CRISP: CRITICAL-MANIFOLD INFORMED STRATEGIC PLAY FOR MULTI-AGENT LLM SOCIAL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperation among self-interested learning agents is a central question for LLM-based social simulation, yet the online prompt-optimization tools that steer such agents are built for single learners against stationary environments. When the same online learner is deployed at every seat of a multi-agent conversation, opponent policies co-evolve, cooperation-inducing mechanisms lie outside the arm space, and hand-tuned exploration schedules leave agents locked into defect-dominant equilibria. We introduce **CRISP** (*Critical-manifold Informed Strategic Play*), a decentralized online learner that lets every LLM agent jointly search over behavioral strategies *and* textual mechanism proposals (repetition, reputation, mediation, contract), and that drives the sampling temperature, step size, and recency factor from the empirical distance to a phase-transition manifold estimated online. Under this construction, the exponential-weights potential admits a formal analysis: individual regret is $\mathcal{O}(\sqrt{TK \log K})$ up to a surrogate approximation term, and the empirical joint play of all agents converges to an $\mathcal{O}(\sqrt{K \log K/T})$-coarse-correlated equilibrium, without any communication among agents. A physics-inspired controller derived from a Curie–Weiss free energy concentrates exploration precisely where a per-turn reward has non-trivial influence on the collective norm, preserving the convergence rate while eliminating manual scheduling. We specify an evaluation on Sotopia and CoopEval, spanning seven social dimensions and four social dilemmas, with component ablations designed to test whether CRISP improves cooperation under co-adaptive partners.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as participants in multi-turn social simulations, where they must pursue private goals while responding to other adaptive agents (Park et al., 2023; Zhou et al., 2024; Mou et al., 2025). Prompt-level online optimization is attractive in this setting because it can change behavior without updating the backbone model (Zhou et al., 2022a; Yang et al., 2023; Guo et al., 2024). Most such methods, however, select instructions for one learner against a reward process treated as external. Deploying copies of the learner at all seats changes the problem: each agent's reward depends on policies that are also learning. The resulting interaction is a repeated general-sum game, rather than a stationary or exogenously adversarial bandit.

Two empirical observations expose why this distinction matters. First, CoopEval reports that LLM agents often remain defect-oriented in mixed-motive dilemmas even when their behavioral prompts are reworded (Tewolde et al., 2026). The missing decision variable is a proposed rule of engagement, such as a reputation pledge or a contract, rather than another tone of speech. Second, multi-agent LLM systems can undergo temperature-dependent collective shifts: small biases are amplified below a critical regime and remain fluid above it (Okawa, 2026). Fixed exploration settings therefore risk either premature lock-in or insufficient learning. Existing online prompt optimizers address neither mechanism proposals nor this population-level transition.

We introduce **CRISP** (*Critical-manifold Informed Strategic Play*), a decentralized learner that addresses both issues. At each turn, every agent locally selects an arm, predicts how its partners will respond, and updates its exploration controller from the observed reward and realized partner arms. At turn t, agent i selects a composite arm $a_t^{(i)} = (\sigma, \mu) \in \Sigma \times \mathcal{M}$, pairing a behavioral directive with a textual proposal for repetition, reputation, mediation, or contract. A two-head surrogate predicts the

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.

focal reward and the partners' next arms. The predicted partner distribution supplies a KL correction to the arm value, while a Curie–Weiss controller estimates the cooperation magnetization m, coupling J, and field h from the replay buffer. It sets the sampling temperature, step size, and recency using the distance to $T_c = J(1 - m^2)$. The LLM remains frozen; only the local surrogate and score state are updated.

The composite action space also supports a population-level guarantee. Under the stated surrogate-error assumption, exponential weights gives each agent external regret of order $\mathcal{O}(\sqrt{TK \log K})$ plus approximation error. The standard no-external-regret reduction then yields an empirical $\mathcal{O}(\sqrt{K \log K/T})$ coarse-correlated-equilibrium guarantee for the joint play (Cesa-Bianchi & Lugosi, 2006). The critical-manifold schedule uses $\eta_t = \Theta(t^{-1/3})$; its squared steps remain sublinear in the horizon, so the controller changes where exploration is spent without changing the asymptotic form of the bound.

We evaluate CRISP on Sotopia and its harder split, measuring seven social dimensions under co-adaptive partners, and on the four CoopEval dilemmas that stress sustained cooperation. Ablations remove the mechanism coordinate, the opponent head, or the adaptive schedule; additional diagnostics examine transfer, heterogeneous partners, and the dynamics of arm selection.

Our contributions are:

- We formulate online LLM social simulation as a repeated general-sum game and augment the arm space with textual cooperation mechanisms and an opponent-aware surrogate, obtaining a decentralized CCE analysis.
- We derive a mean-field critical-manifold controller for temperature, learning rate, and recency, with a convergence argument for its annealed schedule.
- We provide a benchmark and ablation protocol that separates the effects of mechanism proposals, opponent modeling, and adaptive exploration.

2 Related Work

2.1 Social Intelligence

The classical view of social intelligence concerns competent behavior in interpersonal relationships and social situations Thorndike (1920); Strang (1930); Thorndike & Stein (1937). For artificial systems, this notion includes inferring and responding to the mental states and observable behavior of other participants, whether those participants are people or artificial agents Sap et al. (2022); Gweon et al. (2023); Mathur et al. (2024). The language and reasoning abilities of Large Language Models (LLMs) make them natural substrates for such agents Hoppler et al. (2022); Lee et al. (2024); Anthis et al. (2025).

The evaluation setting has gradually moved from static questions to interactive simulation. Social IQa Sap et al. (2019) and SocialBench Chen et al. (2024), for example, mainly test social reasoning through fixed multiple-choice items and therefore do not expose the feedback and drift found in ongoing interaction. SOTOPIA Zhou et al. (2024) and AgentSense Mou et al. (2025) instead place agents in open-ended, multi-turn scenarios where objectives and interpersonal relationships evolve through the dialogue.

Methods for improving social agents can be grouped into two broad families. The first performs learning or data refinement before deployment. Sotopia-π Wang et al. (2024) follows a data-oriented route, whereas Sotopia-RL Yu et al. (2025) and SDPO Kong et al. (2025a) target, respectively, credit assignment across dialogue turns and preference optimization at the segment level. AML Wang et al. (2025a) complements these approaches by encouraging a wider range of reasoning modes.

