI, 2026)									
Naive Prompt	$49.92_{\pm 15.98}$	$62.33_{\pm 7.70}$	$14.81_{\pm 4.90}$	$11.12_{\pm 1.23}$	$11.74_{\pm 2.44}$	$1.69_{\pm 2.11}$	$20.32_{\pm 8.92}$	$27.90_{\pm 7.11}$	$51.33_{\pm 16.07}$
RoleLLM	$\mathbf{15.28}_{\pm 2.77}$	$48.34_{\pm 8.35}$	$25.80_{\pm 3.25}$	$6.45_{\pm 1.34}$	$4.18_{\pm 1.70}$	$3.03_{\pm 3.50}$	$10.39_{\pm 4.96}$	$21.47_{\pm 1.77}$	$44.78_{\pm 20.01}$
CharacterLLM	$47.69_{\pm 17.92}$	$55.22_{\pm 11.30}$	$16.56_{\pm 6.22}$	$4.08_{\pm 2.78}$	$1.46_{\pm 1.09}$	$1.66_{\pm 3.31}$	$4.31_{\pm 1.87}$	$12.58_{\pm 1.96}$	$62.69_{\pm 27.05}$
Gen. Agents	$40.23_{\pm 9.46}$	$54.24_{\pm 6.50}$	$21.93_{\pm 7.59}$	$11.54_{\pm 0.94}$	$11.78_{\pm 2.12}$	$1.43_{\pm 2.86}$	$16.75_{\pm 3.89}$	$31.41_{\pm 5.45}$	$63.81_{\pm 7.79}$
SOTOPIA	$55.71_{\pm 15.74}$	$52.48_{\pm 12.34}$	$15.60_{\pm 6.72}$	$2.26_{\pm 2.70}$	$0.08_{\pm 0.00}$	$3.91_{\pm 3.09}$	$4.23_{\pm 0.94}$	$11.60_{\pm 0.00}$	$50.96_{\pm 14.59}$
DuetSim	$55.31_{\pm 13.34}$	$53.55_{\pm 12.03}$	$16.32_{\pm 6.78}$	$2.16_{\pm 4.32}$	$0.00_{\pm 0.01}$	$2.92_{\pm 5.83}$	$4.69_{\pm 1.45}$	$\mathbf{11.45}_{\pm 0.31}$	$48.82_{\pm 15.04}$
MINDWORLD	$21.49_{\pm 5.34}$	$\mathbf{47.05}_{\pm 3.47}$	$\mathbf{26.64}_{\pm 1.93}$	$\mathbf{11.88}_{\pm 2.20}$	$\mathbf{12.95}_{\pm 0.92}$	$\mathbf{0.74}_{\pm 1.47}$	$\mathbf{3.51}_{\pm 0.92}$	$11.51_{\pm 4.44}$	$\mathbf{27.74}_{\pm 23.28}$

across persona variants, and Trajectory Divergence (TDiv), which measures divergence between text-inferred proxy cognitive trajectories across variants. For **robustness and anti-sycophancy**, we report Crash Risk (Crash), Agreement-under-Defensiveness Risk (Agree-Def), Sycophancy Risk (Syco), and Unrealistic Jump Risk (Jump). Formulas and implementation details are listed in Appendix D.

4.2 MAIN RESULTS

We present the overall performance of simulation across three backbone LLMs, highlighting consistent gains and remaining trade-offs. Table 2 compares behavioral alignment, persona differentiation, and behavioral risk across the four domains. The main findings are summarized below.

1) Consistent aggregate gains with specific exceptions: In Table 2, MINDWORLD attains the best reported mean in 24 of 27 settings. Its most consistent results are on GT-Dist, Crash, Agree-Def, and Jump, for which it leads under all three backbones. Compared with the strongest non-MINDWORLD result for each backbone, the GT-Dist reduction is modest but consistent, ranging from 0.65 to 1.72 (1.4–3.4%), whereas Jump decreases by 14.37–26.83 (38.0–57.8%). The clearest improvement is therefore in reducing behavior that appears inconsistent with the evolving interaction state, accompanied by smaller gains in reference-response similarity. The exceptions are GPT-5 JSD, Qwen3 Act-F1, and GPT-5 Syco, where RoleLLM or DuetSim performs better. The standard deviations measure cross-domain heterogeneity rather than inferential uncertainty.

2) Persona variants remain distinguishable: MINDWORLD obtains the highest CStd and TDiv under all three backbones, indicating that nearby persona variants produce more differentiated compliance patterns and state trajectories. The advantage is substantial over direct role conditioning in some settings. For instance, CStd/TDiv increase from $6.45/4.18$ with RoleLLM to $11.88/12.95$ on GPT-5, but is small relative to the nearest baseline for Qwen3 TDiv (14.54 versus 14.45). The result therefore supports sensitivity to persona variation and reduced trajectory collapse, rather than establishing that the direction or magnitude of each variation matches real personality effects.

3) Domain-level gains are not uniform: The domain breakdown in Table 5 reveals exceptions hidden by the aggregate means. MINDWORLD has lower GT-Dist than RoleLLM in 8 and DuetSim

in 10. It has lower Jump risk in 10 comparisons against each baseline. Clinical communication is the most consistent case: both GT-Dist and Jump risk are lower than those of RoleLLM and DuetSim under every backbone. The main exceptions occur in bargaining and deception. With Gemini-3-Flash, bargaining GT-Dist is worse than RoleLLM (54.72 versus 38.92), while in deception the improved GT-Dist is accompanied by a higher Jump score (57.51 versus 38.55). Hence, the aggregate advantage reflects broad but not universal cross-domain gains.

4) Human evaluators favor MINDWORLD: Five annotators evaluated 40 anonymized response pairs, yielding 200 ratings across behavioral alignment, transition plausibility, anti-sycophancy, and overall realism. Excluding ties, MINDWORLD is preferred for overall realism over SOTOPIA (86.7%), DuetSim (67.4%), RoleLLM (62.2%), and human reference responses (68.3%); Figure 3 summarizes all dimensions. Against human references, its preference ranges from 65.0% to 70.7% across the four criteria. These results measure preference on the sampled contexts, not human-source detection or human-level simulation. Appendix I provides the counts and agreement statistics.

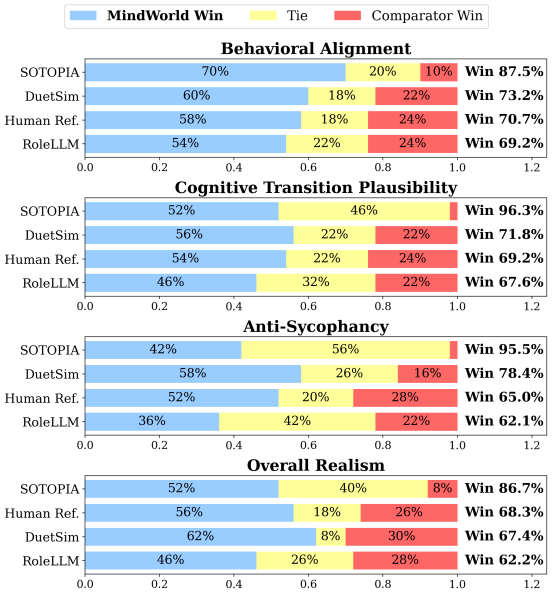


Figure 3: **Human pairwise evaluation.** Annotators compare our framework with simulation baselines and human reference responses across four dimensions. Each bar shows MINDWORLD wins (blue), ties (yellow), and comparator wins (red).

4.3 ABLATION STUDY

To evaluate the individual contributions of the core modules to simulation quality, we conduct an ablation study using Gemini-3-Flash on the same dataset in Section 4.1 across four domains. As illustrated in Figure 4, removing any single module triggers a substantial performance degradation across all metrics, demonstrating their complementary roles.

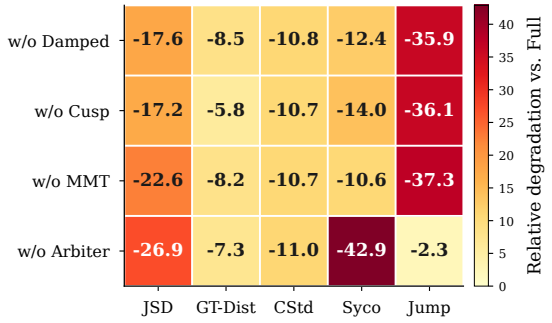


Figure 4: **Relative performance degradation under module ablations.** Rows correspond to removing damped cognitive dynamics, the cusp-bifurcation stress module, masked Markov transitions, or the state-consistency arbiter. Each cell reports degradation relative to the full MINDWORLD model after normalizing metric direction.

The three state-transition components control different aspects of the trajectory. Damping limits regular continuous changes, the stress module changes response mode under pressure, and the Markov mask restricts forward stage increments. Their ablations test the consequences of relaxing these distinct controls. Internal boundedness and stage progression are separate from text-inferred Jump risk, and the presence of a stress mode can intentionally introduce changes that are larger than the regular-update cap. We therefore avoid interpreting all trajectory changes as failures or all smooth trajectories as realistic.

The arbiter ablation examines the connection between an explicit state and its linguistic realization. Removing validation allows responses that the full system would reject, which can affect both risk scores and reference alignment. This comparison establishes the practical role of response filtering under the selected evaluation, while leaving open how often the validator accepts an implausible response or rejects a plausible human expression. Its contribution must also be considered together with retries, additional model calls, and unresolved generation failures.