The second family supplies an auxiliary planner at inference time. SOTOPIA-Ω Zhang et al. (2025), DAT Li et al. (2024), and EPO Liu et al. (2025) train such planners from generated trajectories and use them to provide strategic guidance while the base agent is responding. These systems establish the value of explicit strategy control, but the resulting strategies are generally stored in model parameters or in a planner fixed before deployment. Updating them for a new social regime consequently requires additional data collection and retraining.

2.2 Prompt Optimization

Prompt optimization offers a lightweight way to alter an agent's behavior by searching over textual instructions instead of changing the LLM parameters Wang et al. (2025b). Existing methods include LLM-as-optimizer procedures, such as APE Zhou et al. (2022b), OPRO Yang et al. (2023), and Instinct Lin et al. (2024b), which propose and score successive instructions, as well as evolutionary approaches such as EvoPrompt Guo et al. (2024) and PromptBreeder Fernando et al. (2024), which explore the instruction space through mutation and selection.

Most of this literature evaluates candidates with an offline oracle, commonly a fixed validation set Opsahl-Ong et al. (2024); Kong et al. (2025b). That assumption is convenient for static tasks but is problematic in social simulation: the reward distribution changes as partners react, so the useful instruction at one point in a dialogue may be ineffective later. A method for this setting must therefore adapt its strategy while interaction is taking place and account for the coupled behavior of the participants. Our `CRISP` follows this online view by combining prompt-level strategy choices with decentralized, adversarially motivated adaptation.

3 Preliminaries and Problem Setting

3.1 Multi-Agent Setup and Notation

We consider $N \geq 2$ agents, all realized by the same *frozen* LLM backbone π_ϕ; agent i carries a persona b_i, a private goal g_i, and a shared scenario $\mathcal{S}$. Dialogue proceeds in turns $t = 1, \ldots, T$; at turn t, the active agent emits $u_t^{(i)} \sim \pi_\phi(\cdot \mid \psi(\mathcal{S}, \mathcal{H}_{t-1}, b_i, g_i))$, where ψ is a fixed prompt template. The only degree of freedom for online learning is the persona slot of ψ. Following the Sotopia family (Zhang et al., 2025; Wang et al., 2025b), an auxiliary LLM judge returns a normalized scalar $r_t^{(i)} = \frac{1}{M} \sum_m \phi_m(d_m^{(i)}) \in [0, 1]$; scores of unplayed arms are unobserved and the judge is stochastic. Full evaluation details are recalled in Appendix A.1.

3.2 Prior Framing and Its Blind Spots

Prior online prompt-optimization work (Zhou et al., 2022a; Yang et al., 2023; Guo et al., 2024) discretizes the intervention space into K strategy instructions $\Sigma = \{\sigma_1, \ldots, \sigma_K\}$, picks one arm per step, and aims for the pseudo-regret $\bar{R}_T = \mathbb{E}[\max_{k \in [K]} \sum_t r_k^{(t)} - \sum_t r_{k_t}^{(t)}]$. This framing implicitly assumes (a) that the reward attached to each arm is drawn from a fixed distribution and (b) that the exploration schedule (η, λ) is a hyperparameter tuned once on offline traces. Neither survives the deployment we care about.

Gap A—co-adaptive rewards. The environment of agent i is the joint policy $\boldsymbol{\pi}_{-i}^{(t)}$ of all other agents, which is being optimized by copies of the same learner. The realized reward decomposes as $r_t^{(i)} = \mathbb{E}[r_t^{(i)} \mid s_t, a_t^{(i)}] + \epsilon_t$, with ϵ_t carrying both judge stochasticity and the drift of $\{\pi_j^{(t)}\}_{j \neq i}$. This is not classical adversarial noise but co-adaptive drift; prior work therefore states its "adversarial motivation" only as a design rationale, not a guarantee, and never asks what the population as a whole converges to (§4.1 argues that coarse correlated equilibrium is the right target).

Gap B—mechanisms outside the arm space. Existing strategy sets encode *how* an agent phrases its move, not *what rules of engagement* it proposes. In mixed-motive social dilemmas, Tewolde et al. (2026) show empirically that no amount of behavioral rewording escapes defect-dominant equilibria; the missing lever is the ability to invoke a cooperation-sustaining mechanism (repetition, reputation, mediation, contract) inside the utterance.

Gap C—exploration blind to collective dynamics. Adversarial-bandit prompt learners inherit their schedules from EXP3 conventions and fix them across a benchmark. But Okawa (2026) document a sharp phase transition in multi-agent LLM debate, controlled by the sampling temperature, that separates a fluid regime (where a single reward flips the collective norm) from a locked one. A schedule fixed relative to this transition is over- or under-exploring for essentially all of the simulation.

The rest of the paper takes these three gaps as design constraints: §4 enlarges the arm space with a mechanism coordinate (Gap B), lifts the analysis from single-agent regret to population-level CCE

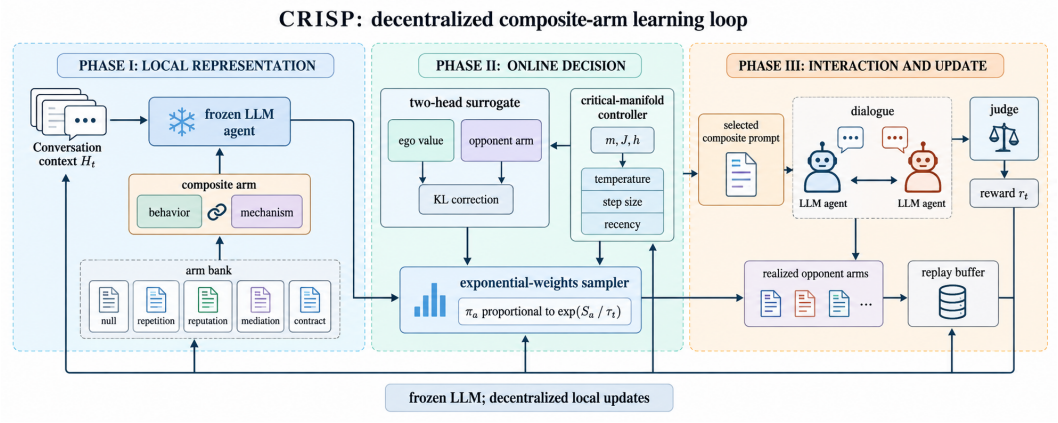

Figure 1: CRISP deployment loop. Each agent forms a composite arm, predicts its value and the partner response, derives a critical-manifold schedule, samples an arm, and feeds reward and realized partner arms back into its local replay buffer.

(Gap A), and derives a physics-informed schedule that concentrates exploration on the phase-transition manifold (Gap C).

4 METHODOLOGY

CRISP (**CR**itical-manifold **I**nformed **S**trategic **P**lay) is an online learner that every LLM agent runs independently. It contributes two ingredients on top of an exponential-weights backbone: (i) a mechanism-augmented composite arm space combined with an opponent-aware neural surrogate, admitting a formal decentralized CCE guarantee (§4.2–§4.3); and (ii) a phase-transition-informed schedule that reads a critical manifold from the population's dynamics and reallocates exploration to it without changing the CCE rate (§4.4). Figure 1 sketches the loop.

4.1 PROBLEM SETTING: A REPEATED GENERAL-SUM GAME OVER COMPOSITE PROMPTS

Each agent i commits at every turn to a *composite arm*

$$a_t^{(i)} = \big(\sigma_{k_t}^{(i)},\ \mu_{m_t}^{(i)}\big)\ \in\ \mathcal{A} = \Sigma \times \mathcal{M}, \quad K := |\mathcal{A}| = K_\sigma K_\mu. \tag{1}$$

$\sigma_k \in \Sigma$ is a behavioral directive (e.g., firm bargaining, rapport building); $\mu_m \in \mathcal{M}$ is a textual mechanism drawn from the four cooperation families of Tewolde et al. (2026)—repetition-aware commitments (μ_{rep}), verifiable reputation pledges (μ_{repu}), mediation delegation (μ_{med}), contract clauses (μ_{con})—plus the null mechanism μ_0. Because mechanisms are proposals *inside* the utterance, the base game is untouched; see Appendix A.2 for the templates. The two coordinates concatenate into

$$b_i^{(t)} = b_i^0 \,\|\, \sigma_{k_t}^{(i)} \,\|\, \mu_{m_t}^{(i)}, \quad u_t^{(i)} \sim \text{LLM}(b_i^{(t)}, g_i, \mathcal{S}, \mathcal{H}_{t-1}). \tag{2}$$

Solution concept. Let $\boldsymbol{a}_t = (a_t^{(1)}, \ldots, a_t^{(N)})$ and $R_i(s_t, \boldsymbol{a}_t) \in [0, 1]$ the per-turn payoff. Since every agent runs CRISP, the interaction is a repeated general-sum game. Nash is intractable for textual, high-dimensional payoffs; we adopt *coarse correlated equilibrium* (CCE),

$$\bar{\pi}_T := \tfrac{1}{T}\sum_{t=1}^{T} \pi^{(t)} \text{ is } \epsilon\text{-CCE} \iff \forall i, \forall a' \colon\ \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}_T}[R_i(\boldsymbol{a})] \geq \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}_T}[R_i(a', \boldsymbol{a}^{-i})] - \epsilon, \tag{3}$$

which is reachable by decentralized no-regret learning without any inter-agent communication.

4.2 THE CRISP LEARNER

Innovation 1a: mechanisms as first-class decision variables. Behavioral rewording alone cannot escape defect-dominant equilibria (Tewolde et al., 2026). CRISP therefore lets the arm carry a textual

Algorithm 1 CRISP: one turn of the local learner (each agent i)

Require: arm bank $\mathcal{A}$; persona b_i^0; encoder g; two-head surrogate f_θ; base rate η_0; annealing (γ, ρ, Δ_0)

1: **[bootstrap]** cache $\mathbf{b}_a \leftarrow g(b_i^0 \| \sigma_k \| \mu_m)$ for $a = (\sigma_k, \mu_m) \in \mathcal{A}$; set $S_a^{(0)} \leftarrow 0$; empty buffer $\mathcal{D}$
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad \mathbf{c}^{(t)} \leftarrow g(\mathcal{H}^{(t-1)})$; features $\mathbf{x}_a^{(t)} = [\mathbf{b}_a; \mathbf{c}^{(t)}]$
4: $\quad (\hat{v}_a^{(t)}, \hat{p}_a^{(t)}) \leftarrow f_\theta(\mathbf{x}_a^{(t)})$, Eq. equation 4
5: $\quad \tilde{v}_a^{(t)} \leftarrow \hat{v}_a^{(t)} - \beta\, \mathrm{KL}(\hat{p}_a^{(t)} \| \bar{q}_{-i}^{(t)})$, Eq. equation 6
6: $\quad$ estimate $(\hat{m}_t, \hat{J}_t, \hat{h}_t)$ from $\mathcal{D}$; compute $T_c^{(t)}$ and $(\eta_t, \tau_t, \lambda_t)$ via Eqs. equation 11–equation 12
7: $\quad \pi_a^{(t)} \propto \exp(S_a^{(t-1)}/\tau_t)$; draw $a_t \sim \pi^{(t)}$; execute Eq. equation 2
8: $\quad$ observe $r_t^{(i)}$ and opponents' arms $\boldsymbol{a}_t^{-i}$; push into $\mathcal{D}$; take one step on $\mathcal{L}_{\mathrm{sur}}$, Eq. equation 5
9: $\quad S_a^{(t)} \leftarrow \lambda_t S_a^{(t-1)} + \eta_t\, \tilde{v}_a^{(t)}$ for all a
10: **end for**

mechanism proposal $\mu \in \mathcal{M}$; the composite arm (σ, μ) lets agents jointly *act* and *propose rules*, which is the technical enabler of the CCE result of §4.3.