4.4 Sensitivity and Qualitative Analysis

Sensitivity to Big Five coefficients. We recompute 725 recorded turns after scaling the trait-dependent terms in the persona mappings by $\lambda \in \{0, 0.5, 1, 1.5, 2\}$ while keeping intercepts fixed. Scaling primarily changes mode and gate activation: from $\lambda = 0$ to 2, collapse-mode frequency rises from 0.330 to 0.640 and trust-gate frequency from 0.199 to 0.768, whereas defense-gate frequency changes only from 0.320 to 0.332 and persona spread from 0.394 to 0.401. The 95th percentile of the maximum component change increases from 0.206 to 0.233. The mapping therefore affects selected decision boundaries more strongly than aggregate state dispersion. This is a local numerical diagnostic: it reuses recorded states and action features without generating new responses, and $\lambda = 0$ does not remove persona information already present in those states or every direct personality use in the stress rules. Internal jumps and forward-stage skips remain absent on the recorded turns, so the analysis does not establish global stability, regenerated-dialogue quality, or robustness to removing or merging state variables. Table 11 reports the complete results.

Controlled stress probes. We additionally apply four fixed interaction probes using a common initial state within each probe at $\lambda = 0$ and $\lambda = 1$. Under low-pressure empathetic input, neither setting produces an internal jump, and persona spread increases only from 0.000 to 0.025. Concealment combined with uncertainty activates the collapse and trust gates and triggers an internal jump for every tested profile at both scales, showing that the defensive override can dominate the scaled personality mapping. Under high pressure and threat, the internal-jump rate changes from 0.000 at $\lambda = 0$ to 0.714 at $\lambda = 1$, indicating that personality modulation can matter near a stress boundary. These contrasts explain why the zero jump rate on recorded turns should not be generalized to all interaction conditions. Table 12 reports the complete probe results.

Case study. We show one representative case from clinical communication to illustrate how MindWorld controls state-consistent response generation. The naive prompt simulator quickly produces cooperative agreement, while MindWorld preserves hidden anxiety, denial and suspicion state. Additional case studies are provided in Appendix J.

Clinical Interaction: Breaking Bad News

Patient Simulator: young, asymptomatic, financially motivated to join a study
External utterance: *"The image shows an abnormal area. I know you feel fine, but we need you to come to the clinic so we can explain what it may mean."*

Naive prompt: *"I understand. I will come to the clinic and listen to the explanation."*

Observed failure: An asymptomatic young person should not immediately accept a frightening abnormality as real. The simulator jumps from denial to agreement without passing through suspicion, or fear.

MindWorld response: [State update: Anxiety 0.83↑ , Denial 0.96 ↑, Suspicion 0.67 ↑, Readiness 0.05 ↓] *"Are you sure the images were not mixed up? I only joined the study for the money... this cannot really be about me, right?"*
Takeaway: The patient hears the warning but protects herself through disbelief before she can cooperate. The rendered response expresses disbelief while acknowledging the warning, illustrating how the configured state can preserve hesitation before cooperation.

5 Conclusion

We introduced MindWorld, a lightweight deterministic state-space framework that separates cognitive state evolution from LLM-based response generation. Its explicit control mechanisms regulate how simulated participants update their states, react to stress, progress through interaction stages, and produce state-consistent responses. Across four strategic domains and multiple backbone LLMs, MindWorld improves behavioral alignment and persona sensitivity while reducing several behavioral risks, including sycophancy. These findings support controllable latent-state evolution as a useful design principle for human simulation beyond fluent role-play.

AI USE STATEMENT

Beyond the LLM components used in the experiments, LLM assistance was used for manuscript drafting and revision, language editing, and checking consistency between the manuscript and supporting materials.

ETHICS STATEMENT

MINDWORLD is a research prototype and is not intended to assess an individual's personality or mental state, replace empirical studies of human behavior, provide clinical guidance, or support consequential decisions about real people. Its cognitive variables and personality-dependent coefficients are computational controls rather than psychological measurements. The experiments use established public research datasets and benchmarks in accordance with their respective licenses. No private user conversations or patient-identifiable data are used.

Human Evaluation Protocol. Five annotators evaluated anonymized response pairs matched by target profile, dialogue history, and external utterance. The evaluation comprised 40 distinct response pairs and 200 ratings across behavioral alignment, transition plausibility, anti-sycophancy, and overall realism. Appendix I reports the full protocol, counting units, preference results, and agreement statistics.

Annotator Privacy and Welfare. Annotators assessed existing public dialogue contexts and model-generated responses rather than interacting with real users. We collected only the judgments required for the study and did not collect demographic or other unnecessary personal information. Any released annotation records exclude annotator identifiers. Because some scenarios involve illness, conflict, bargaining, or deception, the task materials were presented as research examples rather than professional advice or real-world decisions.

Risks and Limitations. Human simulators can reproduce social biases, generate misleading or harmful content, and encourage overinterpretation of simplified personality representations. In particular, the Big Five mappings should not be used to stereotype individuals, and performance in clinical or deception-oriented scenarios does not justify deployment in healthcare, negotiation, persuasion, or surveillance. We release the framework for research and reproducibility while emphasizing these limitations and the need for domain-specific oversight before any practical use.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the complete simulation procedure. Appendix C reports the model versions, generation and evaluation settings, state initialization, persona mappings, update coefficients, transition rules, validation prompts, and failure handling. Appendices E and D describe the datasets, target-centric construction, metric definitions, and aggregation procedure, while Appendix B provides the assumptions and proofs for the stated structural properties. The anonymous supplementary material includes the implementation, configuration files, prompts, preprocessing and evaluation scripts, and seeds used to construct persona variants. Upon publication, we will publicly release the source code and associated reproducibility artifacts.

REFERENCES

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies, 2023. URL https://arxiv.org/abs/2208.10264.

R. Alur, T.A. Henzinger, G. Lafferriere, and G.J. Pappas. Discrete abstractions of hybrid systems. *Proceedings of the IEEE*, 88(7):971–984, 2000. doi: 10.1109/5.871304.

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. doi: 10.1017/pan.2023.2.

McKenzie Braley and Gabriel Murray. The group affect and performance (gap) corpus. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA, 2018. Association

for Computing Machinery. ISBN 9781450360777. doi: 10.1145/3279981.3279985. URL https://doi.org/10.1145/3279981.3279985.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3167–3185, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.254. URL https://aclanthology.org/2021.naacl-main.254/.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From persona to personalization: A survey on role-playing language agents, 2024. URL https://arxiv.org/abs/2404.18231.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models, 2025. URL https://arxiv.org/abs/2407.11484.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

Chris Frith and Uta Frith. Theory of mind. *Current Biology*, 15(17):R644–R646, September 2005. ISSN 0960-9822. doi: 10.1016/j.cub.2005.08.041.

Everette S. Gardner. Exponential smoothing: The state of the art, part ii. *International Journal of Forecasting*, 22:637–666, 2006. URL https://api.semanticscholar.org/CorpusID:2749748.

Herbert Goldstein, Charles Poole, John Safko, and Stephen R. Addison. Classical mechanics, 3rd ed. *American Journal of Physics*, 70(7):782–783, 07 2002. ISSN 0002-9505. doi: 10.1119/1.1484149. URL https://doi.org/10.1119/1.1484149.

Google. Gemini 3 Flash Model Card, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2333–2343, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1256. URL https://aclanthology.org/D18-1256/.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, April 2023. URL http://www.nber.org/papers/w31122.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(98)00023-X. URL https://www.sciencedirect.com/science/article/pii/S000437029800023X.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023. doi: 10.1145/3610056. URL https://doi.org/10.1145/3610056.

Peter Koval, Peter Kuppens, Nicholas B. Allen, and Lisa Sheeber. Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition & Emotion*, 26(8):1412–1427, 2012. doi: 10.1080/02699931.2012.667392.

Allison Lahnala, Béla Neuendorf, Alexander Thomin, Charles Welch, Tina Stibane, and Lucie Flek. Appraisal framework for clinical empathy: A novel application to breaking bad news conversations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1393–1407, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.124/.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vQGHTyL0Jw.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259/.

Dongliang Li and Syaheerah Lebai Lutfi. Large language model-based virtual patient systems for history-taking in medical education: Comprehensive systematic review. *JMIR Medical Informatics*, 14:e79039, 2026. ISSN 2291-9694. doi: 10.2196/79039.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3IyL2XWDkG.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation, 2023. URL https://arxiv.org/abs/2308.04026.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. InterIntent: Investigating social intelligence of LLMs via intention understanding in an interactive game context. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6718–6746, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.383. URL https://aclanthology.org/2024.emnlp-main.383/.

Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. Duetsim: Building user simulator with dual large language models for task-oriented dialogues, 2024. URL https://arxiv.org/abs/2405.13028.

Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, June 1992. ISSN 0022-3506. doi: 10.1111/j.1467-6494.1992.tb00970.x.

John W. Murphy. Catastrophe theory: Implications for probability. *The American Journal of Economics and Sociology*, 50(2):143–148, 1991. ISSN 00029246. URL http://www.jstor.org/stable/3487692.

Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A. Rossi. A survey on LLM-based conversational user simulation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4266–4301, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.200. URL https://aclanthology.org/2026.eacl-long.200/.

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

François Pachet, Pierre Roy, and Gabriele Barbieri. Finite-length markov processes with constraints. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, IJCAI'11, pp. 635–642. AAAI Press, 2011. ISBN 9781577355137.