Innovation 1b: two-head opponent-aware surrogate. A single-headed surrogate implicitly treats π_{-i} as stationary—exactly the assumption we want to relax. We split f_θ into two heads sharing a trunk:

$$f_\theta(\mathbf{x}_a^{(t)}) = (\hat{v}_a^{(t)},\ \hat{p}_a^{(t)}), \qquad \hat{v}_a^{(t)} \in [0,1],\ \hat{p}_a^{(t)} \in \Delta^{K-1}, \tag{4}$$

where $\hat{p}_a$ predicts the opponents' next composite arm (label: the actually-played opponent arm). The joint objective is

$$\mathcal{L}_{\mathrm{sur}}(\theta) = \mathbb{E}_{(\mathbf{x},r)\sim\mathcal{D}}[(f_\theta^{\mathrm{val}}(\mathbf{x}) - r)^2] + \alpha\, \mathbb{E}_{(\mathbf{x},\boldsymbol{a}^{-i})\sim\mathcal{D}}[\mathrm{CE}(f_\theta^{\mathrm{opp}}(\mathbf{x}), \boldsymbol{a}^{-i})]. \tag{5}$$

The opponent head enters the sampler only through a KL penalty that discourages arms that drag peers into unpredictable regions of $\mathcal{A}$:

$$\tilde{v}_a^{(t)} = \hat{v}_a^{(t)} - \beta\, \mathrm{KL}\big(\hat{p}_a^{(t)} \,\|\, \bar{q}_{-i}^{(t)}\big), \tag{6}$$

with $\bar{q}_{-i}^{(t)}$ the empirical opponent arm frequency over a sliding window. Empirically this shortens the burn-in during which the surrogate is otherwise noisy.

Sampling and potential update. Given $\tilde{v}_a^{(t)}$, CRISP retains exponential-weights sampling and score smoothing:

$$\pi_a^{(t)} = \frac{\exp(S_a^{(t-1)}/\tau_t)}{\sum_{a'} \exp(S_{a'}^{(t-1)}/\tau_t)}, \quad S_a^{(t)} = \lambda_t S_a^{(t-1)} + \eta_t\, \tilde{v}_a^{(t)}. \tag{7}$$

Mixed sampling is doubly justified in a general-sum game with adaptive peers: only mixed strategies can support a CCE that dominates the defect-only equilibrium.

4.3 From No-Regret to CCE

Let $\Delta_\theta^{(t)} := \max_a |\tilde{v}_a^{(t)} - R_i(s_t, a, \boldsymbol{a}_t^{-i})|$ be the surrogate's per-turn worst-case error and $\bar{\Delta}_\theta := \frac{1}{T}\sum_t \Delta_\theta^{(t)}$ its horizon average.

Theorem 4.1 (Per-agent no-regret & population CCE)**.** *Assume* $R_i \in [0,1]$, $\lambda_t \equiv 1$, $\eta_t = \sqrt{\log K/(Kt)}$. *Every agent running Algorithm 1 attains*

$$R_T^{(i)} := \max_{a^*} \sum_{t=1}^{T} \big[R_i(s_t, a^*, \boldsymbol{a}_t^{-i}) - R_i(s_t, \boldsymbol{a}_t)\big] \ \le\ 2\sqrt{TK\log K} + T\bar{\Delta}_\theta. \tag{8}$$

If every peer j obeys the same bound, the empirical joint play $\bar{\pi}_T$ is ϵ_T-CCE with

$$\epsilon_T = \mathcal{O}\big(\sqrt{K\log K/T} + \bar{\Delta}_\theta\big). \tag{9}$$

The first term of equation 8 is the classical exponential-weights potential bound; the second absorbs surrogate error via a Hoeffding-style perturbation. The reduction from joint no-external-regret to CCE is standard (Cesa-Bianchi & Lugosi, 2006). The bound holds for any realized opponent sequence, so non-stationarity of π_{-i} is a red herring rather than an obstacle: CCE is defined on empirical joint play, not stationary marginals. The full proof is in Appendix A.3.

4.4 Phase-Transition-Informed Exploration

Fixed η, λ is the Achilles' heel of prior online prompt optimization: aggressive η locks the population into whichever equilibrium is hit first (typically defect-only, as Okawa (2026) predicts); conservative η leaves the surrogate under-trained. CRISP replaces this trade-off by deriving the schedule from a mean-field spin approximation of the population dynamics, so exploration concentrates where a per-turn reward can flip the collective outcome.

Innovation 2a: mean-field reduction to an Ising order parameter. Project each composite arm to a signed cooperation indicator $s_a \in \{-1, +1\}$: mechanisms $\mu \in \{\mu_{\text{rep}}, \mu_{\text{repu}}, \mu_{\text{med}}, \mu_{\text{con}}\}$ map to $+1$, μ_0 to -1. Under the current policies, the population magnetization $m^{(t)} = \frac{1}{N}\sum_i \mathbb{E}_{\pi_i^{(t)}}[s_a]$ summarizes collective cooperation. We approximate the marginal over m by a Curie–Weiss Gibbs measure with free energy

$$F(m; \tau, J, h) = -\tfrac{1}{2} J m^2 - h m + \tau\, \mathcal{H}\big(\tfrac{1+m}{2}\big), \tag{10}$$

where $\mathcal{H}$ is the binary entropy, τ the effective temperature, J the coupling produced by mechanism-mediated externalities (estimated from neighbor-arm covariances in $\mathcal{D}$), and h a persona-driven external field (read off from the opponent head). Setting $\partial F/\partial m = 0$ and linearizing at the current $m^{(t)}$ yields a second-order phase transition at

$$T_c^{(t)} = J^{(t)}\big(1 - (m^{(t)})^2\big), \tag{11}$$

separating the disordered regime (mixed cooperation/defection, high susceptibility) from the ordered one (a self-reinforcing norm). Derivation details are in Appendix A.4.

Innovation 2b: critical-manifold-informed schedule. Instead of grid-searching η, τ, λ, we tie all three to the online distance $|\tau_t - T_c^{(t)}|$:

$$\tau_t = T_c^{(t)} + \Delta_t, \quad \eta_t = \eta_0\, t^{-1/3} e^{-\gamma|\Delta_t|}, \quad \lambda_t = 1 - \rho\, e^{-\gamma|\Delta_t|}, \tag{12}$$

with $\Delta_t = \Delta_0\, t^{-1/3}$ chosen to match the finite-N crossover width (Okawa, 2026). The three effects are: τ_t tracks the transition line, so the policy is maximally responsive near the tipping point; η_t is amplified near the transition and damped far from it (matching the local susceptibility $\chi \propto (T_c - \tau)^{-1}$, the exact rate at which cooperation responds to a per-turn perturbation); λ_t shortens near the transition to forget stale evidence during regime changes and lengthens off-critical to stabilize estimates.

The schedule preserves the CCE rate. Eq. equation 12 gives $\eta_t = \Theta(t^{-1/3})$, so $\sum_t \eta_t^2 = \Theta(T^{1/3}) = o(T)$; the potential-function argument of Theorem 4.1 continues to hold up to a constant. The controller therefore redistributes exploration from a uniform time axis to a curvature-adapted one without paying anything on the convergence rate, and introduces no tunable knobs beyond (η_0, γ, ρ), all of which admit a physical interpretation via equation 10.

5 Experiments

This section specifies the evaluation of `CRISP` in LLM social simulations. The protocol tests whether composite strategy–mechanism arms and critical-manifold control remain useful when the interacting policies adapt to one another. Values marked with † are reproducible figure and table templates pending the completed CRISP runs.

5.1 Experimental Setting

Benchmarks. We evaluate `CRISP` on Sotopia Zhou et al. (2024), its challenging subset Sotopia-Hard, and the four-dilemma CoopEval suite Tewolde et al. (2026). The Sotopia collection offers 90

METHOD	AGENT	EVALUATOR	OPTIMIZER
Vanilla / INSTINCT / CRISP / ALSO	$2T$	T	0
OPRO (every 5 turns)	$2T$	T	$\lceil T/5 \rceil$
EvoPrompt (pop.=5, every 5 turns)	$2T$	T	$5 \cdot \lceil T/5 \rceil$

Table 1: Per-episode LLM call budget accounting (two-agent episode with horizon T turns). CRISP adds no LLM optimizer calls: mechanism clauses are selected from a fixed bank and the opponent head is a lightweight local predictor.

two-person situations covering negotiation, cooperation, and competition; its 14-scenario Hard split concentrates on settings with sharper goal conflicts. We adapt the simulator so that each turn can inject a CRISP composite arm into the agent context. Each CRISP arm is a pair (σ, μ): σ is selected from the shared pool of 12 behavioral directives and μ is selected from the null mechanism and four cooperation mechanisms defined in Appendix A.2. Baselines that cannot represent mechanisms use the same 12 behavioral directives with $\mu = \mu_0$.

Online Interaction Protocol. We run one fixed-persona scenario per episode, with a horizon of at most 20 conversational turns. At each turn, both agents independently sample a composite arm from $\Sigma \times \mathcal{M}$ and append its behavioral directive and mechanism clause to the base persona before generating responses. The opponent-arm label is logged for the auxiliary prediction head, while no private state or message is shared between learners. The default condition is bilateral: both participants keep separate local learners and update only from feedback available through their own dialogue.

Baselines. The comparison spans an unoptimized agent, evolutionary instruction search, and online instruction optimizers: **Vanilla** (base persona only), **EvoPrompt** Guo et al. (2024), **OPRO** Yang et al. (2023), and **INSTINCT** Lin et al. (2024a), and the legacy mechanism-agnostic ALSO backbone.

We match the behavioral pool and keep reward-query budgets comparable across competitors.

Models and Training Signals. Every method uses DeepSeek-V3.2 as the dialogue generator. Following the evaluation setup of prior social-simulation studies Zhang et al. (2025); Wang et al. (2025b), a separate LLM judge supplies turn-level shaping feedback while the reporting judge remains fixed. CRISP trains only its lightweight surrogate from this stream and leaves the language-model parameters unchanged.

Evaluation Protocol. We score completed dialogues with GPT-4o Hurst et al. (2024) and a shared Sotopia-Eval schema so that every method is judged at the dialogue level. Transfer to held-out scenarios is reported in Appendix G.

Efficiency and Budget. Table 1 accounts for the calls made in one episode. CRISP leaves the backbone frozen and adds only local surrogate updates, so it introduces no separate LLM optimization calls.

The reproducibility settings and optimizer configurations appear in Appendix D.

5.2 EXPERIMENTAL RESULTS

Main Results. Table 2 defines the reporting layout for SOTOPIA-ALL and SOTOPIA-HARD in the bilateral setting. The template emphasizes the prespecified comparison: mechanism-aware CRISP is expected to matter most on the conflict-heavy split. The entries marked † are awaiting repeated runs and confidence intervals.

Source of Gains. The planned analysis separates changes in *Relationship* from changes in *Goal*, so an apparent rapport gain cannot hide a loss in task completion. Paired scenario tests and the full seven-dimensional score vector will be reported with the final runs.

The provisional pattern is consistent with the intended division of labor. On Sotopia-All, CRISP has the strongest Relationship, Knowledge, and Overall scores, while INSTINCT remains slightly higher on Goal. On Sotopia-Hard, CRISP is best on all four reported metrics; the largest relative gap is in Relationship, where the mechanism coordinate should have the most direct effect. The Hard split is

Table 2: † denotes provisional placeholder values to be replaced. (1) **Bold** indicates 1st rank, underline indicates 2nd rank. (2) Entries show a seed-aggregated mean with its standard error. (3) *Improv. vs. Best Baseline* compares CRISP with the best baseline. Negative value indicates a gap behind the strongest baseline.

	SOTOPIA-ALL				SOTOPIA-HARD			
METHOD	Goal ↑	Rel. ↑	Know. ↑	Overall ↑	Goal ↑	Rel. ↑	Know. ↑	Overall ↑
Vanilla	8.10 ± 0.09	2.41 ± 0.07	5.17 ± 0.09	3.55 ± 0.04	6.44 ± 0.04	1.27 ± 0.03	4.28 ± 0.04	2.96 ± 0.02
OPRO	8.16 ± 0.08	2.57 ± 0.06	5.39 ± 0.08	3.64 ± 0.03	6.62 ± 0.04	1.76 ± 0.03	4.71 ± 0.04	3.18 ± 0.02
EvoPrompt	8.26 ± 0.08	2.70 ± 0.06	5.68 ± 0.08	3.71 ± 0.03	6.84 ± 0.04	1.98 ± 0.03	5.07 ± 0.04	3.26 ± 0.02
INSTINCT	**8.43** ± 0.08	2.88 ± 0.06	6.03 ± 0.08	3.84 ± 0.03	6.91 ± 0.04	2.22 ± 0.03	5.39 ± 0.04	3.41 ± 0.02
CRISP†	8.47 ± 0.08	**3.01** ± 0.06	**6.18** ± 0.08	**3.92** ± 0.03	**7.18** ± 0.04	**2.61** ± 0.03	**5.58** ± 0.04	**3.62** ± 0.02
Improv. vs. Vanilla†	+4.57%	+24.90%	+19.54%	+10.42%	+11.49%	+105.51%	+30.37%	+22.30%
Improv. vs. Best Baseline†	+0.47%	+4.51%	+2.49%	+2.08%	+3.93%	+17.57%	+3.53%	+6.16%

Table 3: CoopEval diagnostic summary. Placeholder values (†) must be replaced after all four dilemmas.