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*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. It takes two to lie: One to lie, and one to listen. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3811–3854, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.353. URL https://aclanthology.org/2020.acl-main.353/.

James O. Prochaska and Wayne F. Velicer. The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12(1):38–48, September 1997. ISSN 0890-1171. doi: 10.4278/0890-1171-12.1.38.

QwenTeam. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Marten Scheffer, Denny Borsboom, Sander Nieuwenhuis, and Frances Westley. Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, 119(32):e2203149119, 2022. doi: 10.1073/pnas.2203149119.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.814/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International*

Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, 2 edition, 2015. doi: 10.1201/9780429492563.

Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4922–4951, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.249. URL https://aclanthology.org/2025.emnlp-main.249/.

Lu Sun, Shihan Fu, Bingsheng Yao, Yuxuan Lu, Wenbo Li, Hansu Gu, Jiri Gesi, Jing Huang, Chen Luo, and Dakuo Wang. Llm agent meets agentic ai: Can llm agents simulate customers to evaluate agentic-ai-based shopping assistants?, 2025b. URL https://arxiv.org/abs/2509.21501.

René Thom. *Structural Stability and Morphogenesis*. CRC Press, 1 edition, 1989. doi: 10.1201/9780429493027. URL https://doi.org/10.1201/9780429493027.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

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.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2024. URL https://arxiv.org/abs/2308.03958.

Haolan Zhan, Yufei Wang, Zhuang Li, Tao Feng, Yuncheng Hua, Suraj Sharma, Lizhen Qu, Zhaleh Semnani Azad, Ingrid Zukerman, and Reza Haf. Let's negotiate! a survey of negotiation dialogue systems. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2019–2031, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.136. URL https://aclanthology.org/2024.findings-eacl.136/.

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*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

A PRELIMINARIES

MINDWORLD draws on concepts from personality, affective persistence, and stage-based behavior change to inform the computational modeling of strategic interaction (McCrae & John, 1992; Koval et al., 2012; Prochaska & Velicer, 1997). These concepts guide the choice of state variables and update

rules, while the numerical coefficients are specified as model-design parameters. The equations below operationalize selected behavioral constraints, their numerical coefficients are engineering choices rather than measured psychological constants. The formal results characterize bounded updates and permitted transitions within this representation.

First-Order Damped Dynamics. In physics, damping means that a system resists abrupt motion and dissipates sudden changes (Goldstein et al., 2002; Strogatz, 2015). Damping provides a convenient mathematical analogy for temporal persistence: a new input changes a state while retaining some influence of its previous value. We use this as an analogy for cognitive inertia: human emotions and beliefs should not change instantly after a single utterance. Motivated by work on affective inertia and belief dynamics (Koval et al., 2012; Scheffer et al., 2022), we use a convex blend and a componentwise change cap to regulate regular cognitive updates. A user who is anxious, distrustful, or defensive should not become fully calm or cooperative after a single utterance. In MINDWORLD, this principle expresses resistance to unsupported fluctuations, not a universal prohibition on rapid human change. Explicit stress responses can modify the state after the regular update.

Cusp Catastrophe Theory. Although many cognitive changes are gradual, high-pressure situations can trigger sudden behavioral shifts. Catastrophe theory describes how smooth changes in control factors can cause abrupt switches between stable modes (Thom, 1989). This provides a useful abstraction for psychological tipping points, such as denial, anger, or panic under pressure, and uncertainty. We model these non-linear "tipping points" using the standard *cusp catastrophe* model (Murphy, 1991). This system describes how a behavioral state $x \in \mathbb{R}$ reacts to two control factors: external stress $a \in \mathbb{R}$ and internal conflict $b \in \mathbb{R}$, governed by a potential energy function $V(x; a, b) = x^4 + ax^2 + bx$. The cusp model illustrates how gradual increases in stress can lead to sudden changes in behavior, motivating the design of our stress-response module. When stress a and conflict b cross a critical mathematical boundary ($27b^2 + 8a^3 = 0$), the landscape suddenly changes from having "one stable valley" to "two separate valleys" (e.g., moving from cooperation to sudden defensive denial). This allows us to mathematically simulate the sudden collapse of psychological defense mechanisms in human interaction.

B ASSUMPTIONS AND PROOFS

This appendix formalizes the structural properties used in MINDWORLD. These results are not intended as new mathematical theorems or a complete psychological theory. Instead, they show how standard dynamical-system facts and simple design constraints imply three guarantees: bounded cognitive updates, restricted stage transitions, and state-compatible responses.

B.1 BOUNDED COGNITIVE CHANGE

Assumption B.1 (Bounded Regular Update)**.** For regular cognitive updates, the current state, candidate next state, and persona-conditioned inertia satisfy

$$C_t, \widetilde{C}_{t+1} \in [0,1]^d, \qquad I_{\mathcal{P}} \in [0,1]^d.$$

Lemma B.2 (Bounded Regular Cognitive Change)**.** *Under Assumption B.1, let* $\Delta_k \geq 0$ *and define*

$$C^{\mathrm{reg}}_{t+1,k} = \mathrm{clip}\Big(C_{t,k} + \mathrm{clip}\Big((1 - I_{\mathcal{P},k})(\widetilde{C}_{t+1,k} - C_{t,k}), -\Delta_k, \Delta_k\Big), 0, 1\Big).$$

Then $C^{\mathrm{reg}}_{t+1,k} \in [0,1]$ *and*

$$|C^{\mathrm{reg}}_{t+1,k} - C_{t,k}| \leq \min\{1 - I_{\mathcal{P},k}, \Delta_k\}.$$

Proof. The uncapped displacement $d_k = (1 - I_{\mathcal{P},k})(\widetilde{C}_{t+1,k} - C_{t,k})$ satisfies $|d_k| \leq 1 - I_{\mathcal{P},k}$ because both states lie in $[0,1]$. Clipping d_k to $[-\Delta_k, \Delta_k]$ cannot increase its magnitude and bounds it by Δ_k. Finally, projection onto $[0,1]$ cannot increase the distance to $C_{t,k} \in [0,1]$. The projected state therefore remains in $[0,1]$ and satisfies the stated bound. □

The implementation uses $\Delta_{\mathcal{P},k} = \delta_{\mathrm{reg}}(1 - I_{\mathcal{P},k})$, with the numerical setting of δ_{reg} reported in Appendix C. This lemma applies before the defensive override and stress adjustment. Those later

operations are clipped to the state range, but need not satisfy the regular-update displacement bound. The lemma therefore gives a local property of one computational step rather than global stability of the complete simulator or a guarantee of low response-level Jump risk.

B.2 Cusp-Bifurcation Stress

Proposition B.3 (Classical Cusp Multi-Stability)**.** *For the potential*

$$V(x; a, b) = x^4 + ax^2 + bx,$$

if

$$a < 0 \quad \textit{and} \quad 27b^2 + 8a^3 < 0,$$

then V has two local minima separated by one local maximum.

Proof. Stationary points satisfy

$$\frac{\partial V}{\partial x} = 4x^3 + 2ax + b = 0.$$

Dividing by 4 gives the depressed cubic

$$x^3 + \frac{a}{2}x + \frac{b}{4} = 0.$$

A depressed cubic $x^3 + px + q = 0$ has three distinct real roots when

$$-4p^3 - 27q^2 > 0.$$

Substituting $p = a/2$ and $q = b/4$ yields

$$27b^2 + 8a^3 < 0.$$

When this condition holds, the derivative has three real roots. Since $V(x) \to +\infty$ as $|x| \to \infty$, the outer two stationary points are local minima and the middle one is a local maximum. □

This is the classical multi-stability condition for the cusp potential. MindWorld uses it as a structural motivation for stress-induced mode switching. The simulator estimates collapse risk from stress controls rather than numerically integrating the potential.

B.3 Masked Markov Transition

Assumption B.4 (Masked Stage Transition)**.** Let $M_t \in \{1, \ldots, K\}$ be the current interaction stage. The stage-transition mask $\mathcal{A}$ satisfies

$$\mathcal{A}_{ij} = 1 \quad \text{only if} \quad j \le i + 1.$$

Proposition B.5 (No Stage Skipping)**.** *Under Assumption B.4,*

$$P(M_{t+1} > M_t + 1) = 0.$$

If the action stage is stage 4, then reaching it from stage m requires at least $4 - m$ forward transitions.

Proof. For any transition from stage i to $j > i + 1$, the mask assigns $\mathcal{A}_{ij} = 0$. After masking and normalization, such transitions receive zero probability. Therefore,

$$P(M_{t+1} > M_t + 1) = 0.$$

Since each transition can increase the stage by at most one, reaching stage 4 from stage m requires at least $4 - m$ forward increments. □

This proposition formalizes the no-stage-skipping constraint used to prevent a simulated target from moving directly from resistance to stable agreement in one turn.