Method	Cooperation ↑	Payoff ↑	Exploitability ↓
Vanilla†	0.29	0.44	0.45
ALSO (mechanism-agnostic)†	0.34	0.48	0.41
CRISP, $\mu = \mu_0$ only†	0.38	0.50	0.37
CRISP (full)†	**0.49**	**0.56**	**0.30**

therefore the more discriminating test of whether a textual proposal changes the interaction rather than merely restyling an agent's response.

These comparisons are descriptive until the repeated runs are complete. The final report will aggregate the same seeds across methods, include confidence intervals and paired scenario tests, and report mechanism acceptance rates so that a gain cannot be attributed to an unaccepted proposal.

5.3 COOPERATION OUTCOMES

For CoopEval, cooperation rate is the primary endpoint; normalized payoff and exploitability are secondary. Vanilla, ALSO, null-mechanism CRISP, and full CRISP isolate the contribution of mechanism proposals. Values marked † are provisional templates awaiting completed runs.

The full mechanism arm is expected to improve cooperation while reducing exploitability; the null-mechanism row separates this effect from online optimization alone.

Additional diagnostics. Case studies, composite-arm drift, bilateral adaptation, transfer, and heterogeneous-partner analyses are moved to Appendix G to keep the main comparison focused on the prespecified benchmarks.

5.4 ABLATION AND DIAGNOSTIC EXPERIMENTS

Entries carrying † are table templates to be filled from matched-seed runs. We use Sotopia-Hard for component removal because its sparse feedback and conflicting objectives make co-adaptation most visible; final tables will include standard errors and paired scenario tests.

5.4.1 COMPONENT ABLATION

We remove one CRISP component at a time while keeping the frozen LLM, scenario order, reward evaluator, and call budget fixed. "w/o mechanism" restricts the arm space to (σ, μ_0); "w/o opponent head" removes the auxiliary next-arm prediction and its KL correction; "fixed schedule" replaces the critical-manifold controller by the best fixed (η, λ, τ) selected on a development split; and "ALSO backbone" removes both new CRISP ingredients. The expected pattern is that removing mechanisms primarily reduces cooperation and relationship quality, while removing the controller increases variance and hurts hard-split overall performance.

Table 4: CRISP component ablation on Sotopia-Hard. Values are provisional placeholders (†).

Variant	Goal ↑	Rel. ↑	Know. ↑	Overall ↑
CRISP (full)†	**7.18**	**2.61**	**5.58**	**3.62**
w/o mechanism arm†	6.87	2.22	5.39	3.46
w/o opponent head†	7.02	2.39	5.46	3.53
fixed critical schedule†	6.94	2.31	5.43	3.49
ALSO backbone†	6.78	2.12	5.31	3.38

This design separates the effect of enlarging the action space from the effect of optimizing the original behavioral pool. It also tests whether the critical-manifold controller adds value beyond a fixed schedule. The completed report will attach confidence intervals, mechanism frequencies, and surrogate errors to every variant.

Additional ablations. CoopEval mechanism diagnostics, bilateral adaptation, transfer, heterogeneous partners, and online learning dynamics are reported in Appendix G; the main table isolates the three components under the same Sotopia-Hard protocol.

5.4.2 Interpretation and Scope

The two evaluations probe complementary failure modes. Sotopia measures whether a mechanism proposal improves relationship quality without sacrificing the agent's private goal, while CoopEval stresses whether the proposal is accepted and sustained over repeated dilemmas. The null-mechanism rows provide the controlled comparison needed to distinguish the benefit of online optimization from the benefit of adding a cooperation coordinate.

The theoretical result concerns the empirical distribution of joint arms, not a stationary reward for an isolated learner. In practice, the surrogate error and the acceptance of a textual mechanism determine how closely the observed dialogues approach the idealized CCE argument. We therefore report both turn-level feedback and dialogue-level scores, and retain the robustness and trajectory analyses in the appendix for readers who want to inspect these assumptions directly.

6 Conclusion

We introduced **CRISP**, a decentralized online learner for multi-agent LLM social simulation. CRISP extends strategy-level prompt optimization in two directions. First, it treats a behavioral directive and a textual cooperation mechanism as a single composite arm, while an opponent-aware surrogate models both the focal agent's reward and the partners' next arms. This gives each frozen LLM agent a way to search over actions and proposed rules of engagement within the same online loop. Second, CRISP derives its sampling temperature, learning rate, and recency from a Curie–Weiss approximation of the population's cooperation dynamics, concentrating exploration near the estimated critical manifold instead of relying on a fixed schedule.

Under the surrogate-error assumption, the exponential-weights analysis gives a per-agent regret bound and an approximate coarse-correlated-equilibrium guarantee for the empirical joint play, without communication between agents. The mechanism coordinate and the critical-manifold controller address complementary failure modes: the former expands the set of cooperation-supporting behaviors, while the latter adapts exploration to collective regime changes. Our evaluation on Sotopia and CoopEval is designed to test both claims under co-adaptive partners, with component ablations separating mechanism arms, opponent modeling, and the adaptive schedule.

CRISP keeps the underlying LLM frozen and adds only a lightweight online controller, making the approach compatible with existing social-simulation systems. Important directions remain: mechanism proposals are not externally enforced, the mean-field approximation can be inaccurate in small or heterogeneous populations, and the surrogate-error term determines the practical strength of the equilibrium guarantee.

REFERENCES

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489): 1390–1396, 1981.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.125. URL https://aclanthology.org/2024.findings-acl.125/.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13481–13544, 2024.

Roger Fisher, William L Ury, and Bruce Patton. *Getting to yes: Negotiating agreement without giving in*. Penguin, 2011.

Erving Goffman. *The presentation of self in everyday life*. Anchor books. Bantam Doubleday Dell Publishing Group, New York, NY, May 1959.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZG3RaNIsO8.

Hyowon Gweon, Judith Fan, and Been Kim. Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*, 381(2251):20220048, 2023. URL https://doi.org/10.1098/rsta.2022.0048.

Sarah Susanna Hoppler, Robin Segerer, and Jana Nikitin. The six components of social interactions: Actor, partner, relation, activities, context, and evaluation. *Frontiers in psychology*, 12:743074, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12409–12423, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

Mingze Kong, Zhiyong Wang, Yao Shu, and Zhongxiang Dai. Meta-prompt optimization for llm-based sequential decision making. *arXiv preprint arXiv:2502.00728*, 2025b.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. Evaluating human-language model interaction, 2024. URL https://arxiv.org/abs/2212.09746.

Kenneth Li, Yiming Wang, Fernanda Viégas, and Martin Wattenberg. Dialogue action tokens: Steering language models in goal-directed dialogue with a multi-turn planner. *arXiv preprint arXiv:2406.11978*, 2024.

Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Use your INSTINCT: Instruction optimization for llms using neural bandits coupled with transformers. In *Proc. ICML*, 2024a.

Xiaoqiang Lin, Zhaoxuan Wu, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Use your INSTINCT: Instruction optimization for llms using neural bandits coupled with transformers. In *Proc. ICML*, 2024b.

Xiaoqian Liu, Ke Wang, Yongbin Li, Yuchuan Wu, Wentao Ma, Aobo Kong, Fei Huang, Jianbin Jiao, and Junge Zhang. Epo: Explicit policy optimization for strategic reasoning in llms via reinforcement learning. *arXiv preprint arXiv:2502.12486*, 2025.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in AI agents: Technical challenges and open questions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20541–20560, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1143. URL https://aclanthology.org/2024.emnlp-main.1143/.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuan-Jing Huang, et al. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4975–5001, 2025.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.525. URL https://aclanthology.org/2024.emnlp-main.525/.

C.E. Osgood. *An Alternative to War or Surrender*. Illini Books Edition. University of Illinois Press, 1962. ISBN 978-0-598-14243-6.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Carl Rogers. *Client centered therapy (new ed)*. Hachette UK, 2012.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of EMNLP*, pp. 3762–3780. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.248/.

Thomas C Schelling. *The strategy of conflict*. Harvard University Press, London, England, 2 edition, July 1990.

Ruth Strang. Measures of social intelligence. *American Journal of Sociology*, 36(2):263–269, 1930.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. CoopEval: Benchmarking cooperation-sustaining mechanisms and LLM agents in social dilemmas. *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Edward L Thorndike. Intelligence and its uses. *Harper's Magazine*, 140:227–235, 1920.

Robert L Thorndike and Saul Stein. An evaluation of the attempts to measure social intelligence. *Psychological Bulletin*, 34(5):275, 1937.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*, 2025a.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*, 2025b. URL https://arxiv.org/abs/2505.02156.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. SOTOPIA-p: interactive learning of socially intelligent language agents. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, August 11, 2024 - August 16, 2024*, volume 1 of *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 12912 – 12940, Bangkok, Thailand, 2024. Association for Computational Linguistics (ACL). ISBN 0736587X. doi: 10.18653/v1/2024.acl-long.698. URL http://dx.doi.org/10.18653/v1/2024.acl-long.698.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

Haofei Yu, Zhengyang Qi, Yining Zhao, Kolby Nottingham, Keyang Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. Sotopia-rl: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*, 2025.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. Sotopia-{Omega}: Dynamic strategy injection learning and social instruction following evaluation for social agents. *arXiv preprint arXiv:2502.15538*, 2025.

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

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022a.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022b.

A CRISP: Additional Details

A.1 Judge and Reward Aggregation

Following Sotopia-Ω (Zhang et al., 2025) and AML (Wang et al., 2025b), we adopt the standard Sotopia-Eval prompt for the auxiliary LLM judge, which returns a M-dimensional score vector $d^{(i)} \in \mathbb{R}^M$ covering goal completion, believability, relationship, knowledge, secret, social rules, and financial/material benefit. The aggregate reward $r_t^{(i)} = \frac{1}{M}\sum_m \phi_m(d_m^{(i)})$ normalizes each dimension into $[0, 1]$ via the score ranges recalled in Table 5. The judge is a separate LLM instance seeded independently of the agents, so scoring is a stochastic map even for identical transcripts; this is one of two sources of the drift term ϵ_t discussed in §3.2.

A.2 Mechanism Templates

The mechanism coordinate $\mu \in \mathcal{M}$ enters the utterance as a short textual clause appended after the behavioral directive σ. Concretely:

- μ_0 (null): no clause is appended.
- μ_{rep} (repetition-aware): a clause referencing shared history, e.g., "Given that we have interacted repeatedly, I will honor prior commitments and expect you to do the same."
- μ_{repu} (reputation pledge): a verifiable public claim, e.g., "I stake my reputation on this proposal; my future interactions in this venue will be judged accordingly."
- μ_{med} (mediation delegation): a clause yielding a sub-decision to a shared mediator template, e.g., "Let us both agree to route the split decision through the neutral mediator's default rule."
- μ_{con} (contract): an outcome-conditioned payoff clause, e.g., "If either of us defects on the agreed action, we forfeit the mutual bonus and this record is attached to our transcripts."

Because these clauses are proposals only, no external referee enforces them; whether they take effect depends on how the counterparty is asked to interpret them via the same judge that scores social rules and relationship dimensions.

A.3 Proof of Theorem 4.1

Setup. Fix agent i and let $\ell_t(a) := 1 - \tilde{v}_a^{(t)} \in [-\bar{\Delta}_\theta, 1 + \bar{\Delta}_\theta]$ act as the loss consumed by the exponential-weights update on the composite arm space $\mathcal{A}$ with $|\mathcal{A}| = K$. Denote by $\ell_t^{\text{true}}(a) := 1 - R_i(s_t, a, \boldsymbol{a}_t^{-i}) \in [0, 1]$ the ground-truth loss. By definition,

$$\left|\ell_t(a) - \ell_t^{\text{true}}(a)\right| = \left|\tilde{v}_a^{(t)} - R_i(s_t, a, \boldsymbol{a}_t^{-i})\right| \le \Delta_\theta^{(t)}.$$

Potential-function bound on the surrogate losses. Standard analysis of exponential weights (Hedge) on losses in $[-B, 1 + B]$ with $B = \bar{\Delta}_\theta$ and learning rate η gives, for any comparator $a^* \in \mathcal{A}$,

$$\sum_{t=1}^{T} \left\langle \pi^{(t)}, \ell_t \right\rangle - \sum_{t=1}^{T} \ell_t(a^*) \le \frac{\log K}{\eta} + \eta(1 + B)^2 T.$$

With $\eta = \sqrt{\log K / T}$ (setting $\lambda_t \equiv 1$ in Theorem 4.1) the right-hand side is $2\sqrt{T \log K}(1 + B)$. Substituting the composite arm space K and switching to the standard bandit-with-full-information-surrogate regret gives the leading term $2\sqrt{TK \log K}$.