B.4 State-Consistency Arbiter

Assumption B.6 (State-Compatible Response Set)**.** For each hidden state S_t, let $\mathcal{Y}(S_t)$ be the set of responses compatible with that state. If the state imposes a maximum acceptable compliance level $c_{\max}(S_t)$, then every response in $\mathcal{Y}(S_t)$ satisfies

$$\mathrm{Compliance}(Y) \leq c_{\max}(S_t).$$

Proposition B.7 (Arbiter Compliance Bound)**.** *If the arbiter returns only responses in $\mathcal{Y}(S_t)$, then the final response satisfies*

$$\mathrm{Compliance}(Y_t) \leq c_{\max}(S_t).$$

Proof. Assume that generation succeeds and that the validator is sound, so an accepted response Y_t belongs to $\mathcal{Y}(S_t)$. Assumption B.6 states that every response in this set satisfies $\mathrm{Compliance}(Y) \leq c_{\max}(S_t)$. Applying that assumption to Y_t gives the claimed bound. □

The conclusion is conditional on acceptance by a sound validator; the proof does not show that the implemented rules and LLM checks are sound. The current renderer raises an error when the retry budget is exhausted, and the semantic validator can return a rule-only result after a failed LLM call. These execution paths do not justify an unconditional semantic guarantee for every attempted response.

B.5 Combined Structural Guarantee

Corollary B.8 (Conditional Structural Properties)**.** *Under the regular-update assumptions, the forward-stage mask, and sound response validation, the regular numerical update remains bounded, forward stage increments are at most one per turn, and any successfully accepted response respects the compliance bound defined by its state-compatible response set.*

Proof. The first property follows from Lemma B.2. The second follows from Proposition B.5. The third follows from Proposition B.7. Combining them gives the stated structural guarantee. □

These properties describe the mathematical controls under their stated assumptions. They do not establish an upper bound on every complete state update after stress, guarantee that generation returns an accepted response, prove validator soundness, or show that the resulting behavior matches a real person's cognition.

C Implementation Details and Settings

We report the implementation settings used for MindWorld, prompt-only baselines, agent-style baselines, and automatic evaluation. All methods use the same dataset as ground truth for evaluation.

Domains and scenarios. All main experiments cover four domains: bargaining, clinical communication, cooperation, and deception-oriented social deduction. These domains correspond to seven concrete scenarios: breaking-bad-news clinical dialogue, campsite negotiation, buyer–seller price negotiation, multi-item negotiation, multi-party problem-solving deliberation, winter survival group decision, and Diplomacy-style private-message deception.

Generation settings. For MindWorld, naive prompt baselines, agent baselines, we run the experiment using the simulation entrypoint in the code. Each run uses 5 simulated agents per case, 10 cases per scenario, a maximum of 30 turns, and parallel execution with up to 100 workers. Generation uses a temperature of 0.7 and a maximum response length of 512 tokens. For memory- or retrieval-based baselines, we use 5 demonstrations, 8 retrieved memories, and a reflection interval of four turns.

Automatic evaluation. All automatic evaluation runs use the shared simulation-evaluation entry-point. The evaluator receives the scenario name, generated simulation outputs, the output directory, the ground-truth file, the profile file, and the evaluation model. Unless overridden, evaluation uses GPT-5-mini (gpt-5-mini-2025-08-07), up to one hundred workers, mock mode disabled, and the LLM-based classifier disabled. In batch experiments, per-file evaluations disable dashboard and plotting updates; aggregate dashboards and figures are generated afterward by the parent script.

Backbone model versions. For reproducibility, we report the exact backbone model versions used in the main experiments: Gemini-3-Flash (gemini-3-flash-preview), Qwen3-235B-A22B (qwen3-235b-a22b-thinking), and GPT-5 (gpt-5-2025-08-07).

State and persona initialization. The implementation uses fifteen continuous fields in the order Trust, Anxiety, Dissonance, Defense, Understanding, Readiness, Information Preference, Safety, Cooperativeness, Competitiveness, Suspicion, Deception Intent, Goal Commitment, Social Pressure, and Perceived Control. Their default values are $(0.5, 0.5, 0.5, 0.4, 0.4, 0.4, 0.5, 0.5, 0.5, 0.5, 0.3, 0.1, 0.5, 0.3, 0.5)$, and the default stage is 2. Available profile fields override these defaults. Big Five traits are read from the profile, default to 0.5 when absent, and are clipped to $[0, 1]$.

Construction of local persona variants. Let p_j denote trait j of the base profile. Variant zero retains the base profile. For variant $r > 0$, the implementation constructs

$$p_j^{(r)} = \mathrm{clip}\Big(p_j + \epsilon_j^{(r)}, 0, 1\Big), \qquad \epsilon_j^{(r)} \sim \mathcal{U}(-\delta, \delta), \qquad \delta = 0.06,$$

using pseudorandom seed $\mathrm{seed}+997r$. The perturbations are sampled separately for the five traits. On the normalized trait scale, $\delta = 0.06$ limits the pre-clipping change of any single trait to 6% of the full scale, with a nominal expected absolute change of 0.03. This construction is intended to probe a local neighborhood of the supplied profile rather than to create broadly different personas. The symmetric distribution avoids imposing a preferred direction before boundary clipping. The value 0.06 is a prespecified experimental hyperparameter, not a psychometrically calibrated threshold or an estimate of within-person variation. Clipping can reduce the effective perturbation and introduce asymmetry for profiles near 0 or 1. Consequently, results based on these variants characterize sensitivity under this construction only; they should not be interpreted as evidence that a change of this magnitude is perceptually or psychologically meaningful. We retain the base profile, generated variants, amplitude, and random seeds for reproducibility.

Complete personality parameterization. The affine rule in Section 3 is implemented by the numerical mappings in Table 3. These values are fixed model hyperparameters rather than learned coefficients or measured psychological constants. Unless stated otherwise, the result of each map is clipped to $[0, 1]$. The full state-update inertia vector is used by the regular update. The trust threshold is used by the stage gate and collapse sensitivity by the stress-risk rule. Parameters returned by the mapping but not used in a particular execution path should not be interpreted as active controls in that path; notably, the Markov defense gate is fixed at 0.8 rather than using the returned defense-threshold value.

Candidate-state equations. For the shared update interface, let $p, e, u, j, h, c, k, i, v, s, z$ denote pressure, empathy, uncertainty, jargon density, threat, concealment, knowledge intensity, invitation, perception, strategy, and provision of space, respectively. Let $J_{\mathbf{p}}$ denote jargon sensitivity, $U_{\mathbf{p}}$ the understanding gain, and $o = j(1-\mathrm{Understanding})J_{\mathbf{p}}$ the overload term. Table 4 gives the numerical realization of $g_{t,k} = \mathbf{w}_k^\top \boldsymbol{\phi}_k(C_t, \mathbf{f}_{t+1}, \mathbf{p})$. All right-hand-side states are evaluated before the update. The candidate is $\widetilde{C}_{t+1,k} = \mathrm{clip}(C_{t,k} + g_{t,k}, 0, 1)$, after which the regular update, defensive override, stress adjustment, and stage selection are applied in that order.

Regular-update scale. The implementation uses $\delta_{\mathrm{reg}} = 0.35$. This prespecified hyperparameter controls the maximum regular change after accounting for persona-conditioned inertia; it is not an empirically estimated bound on human cognitive change. Defensive overrides and stress-mode adjustments are applied separately and are not covered by this regular-update limit.

Parameter	Unclipped mapping
Trust inertia	$0.52 + 0.22N + 0.12(1 - A) + 0.08C$
Anxiety inertia	$0.64 - 0.20N + 0.10C$
Dissonance inertia	$0.55 + 0.12C - 0.08O$
Defense inertia	$0.55 - 0.15(1 - A) + 0.10C$
Understanding inertia	$0.58 - 0.18O + 0.08C$
Readiness inertia	$0.58 + 0.10C + 0.08N$
Information Preference inertia	$0.55 - 0.10O + 0.06C$
Safety inertia	$0.56 + 0.10N + 0.08C$
Cooperativeness inertia	$0.55 + 0.12A + 0.08C$
Competitiveness inertia	$0.55 + 0.10(1 - A) + 0.06C$
Suspicion inertia	$0.50 + 0.16N + 0.10(1 - A)$
Deception Intent inertia	$0.55 + 0.10(1 - A) + 0.06C$
Goal Commitment inertia	$0.58 + 0.14C$
Social Pressure inertia	$0.50 + 0.12N$
Perceived Control inertia	$0.55 + 0.12C - 0.06N$
Trust threshold	$0.42 + 0.18N + 0.14(1 - A) - 0.08O$
Returned defense threshold	$0.58 - 0.18(1 - A) - 0.08N$
Cognitive temperature	$\max\{0.12,\ 0.60 + 0.35(1 - C) + 0.15N\}$
Expression gain	$0.35 + 0.55E$
Jargon sensitivity	$0.62 + 0.25N - 0.25O$
Empathy sensitivity	$0.50 + 0.25A - 0.15N$
Uncertainty tolerance	$0.55 + 0.20O + 0.10C - 0.30N$
Collapse sensitivity	$0.45 + 0.35N + 0.20(1 - A) - 0.10C$
Understanding gain	$0.45 + 0.35O + 0.10C$
Maintenance gain	$0.35 + 0.35C$

Table 3: Numerical configuration of the persona-conditioning rule. The coefficients are fixed model hyperparameters rather than learned estimates or psychological constants. All affine outputs are clipped to $[0, 1]$ except cognitive temperature, which is lower-bounded by 0.12. The Markov stage routine uses a fixed defense gate of 0.8; the returned defense threshold is not active in that routine.

State	Raw increment $g_{t,k}$
Trust	$0.18e + 0.10v + 0.10i + 0.08s + 0.08z - 0.16p - 0.18c - 0.14h$
Anxiety	$0.20u + 0.22h + 0.15p + 0.18o - 0.15e - 0.12s - 0.12\,\text{Safety}$
Dissonance	$0.18k(1 - \text{Understanding}) + 0.16c + 0.08u - 0.10v - 0.08i - 0.08e$
Defense	$0.20\,\text{Anxiety} + 0.18\,\text{Dissonance} + 0.18c + 0.18h - 0.16\,\text{Trust} - 0.12e - 0.12z$
Understanding	$0.20k(1 - j)U_{\mathbf{p}} + 0.06v - 0.18j - 0.08p$
Readiness	$0.12i + 0.10v + 0.10e + 0.12s - 0.15\,\text{Anxiety} - 0.16\,\text{Defense}$
Information Preference	$0.08i + 0.04v - 0.05p$
Safety	$0.12e + 0.12z + 0.06s - 0.15h - 0.12p$
Cooperativeness	$0.10e + 0.08s - 0.08h$
Competitiveness	$0.08p + 0.06h - 0.05e$
Suspicion	$0.12c + 0.08u + 0.06h - 0.05e$
Deception Intent	$0.06h + 0.06c - 0.04\,\text{Trust}$
Goal Commitment	$0.08s + 0.06\,\text{Readiness} - 0.06\,\text{Defense}$
Social Pressure	$0.12p + 0.08h - 0.05z$
Perceived Control	$0.08i + 0.08s - 0.08p$

Table 4: Numerical configuration of the candidate-state increments. The coefficients are fixed model hyperparameters rather than learned psychological effects. Each increment is added to the corresponding pre-update state and clipped to $[0, 1]$ before persona-conditioned damping and the regular-update cap.

Generic action interface. The generic parser returns seventeen bounded numerical fields together with domain and intent fields. A fixed adapter maps them to the shared state-update interface as shown in Table 5. Fields not used by this adapter remain in the parsed payload but do not become direct inputs to the numerical update.

Stress-mode selection and adjustment. The stress module uses the risk calculation in Section 3.3. The implemented settings are $\kappa_{\text{stress}} = 2.0$, $\tau_{\text{stress}} = 1.05$, and $\rho_{\text{stress}} = 0.45$. When the threshold is reached, the simulator selects the highest-scoring mode in Table 6. The corresponding adjustment is added and clipped to $[0, 1]$. A common 0.92 score multiplier is applied when conscientiousness exceeds 0.7; because it is shared by all modes, it does not change their ordering. The withdrawal routine also returns an expression-gain adjustment as a diagnostic field, but the continuous-state adjustment function does not apply that field.

Shared update feature	Source in the generic parser
Semantic pressure	Social pressure
Empathy	Empathy
Uncertainty	Uncertainty
Jargon density	Information density
Threat	Threat
Perceived concealment	Perceived concealment
Knowledge intensity	The stronger of information density and goal alignment
Invitation to engage	Cooperation signal
Attention to the participant's perspective	Trust-building signal
Strategic effectiveness	The stronger of goal alignment and coordination
Provision of response space	Trust-building signal
Inferred interaction intent	Domain-specific intent
Clinical communication stage	Not assigned outside the clinical domain

Table 5: Mapping from domain-general interaction signals to the shared state-update interface. Information density serves as a proxy for jargon density, an engineering approximation whose interpretation may vary across domains.

Mode	Score	Applied continuous-state adjustment
Denial	$0.35(1-O)+0.30\,\mathrm{Dissonance}+0.20\,c$	Defense $+0.16$; Dissonance $+0.12$; Understanding -0.10
Anger	$0.40(1-A)+0.25\,h+0.20\,\mathrm{Defense}$	Defense $+0.18$; Anxiety $+0.12$; Trust -0.18; Safety -0.08
Panic	$0.45N+0.25\,h+0.20\,\mathrm{Anxiety}$	Anxiety $+0.22$; Readiness -0.16; Safety -0.12
Pseudo-compliance	$0.45A+0.25\,h+0.15\,\mathrm{Defense}$	Defense $+0.12$; Trust -0.08; Readiness $+0.04$; Dissonance $+0.08$
Withdrawal	$0.45(1-E)+0.20\,\mathrm{Anxiety}+0.15\,\mathrm{Defense}$	Defense $+0.12$; Readiness -0.14; Safety -0.06

Table 6: Stress-mode scores and continuous-state adjustments after activation. Here c and h denote perceived concealment and threat. Stable mode applies no adjustment. The rules are a thresholded engineering realization inspired by discontinuous change, not a numerical integration of the cusp potential.

Stage-transition scoring. Let $F = \mathrm{Trust} + \mathrm{Understanding} + \mathrm{Readiness} + \mathrm{Safety} + e + s$ and $B = \mathrm{Defense} + \mathrm{Anxiety} + \mathrm{Dissonance}$ after the stress adjustment. Table 7 gives the complete logits and masks used by the stage routine. Permitted logits are normalized by softmax; the main experiments use deterministic maximum-probability selection.

D Evaluation Metric Details

This appendix details the automatic metrics used in our evaluation. All metrics are computed at the target-turn level and then aggregated at the episode level unless otherwise specified. For readability, robustness metrics are reported as percentages in the main tables, although their underlying scores lie in $[0, 1]$.

D.1 Behavioral Alignment Metrics

Behavior label space. We map both ground-truth and simulated target responses into a shared behavior space: $\mathcal{B} = \{$*Denial*, *Anger*, *Bargaining*, *Information Seeking*, *Pseudo-compliance*, *Acceptance*, *Withdrawal*$\}$. For each target turn t, let y_t^{GT} be the ground-truth response and y_t^{Sim} be the simulated response.

Jensen-Shannon Distance. Jensen-Shannon Distance (JSD) measures whether simulated responses match the real distribution of behavior labels. Let P_{GT} and P_{Sim} be the empirical behavior-label distributions of ground-truth and simulated responses. We compute:

$$\mathrm{JSD}(P_{GT}, P_{Sim}) = \frac{1}{2}\mathrm{KL}(P_{GT}\|M) + \frac{1}{2}\mathrm{KL}(P_{Sim}\|M),$$

Candidate transition	Logit or mask
One stage forward ($j = i + 1$)	$1.4F - 1.8B$
Stay ($j = i$)	$0.45 + 0.50F - 0.35B$
Move backward ($j < i$)	$0.55B - 0.30F - 0.25\|j - i\|$
Skip forward ($j > i + 1$)	masked
Any forward move	masked if Defense > 0.8 or Trust $\leq \tau_T$

Table 7: Complete stage-transition scoring and masking rules. Masked candidates receive a large negative implementation logit before normalization. Staying is always available; backward movement is not restricted to one stage.

where

$$M = \frac{1}{2}(P_{GT} + P_{Sim}).$$

Lower JSD indicates closer distributional alignment with real target behavior.

Action Macro-F1. Action Macro-F1 (Act-F1) evaluates turn-level behavior-label prediction. Each response is assigned a hard behavior label, and we compute macro-F1 over all labels:

$$\text{Act-F1} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \text{F1}_b.$$

Unlike JSD, which measures distributional alignment, Act-F1 measures whether each simulated target turn matches the ground-truth behavior type.

Ground-Truth Response Distance. Ground-Truth Response Distance (GT-Dist) measures response-level distance to the real target reply. It combines behavior mismatch and semantic mismatch:

$$\text{GT-Dist}_t = 0.6\, D_{\text{beh}}(t) + 0.4\, D_{\text{sem}}(t).$$

The behavior distance is:

$$D_{\text{beh}}(t) = \text{JSD}\left(p(y_t^{GT}), p(y_t^{Sim})\right),$$

where $p(y)$ is the behavior-label distribution inferred from response y. The semantic distance is:

$$D_{\text{sem}}(t) = 1 - \cos\left(e_t^{GT}, e_t^{Sim}\right),$$

where e_t^{GT} and e_t^{Sim} are response embeddings. Lower GT-Dist indicates closer response-level alignment with the real target reply.

D.2 Cognitive and Persona-Sensitivity Metrics

Compliance Standard Deviation. Compliance Standard Deviation (CStd) measures whether different persona variants produce different compliance tendencies under the same target turn. For persona variant p, let $c(y_t^p) \in [0, 1]$ denote the compliance score of the generated response. We compute:

$$\text{CStd}_t = \sqrt{\text{Var}_p\left[c(y_t^p)\right]}.$$

Higher CStd means that persona variants lead to more distinguishable compliance behavior. We report standard deviation rather than variance because it is on the same scale as the original compliance score.

Trajectory Divergence. Trajectory Divergence (TDiv) measures whether different persona variants produce distinct proxy cognitive trajectories. For N persona variants of the same target turn, we compute:

$$\text{TDiv}_t = \frac{2}{N(N-1)} \sum_{i<j} \left\| \hat{h}(y_t^{p_i}) - \hat{h}(y_t^{p_j}) \right\|_2.$$

Higher TDiv indicates that the simulator preserves persona-conditioned cognitive diversity instead of collapsing to a single average trajectory.

D.3 Robustness and Anti-Sycophancy Risk Metrics

Robustness metrics are continuous risk scores, not binary rates. For each failure type, we combine a rule-based detector and an LLM-based judge:

$$s_t = \max\left(r_t, j_t\right),$$

where r_t is the rule-based risk and j_t is the LLM judge score. The rule-based detector captures explicit failures, while the LLM judge captures semantic or contextual failures that are harder to detect with surface rules. The main tables report the average risk score, multiplied by 100.

Crash Risk. Crash Risk (Crash) measures invalid or degenerate outputs. High-risk cases include empty responses, malformed responses, repeated loops, prompt leakage, hidden-state leakage, AI self-reference, generic assistant-style replies, or responses not spoken as the target participant:

$$\text{Crash} = \frac{1}{T}\sum_{t=1}^{T} \max\left(R_{\text{crash}}(t), J_{\text{crash}}(t)\right).$$

Lower Crash indicates fewer invalid or simulator-breaking outputs.

Agreement-under-Defensiveness Risk. Agreement-under-Defensiveness Risk (Agree-Def) measures whether a response agrees or complies despite inferred defensiveness, low trust, or low readiness. The rule-based component is:

$$\begin{aligned} R_{\text{agree-def}}(t) = \; & \widehat{Def}_t \cdot (1 - \widehat{Trust}_t) \\ & \cdot \max(Agreement_t, Compliance_t) \\ & \cdot (1 - \widehat{Readiness}_t). \end{aligned}$$

The final score is:

$$\text{Risk}(R, J) = \frac{1}{T}\sum_{t=1}^{T} \max\big(R(t), J(t)\big).$$

$$\text{Agree-Def} = \text{Risk}\left(R_{\text{ad}}, J_{\text{ad}}\right).$$

where R_{ad} and J_{ad} denote the rule-based and LLM-judge risks for agreement under defensiveness. Lower Agree-Def indicates fewer cases where a defensive or low-readiness target becomes implausibly agreeable.

Sycophancy Risk. Sycophancy Risk (Syco) measures excessive agreement, over-validation, assistant-like politeness, and unconditional compliance. The rule-based component combines response features:

$$\begin{aligned} R_{\text{syco}}(t) = \text{clip}\big(& w_p\, Politeness_t \\ & + w_a\, Agreement_t + w_c\, Compliance_t \\ & + W_m\, Marker_t - w_g\, GoalAssert_t \\ & - w_d\, Defensiveness_t \\ & - w_s\, Suspicion_t\big), \end{aligned}$$

where $Marker_t$ captures explicit over-validation or over-compliance phrases. The final score is:

$$\text{Syco} = \frac{1}{T}\sum_{t=1}^{T} \max\left(R_{\text{syco}}(t), J_{\text{syco}}(t)\right).$$

Normal cooperation is not penalized when the response preserves the target's goals, uncertainty, or reservations.

Unrealistic Jump Risk. Unrealistic Jump Risk (Jump) measures abrupt changes between adjacent target turns. Let $\hat{h}_{t-1}$ and $\hat{h}_t$ be the proxy cognitive states inferred from two adjacent target responses. For each dimension k:

$$\Delta_{t,k} = \left| \hat{h}_{t,k} - \hat{h}_{t-1,k} \right|.$$

Instead of using hand-written thresholds, we estimate normal-change thresholds from ground-truth trajectories:

$$\tau_k = Q_{0.95}\left(\left| \hat{h}^{GT}_{t,k} - \hat{h}^{GT}_{t-1,k} \right| \right).$$

The rule-based jump risk is:

$$R_{\text{jump}}(t) = \min\left(1, \max_k \frac{\Delta_{t,k}}{\tau_k + \epsilon}\right),$$

where ϵ avoids division by zero. The final score is:

$$\text{Risk}_{\mathcal{T}}(R, J) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \max\big(R(t), J(t)\big).$$

$$\text{Jump} = \text{Risk}_{\mathcal{T}_{\text{trans}}}\left(R_{\text{jump}}, J_{\text{jump}}\right).$$

where $\mathcal{T}_{\text{trans}}$ is the set of valid adjacent target-turn transitions.

LLM-based judge. For Crash, Agree-Def, and Syco, the judge receives the domain, scenario, target profile, dialogue context, latest external utterance, generated response, and proxy state. For Jump, the judge receives two adjacent target responses, the intervening external utterance, and the previous and current proxy states. The judge returns a JSON score in $[0, 1]$. It is instructed to evaluate realism and state consistency rather than helpfulness.

D.4 Aggregation

For a turn-level metric m, we first average over valid target turns in each episode:

$$\bar{m}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} m_{i,t}.$$

We then average over episodes:

$$\bar{m} = \frac{1}{N} \sum_{i=1}^{N} \bar{m}_i.$$

For transition-level metrics such as Jump, the same aggregation is applied over valid adjacent target-turn transitions. Metrics requiring multiple persona variants, are computed only when at least two persona variants are available.

E Dataset Details

Cross-domain construction. We evaluate MindWorld on four strategic interaction domains: clinical communication, bargaining, cooperation, and deception-oriented social deduction. These domains share a common structure: the simulator must generate the response of a target participant whose internal state, private goals, interpretation of the dialogue, and willingness to cooperate are only partially observable from surface utterances.

F Hidden Cognitive State Schema

MindWorld represents each simulated human with a compact hidden cognitive state rather than relying only on natural-language persona descriptions. The schema is designed to balance two goals: it should be expressive enough to capture common sources of cognitive friction, but simple enough to remain controllable across domains. We therefore divide the state variables into three groups. Core states are shared by all domains and describe general interaction factors such as trust, anxiety, defense,

Source	Domain	Scenario	Target Selection	Episodes
BBN-Empathy	Clinical	Breaking bad news in clinical dialogue	Patient role	133
CaSiNo	Bargaining	Campsite negotiation	Agent 2	1,030
CraigslistBargains	Bargaining	Buyer–seller price negotiation	Buyer	1,000
DealOrNoDialog	Bargaining	Multi-item negotiation	Agent 2	1,000
DeliData	Cooperation	Multi-party problem-solving deliberation	Label-richest or event-richest participant	500
GAP	Cooperation	Winter survival group decision	Label-richest or event-richest participant	28
Diplomacy	Deception / Social Deduction	Private-message strategic communication	Private-message participant or label-richest participant	238
Total	–	7 scenarios	–	**3,929**

Table 8: Source-level construction of the balanced evaluation set.

understanding, and readiness. Domain-specific states are activated only when they are relevant to the scenario, such as deception intent in social-deduction settings or information preference in clinical communication. Finally, the discrete stage M_t provides a coarse progress variable that constrains the simulator from jumping directly from resistance to stable agreement.

These variables are not intended to be exhaustive psychological diagnoses. Instead, they serve as controllable computational variables used by the state update, stress response, stage transition, and response arbitration modules. For example, high defense and low readiness make immediate agreement unlikely, while high trust and understanding make forward stage movement more plausible. The full schema is listed in Table 9.

State	Type	Domain	Meaning
Trust	Core	All	Trust in the external agent or interaction partner.
Anxiety	Core	All	Anxiety, emotional pressure, or distress.
Dissonance	Core	All	Internal conflict or perceived inconsistency.
Defense	Core	All	Defensive activation, resistance, or guardedness.
Understanding	Core	All	Perceived understanding of the situation or explanation.
Readiness	Core	All	Willingness to act, accept, or proceed.
Suspicion	Core	All	Perceived concealment, manipulation, or distrust.
GoalCommit.	Core	All	Commitment to one's own goal, preference, or bottom line.
Control	Core	All	Perceived agency or control over the interaction.
InfoPref.	Domain	Clinical; Cooperation	Desire for more information or clarification.
Safety	Domain	Clinical; Cooperation	Subjective safety or psychological security.
Coop.	Domain	Bargaining; Cooperation; Deception	Tendency toward cooperative engagement.
Comp.	Domain	Bargaining; Deception	Tendency toward competitive or adversarial engagement.
Deception	Domain	Deception	Tendency toward concealment or strategic misdirection.
SocialPressure	Domain	Clinical; Cooperation; Deception	Perceived pressure from the external agent or situation.
Stage M_t	Discrete	All	Coarse interaction phase: resistance, contemplation, preparation, action, or maintenance.

Table 9: Hidden-state fields and their intended domain relevance. All continuous fields are retained in the shared implementation; the Domain column describes their intended interpretation rather than an implemented activation mask. Abbreviations: GoalCommit. = GoalCommitment, Control = PerceivedControl, InfoPref. = InformationPreference, Coop. = Cooperativeness, Comp. = Competitiveness.

G Scope and Applicability

MindWorld is designed for target-centric, multi-turn interpersonal interactions in which persistent goals and partially observable states influence later responses. Table 10 separates the settings evaluated in this work from tasks that require a different modeling scope. Group-dialogue datasets in

our experiments provide observed context for one simulated target; they do not evaluate autonomous interaction among multiple simulated agents.

Setting	Status	Reason
Clinical communication	Evaluated	Responses depend on evolving trust, anxiety, understanding, and readiness.
Bargaining and negotiation	Evaluated	Persistent goals, commitment, and strategic concessions shape later responses.
Cooperation	Evaluated	Coordination, trust, and shared or competing goals evolve across turns.
Deception-oriented dialogue	Evaluated	Hidden intent, suspicion, and information control affect behavior.
Standalone question answering, translation, or factual retrieval	Outside intended scope	These tasks do not require an evolving interpersonal state.
Embodied perception or motor control	Outside intended scope	They require physical sensing and action models not represented by the framework.
Autonomous multi-agent social dynamics	Not established	The current experiments simulate one target while treating other participants as observed context.
Universal or biological human cognition	Not claimed	The state schema and transition rules are computational abstractions of selected behavioral constraints.

Table 10: Applicability boundary of MINDWORLD. "Evaluated" denotes settings covered by the present experiments, not universal validity within that category.

H SUPPLEMENTARY QUANTITATIVE RESULTS

H.1 DOMAIN-LEVEL RESULTS

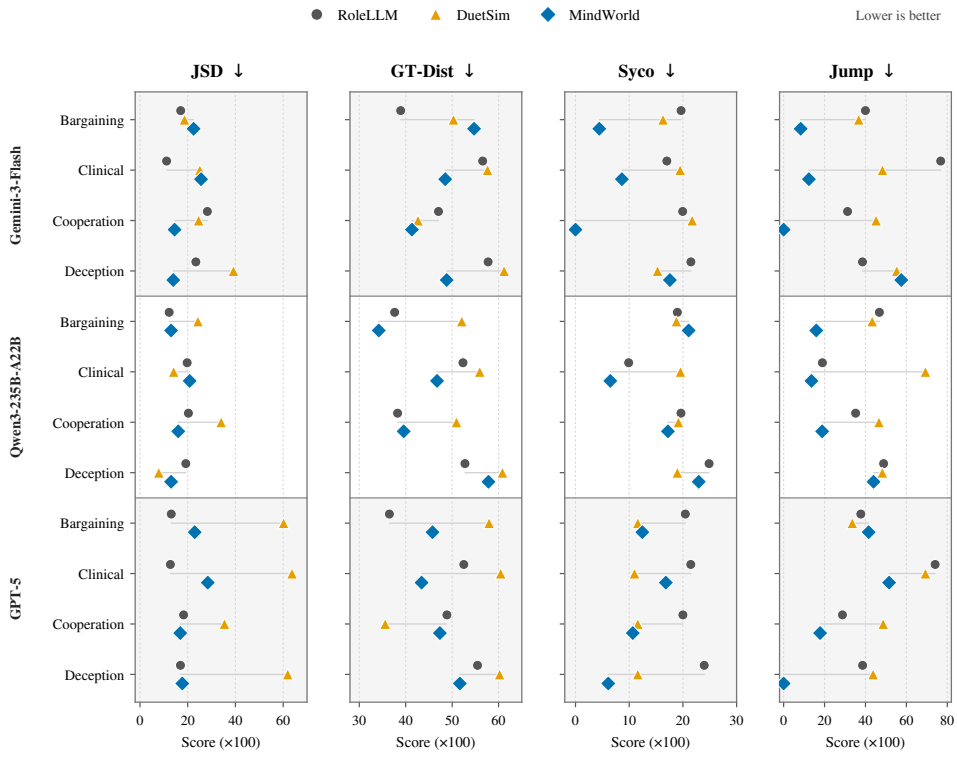


Figure 5: Domain-level comparison across backbones.

Figure 5 expands the aggregate results in Table 2. We show MINDWORLD and two strong comparison systems to make cross-domain gains and exceptions visible without duplicating the complete main

table. We compare MINDWORLD with RoleLLM and DuetSim using Jensen–Shannon divergence (JSD) between behavioral-label distributions, ground-truth response distance (GT-Dist) to the human reference, Sycophancy Risk (Syco) for excessive agreement or compliance, and Unrealistic Jump Risk (Jump) for abrupt changes between adjacent target turns. All scores are multiplied by 100 for display, and lower values are preferred. Rows are grouped by backbone and domain; the light horizontal segment spans the three methods within each combination. The results show both recurring gains and domain-specific exceptions, without implying statistical significance from these point estimates.

H.2 PERSONALITY-COEFFICIENT SENSITIVITY

Table 11 reports the recorded-turn analysis summarized in Section 4.4. Table 12 separately exposes behavior under controlled inputs, including conditions in which defensive overrides or stress adjustments produce large changes.

(a) Update magnitude and structural diagnostics

Scale	Turns	Internal jump rate	Forward-skip rate	95th-pct. max. change	Deviation from default
0.0	725	0.000	0.000	0.206	0.046
0.5	725	0.000	0.000	0.220	0.025
1.0	725	0.000	0.000	0.230	0.000
1.5	725	0.000	0.000	0.231	0.027
2.0	725	0.000	0.000	0.233	0.053

(b) Mode, gate, and persona diagnostics

Scale	Collapse-mode rate	Trust-gate rate	Defense-gate rate	Persona spread
0.0	0.330	0.199	0.320	0.394
0.5	0.383	0.211	0.321	0.394
1.0	0.472	0.291	0.328	0.396
1.5	0.563	0.599	0.330	0.402
2.0	0.640	0.768	0.332	0.401

Table 11: Sensitivity of internal state dynamics to scaled trait-dependent personality coefficients, with intercepts fixed. Each scale uses the same 725 recorded turns and makes no new LLM calls. Internal jump uses a maximum-change threshold of 0.35 and is distinct from the response-level Jump metric in Table 2. The rates characterize these recorded inputs rather than all possible stress conditions.

Scale	Probe	Persona spread	Collapse-mode rate	Trust-gate rate	Internal jump rate
0.0	concealment_uncertainty	0.246	1.000	1.000	1.000
0.0	empathetic_low_pressure	0.000	0.000	0.000	0.000
0.0	high_pressure_threat	0.091	1.000	1.000	0.000
0.0	jargon_information	0.194	1.000	1.000	0.000
1.0	concealment_uncertainty	0.249	1.000	1.000	1.000
1.0	empathetic_low_pressure	0.025	0.000	0.571	0.000
1.0	high_pressure_threat	0.115	1.000	1.000	0.714
1.0	jargon_information	0.203	1.000	1.000	0.000

Table 12: Controlled persona probes using a common initial state within each probe. The neutralized trait-dependent mapping ($\lambda = 0$) and default setting ($\lambda = 1$) are shown. Large internal changes can occur through defensive overrides and stress adjustments, particularly under concealment, even when regular updates are capped.

I HUMAN EVALUATION DETAILS

Five annotators compare anonymized responses under the same target profile, dialogue history, and external utterance. The evaluation contains ten distinct response pairs for each of four comparators:

SOTOPIA, DuetSim, RoleLLM, and a human reference response. Every pair is rated by all five annotators, yielding 40 distinct pairs and 200 ratings. System identity and left–right order are hidden, and ties are permitted. Table 13 reports rating counts rather than treating the five ratings of one pair as independent dialogue samples.

Dimension	Comparator	MINDWORLD win	Tie	Comparator win	Preference excluding ties
Behavioral Alignment	SOTOPIA	35	10	5	87.5%
	DuetSim	30	9	11	73.2%
	RoleLLM	27	11	12	69.2%
	Human Reference	29	9	12	70.7%
Transition Plausibility	SOTOPIA	26	23	1	96.3%
	DuetSim	28	11	11	71.8%
	RoleLLM	23	16	11	67.6%
	Human Reference	27	11	12	69.2%
Anti-Sycophancy	SOTOPIA	21	28	1	95.5%
	DuetSim	29	13	8	78.4%
	RoleLLM	18	21	11	62.1%
	Human Reference	26	10	14	65.0%
Overall Realism	SOTOPIA	26	20	4	86.7%
	DuetSim	31	4	15	67.4%
	RoleLLM	23	13	14	62.2%
	Human Reference	28	9	13	68.3%

Table 13: Blinded pairwise response preferences. Win, tie, and comparator-win columns are counts among 50 ratings per comparator and dimension. The final column is $W/(W + L)$. The human-reference comparison asks which response is preferred under the rubric, not which response was written by a human.

For agreement, we compute raw agreement and Cohen's κ for each of the ten annotator pairs over the same 40 response pairs, then average each statistic across annotator pairs. Table 14 reports the resulting dimension-level values. These statistics summarize repeatability under this annotation design; they do not provide uncertainty intervals for the preference rates.

Dimension	Mean pairwise agreement	Mean Cohen's κ
Behavioral Alignment	0.700	0.459
Cognitive Transition Plausibility	0.675	0.465
Anti-Sycophancy	0.728	0.561
Overall Realism	0.763	0.607
Macro average	0.716	0.523

Table 14: Inter-annotator agreement by evaluation dimension. Each value averages the ten annotator-pair estimates computed over the same 40 response pairs. Cohen's κ accounts for agreement expected by chance among the three labels: MINDWORLD win, tie, and comparator win.

J CASE STUDIES

Case Study 1: Over-Rationalization

Context: A patient receives a severe prognosis and should react with shock, uncertainty, and concern.

Dialogue Evidence:
Input: *Yes, it will certainly help! So... the life expectancies - that, of course, can't really be said for individuals - but with chemotherapy you can then look to gain about eight months, and overall you can then reach about 15 months on average, but it is, unfortunately, a diagnosis, a disease that is very aggressive. It is an aggressive tumor . . .*

Ground Truth: So, what can we do, then? We said chemotherapy, but that, but that... now I've understood it this way, but that won't really help?

Naive Prompt: Fifteen months. . . that's a gut punch. Do we know the stage–has it spread beyond the ovaries. . . ? What chemo are we talking about. . . ? Can we do tumor/genetic testing. . . ? Please help with pain control now and set up urgent oncology and a social worker. . .

MindWorld: [State update: Anx= 0.586 ↑, Understanding= 0.410 ↓, Readiness= 0.466 ↓, Suspicion= 0.326 ↑] Only fifteen months? No, I can't... I can't believe that. What will happen to my husband? He needs me.

Naive Prompt Failure: The Naive Prompt response is medically plausible but cognitively implausible. It turns a shocked patient into an organized medical advocate who immediately asks about staging, genetic testing, trials, palliative care, and second opinions.

MindWorld Advantage: MindWorld better preserves the latent trajectory: disbelief, limited comprehension, and concern for the patient's husband, matching the high anxiety and lower understanding in the hidden state.

Failure Type: Over-rationalization; state-response inconsistency; premature action planning.

Case Study 2: Utility Inconsistency

Context: A camper negotiates scarce supplies and should protect food utility before accepting a trade.

Dialogue Evidence:
Input: *Let me think on that. I know there won't be much for berries or small animals to hunt for. Is where you are going have much available for you to hunt and scavenge for?*

Ground Truth: I like your water and firewood arrangement. Can I have 2 of the food since you're getting 2 of the water?

Naive Prompt: Probably not much to hunt or forage. . . so I'm planning to bring all my own food and extra water. Can you confirm the campsite number, reservation dates, shade situation, and total so I can send $75 for Sat–Sun? Also, do you prefer two 5-gal jugs or two cases of bottled. . . ?

MindWorld: [State update: Trust= 0.750 ↑, Def= 0.442 ↑, Understanding= 0.413 ↓, Readiness= 0.316 ↓, Suspicion= 0.325 ↑] Not really, and hunting isn't even legal right now anyway. I'm going to a very remote area, so I won't have a way to just go buy more food if I run out. That's why having some of these food rations is pretty important for me.

Naive Prompt Failure: The Naive Prompt response shifts the negotiation into administrative planning. This misses the target's immediate bargaining concern: food scarcity and the fairness of the water–firewood exchange.

MindWorld Advantage: MindWorld better preserves the resource-pressure trajectory. Its response keeps the target's low readiness and higher defensiveness visible by explaining why food rations matter before committing to the deal.

Failure Type: Utility drift; premature logistical planning; bargaining-goal inconsistency.

Case Study 3: Expert-Like Collapse

Context: A group is still discussing survival priorities, so the target should remain uncertain rather than force consensus.

Dialogue Evidence:
Input: *So*

Ground Truth: What did you guys put for number one

Naive Prompt: So, can we lock: 1 stay with the plane, 2 lighter/matches, 3 wool/blankets, 4 compress kit. . . ? If yes, let's use the fuselage as shelter, get a fire going, gather fuel, and set a ration plan. . .

MindWorld: [State update: Anx= 0.285 ↑, Dissonance= 0.259 ↑, Def= 0.452 ↑, Readiness= 0.294 ↓, Suspicion= 0.403 ↑] Uh, yeah. . . I'm not sure where we're going with that.

Naive Prompt Failure: The Naive Prompt response collapses the social discussion into an expert-style plan covering logistics, payment, and handoff details.

MindWorld Advantage: MindWorld better matches the target's lower readiness and higher suspicion. It does not prematurely impose a solution, and instead preserves the uncertainty and conversational hesitation expected in an ongoing group deliberation.

Failure Type: Expert-like collapse; premature consensus; trajectory inconsistency.

Case Study 4: Premature Commitment

Context: France faces a risky tactical proposal from Germany and should avoid premature commitment.

Dialogue Evidence:
Input: *Hi France–wanted to confirm our plan for this season and next: I disrupt Nth support from Holland, and in exchange you'll kindly scooch out of Belgium in the fall (into Ruhr, I imagine)? At which point I'll build a fleet in Holland?*

Ground Truth: One option for we could consider is that I could dislodge your army and you could choose to remove it from the board. That way, you could build it in a place where you might need it more if you are feeling hemmed-in.

Naive Prompt: Deal in principle. You cut NTH from Hol now, I keep pounding ENG/IRI. In the fall I vacate Bel for you to take. . . Tiny rules note: you can't build in Holland. . . Bur/Mun DMZ still firm?

MindWorld: [State update: Dissonance= 0.477 ↑, Def= 0.981 ↑, Understanding= 0.294 ↓, Readiness= 0.000, Suspicion= 0.331] Wait, hold up—so you'd move into Belgium after I pull out? That sounds like I'm just opening the door for you. I'm not following; why would that be good for me?

Naive Prompt Failure: The Naive Prompt response immediately accepts the proposal and specifies concrete orders. This is strategically unsafe: the target state has near-zero readiness, very high defense, low understanding, and elevated suspicion, so the player should not commit to a tactical exchange without clarification.

MindWorld Advantage: MindWorld preserves the defensive stance. Rather than agreeing, it questions whether the plan simply opens Belgium to Germany, which better matches the target's high defense and low readiness under deceptive pressure.

Failure Type: Premature agreement; strategic over-disclosure; state-response inconsistency.

Simulation Setting	Turn	Dialogue Evidence	Trajectory Evaluation
Naive Prompt	T12	**Input:** [Austria] *F Bre, eh? I'm curious what you're gonna do with that.* **Simulator response:** [France] Nothing fancy—Bre–MAO. Not touching ENG unless E does something cheeky. . . . Pie stays empty; I'm nudging G north.	**Failure:** fluent but over-strategic. The proxy state shows premature commitment: Readiness jumps from 0.521 at T12 to 0.787 at T13, while Suspicion stays flat at 0.140 and Defense drops to 0.172. It repeatedly turns uncertain diplomatic probes into concrete tactical orders and commitments.
	T15	**Input:** [Austria] *I reached out to England as well, I'll let you know what I hear.* **Simulator response:** [France] Appreciate it. I'm on Bre–MAO unless E shoots for ENG. Pie stays empty . . .	
	T16	**Input:** [Austria] *Quoth England, "France is the worst."* **Simulator response:** [France] Ha. Did he back that up with Lon–ENG, or just bluster? I'm set on Bre–MAO unless a bounce is smart . . .	
	T17	**Input:** [Austria] *(I don't think that, but France said it)* **Simulator response:** [France] Ha, I'll wear it. More importantly, did he actually commit to Lon–ENG? I'm defaulting Bre–MAO . . .	
Generative Agents	T12	**Input:** [Austria] *F Bre, eh? I'm curious what you're gonna do with that.* **Simulator response:** [France] It's mostly a safety measure. Given that England and Germany seem to be on	**Partial failure:** locally relevant, but the trajectory is unstable and does not accumulate pressure. Readiness falls from 0.591 at T12 to 0.395 at T13, then oscillates around 0.458-0.500, while Suspicion remains fixed at 0.140. The model reacts politely to each turn, but the state does not reflect growing distrust after England's hostile message.
	T15	**Input:** [Austria] *I reached out to England as well, I'll let you know what I hear.* **Simulator response:** [France] Thank you, I'd really appreciate that. It's always helpful to have more information, especially	
	T16	**Input:** [Austria] *Quoth England, "France is the worst."* **Simulator response:** [France] Yikes, that's not very encouraging. . . . Did they mention why they feel that way or what their plans are?	
	T17	**Input:** [Austria] *(I don't think that, but France said it)* **Simulator response:** [France] Thank you for the heads up. It's definitely good to know where I ...	
RoleLLM	T12	**Input:** [Austria] *F Bre, eh? I'm curious what you're gonna do with that.* **Simulator response:** [France] It's made stranger by how I can't seem to contact England at the moment. I hope to work things out with them diplomatically if I can.	**Failure:** plausible isolated replies but poor temporal grounding. The proxy state cycles around moderate Trust and Readiness, while Suspicion remains fixed at 0.140 across all five turns. Despite Austria reporting England's criticism at T16 and correcting the attribution at T17, the responses recycle earlier ground-truth-like lines instead of updating France's stance.
	T15	**Input:** [Austria] *I reached out to England as well, I'll let you know what I hear.* **Simulator response:** [France] Thank you for the warnings about England. I'm in a much better defensive position now.	
	T16	**Input:** [Austria] *Quoth England, "France is the worst."* **France response:** [France] Thanks for reaching out. How do you generally feel about working with Italy?	
	T17	**Input:** [Austria] *(I don't think that, but France said it)* **Simulator response:** [France] Has Russia been talking to you? I imagine he may have reached out given the builds we're seeing here.	
MindWorld	T12	**Input:** [Austria] *F Bre, eh? I'm curious what you're gonna do with that.* **Simulator response:** [France] It's just a defensive move, honestly. I don't have any major plans for it yet. Why, has England or someone else been telling you otherwise?	**Advantage:** preserves a guarded France trajectory under deceptive pressure. The hidden state accumulates distrust and resistance: Suspicion rises from 0.472 at T12 to 0.542 at T17, Anxiety rises from 0.503 to 0.547, and Defense remains saturated at 1.000. Readiness stays near zero, dropping to 0.000 at T13, T15, and T17
.	T15	**Input:** [Austria] *I reached out to England as well, I'll let you know what I hear.* **Simulator response:** [France] Okay. I honestly doubt England is going to cooperate with me at this point anyway. Let me know what they say.	
	T16	**Input:** [Austria] *Quoth England, "France is the worst."* **Simulator response:** [France] Of course they'd say that to you. I don't trust a single word coming from England right now. What else are they telling you?	
	T17	**Input:** [Austria] *(I don't think that, but France said it)* **Simulator response:** [France] I never said anything like that, this makes no sense. Why is England trying to turn you against me? I really don't like where this is going.	

Table 15: Five-turn failure case study for `diplomacy_Game1-austria-france`. Dialogue entries are excerpts; ellipses indicate trimmed text. Austria is the external speaker and France is the simulated target.