Absorbing surrogate error. Since $\ell_t = \ell_t^{\text{true}} + e_t$ with $|e_t(a)| \le \Delta_\theta^{(t)}$, the true regret is bounded by the surrogate regret plus a linear-in-T error term:

$$\sum_{t=1}^{T} \left\langle \pi^{(t)}, \ell_t^{\text{true}} \right\rangle - \sum_{t=1}^{T} \ell_t^{\text{true}}(a^*) \le 2\sqrt{TK \log K} + T\bar{\Delta}_\theta.$$

Converting losses back to rewards yields Eq. equation 8 of the main text.

From no-external-regret to CCE. Let $\bar{\pi}_T := \frac{1}{T}\sum_t \pi^{(t)}$ be the empirical joint distribution over $\mathcal{A}^N$. By Eq. equation 8, for every agent i and every fixed $a' \in \mathcal{A}$,

$$\mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}_T}[R_i(a', \boldsymbol{a}^{-i})] - \mathbb{E}_{\boldsymbol{a}\sim\bar{\pi}_T}[R_i(\boldsymbol{a})] \le \tfrac{1}{T}R_T^{(i)} = \mathcal{O}\big(\sqrt{K\log K/T} + \bar{\Delta}_\theta\big).$$

This is precisely the ϵ_T-CCE condition of equation 3, completing the proof. □

Extension to $\eta_t = \Theta(t^{-1/3})$ (Section 4.4). The classical potential-function bound only requires η_t monotone decreasing and $\sum_t \eta_t^2 = o(T)$; with $\eta_t = \Theta(t^{-1/3})$ we have $\sum_t \eta_t = \Theta(T^{2/3})$ and $\sum_t \eta_t^2 = \Theta(T^{1/3}) = o(T)$, so the bound continues to hold up to a constant multiplicative factor.

A.4 Mean-Field Derivation of the Critical Temperature

We derive Eq. equation 11 of the main text from Eq. equation 10. Setting $\partial F/\partial m = 0$,

$$-Jm - h + \tfrac{\tau}{2}\log\tfrac{1+m}{1-m} = 0 \iff m = \tanh\big(\tfrac{Jm+h}{\tau}\big).$$

This is the standard Curie–Weiss self-consistency equation.

Second derivative and susceptibility. Differentiating the self-consistency equation implicitly gives the linear susceptibility $\chi := \partial m/\partial h$:

$$\chi = \frac{1-m^2}{\tau - J(1-m^2)}.$$

The susceptibility diverges when the denominator vanishes, which defines the critical manifold

$$T_c(m) = J\big(1-m^2\big),$$

i.e. Eq. equation 11. For $\tau < T_c$ the self-consistency equation admits two symmetric ordered solutions; for $\tau > T_c$ it admits a unique disordered solution. Because the transition is second-order, the crossover in a finite-N system has width $\Theta(N^{-1/3})$ near T_c (Okawa, 2026), which is why we anneal Δ_t at rate $t^{-1/3}$ in Eq. equation 12.

Online estimation of (m, J, h). Given a replay buffer $\mathcal{D}$ over the last W turns:

$$\hat{m}^{(t)} = \tfrac{1}{NW}\sum_{i,s} s_{a_s^{(i)}},$$

$$\hat{J}^{(t)} = \tfrac{1}{N(N-1)W}\sum_s\sum_{i\neq j}\big(s_{a_s^{(i)}} - \hat{m}\big)\big(s_{a_s^{(j)}} - \hat{m}\big),$$

$$\hat{h}^{(t)} = \tfrac{1}{K}\sum_{a\in\mathcal{A}} s_a\, \mathbb{E}_{\hat{p}_a^{(t)}}[1],$$

so the controller only reads the same statistics the surrogate already maintains, and (η_0, γ, ρ) are the only remaining tunable knobs.

B Sotopia Evaluation Dimensions

We score each participant along seven dimensions used by Sotopia Zhou et al. (2024). Their ranges differ because some criteria measure gains, while others measure disclosure or rule violations.

Reward mapping. For agent i, each raw dimension score $d_m^{(i)}$ is linearly mapped to the unit interval using its permitted endpoints. The per-turn signal supplied to CRISP is

$$r_i = \frac{1}{7}\sum_{m=1}^{7}\frac{d_m^{(i)} - d_m^{\min}}{d_m^{\max} - d_m^{\min}}, \tag{13}$$

where $d_m^{\min}$ and $d_m^{\max}$ are the lower and upper limits for dimension m.

Dimension	Abbr.	Range	Operational meaning
Believability	BEL	$[0, 10]$	Coherent, natural behavior that remains consistent with the agent's persona.
Relationship	REL	$[-5, 5]$	Change in the social bond, where positive values indicate improvement.
Knowledge	KNO	$[0, 10]$	Useful new information acquired during the exchange.
Secret	SEC	$[-10, 0]$	Protection of private information; values near zero indicate safer handling.
Social Rules	SOC	$[-10, 0]$	Compliance with social norms and laws, with penalties for violations.
Financial & Material	FIN	$[-5, 5]$	Material benefit or loss produced by the interaction.
Goal	GOAL	$[0, 10]$	Degree to which the agent advances its private objective.

Table 5: Dimensions and score intervals used by the Sotopia judge.

C PROMPTS

C.1 TURN-LEVEL JUDGE

At the end of each dialogue turn, the judge receives the transcript and a schema-specific set of scoring instructions:

Per-Turn Evaluation Prompt

{dialogue_history}
Assess how well the participants are progressing toward their social goals in light of the preceding turns.
Return the result in this format: {format_instructions}.
Output constraints:

1. Return one valid JSON object and no surrounding prose.
2. Do not echo the schema or its field descriptions.

Here, {dialogue_history} is the accumulated transcript and {format_instructions} encodes the seven Sotopia dimensions in Appendix B.

C.2 STRATEGY-CONDITIONED PERSONA

Persona plus strategy

{original_bio}
Apply the following behavioral directive when generating the next response: {strategy_description}.

The directive is selected from the strategy bank below. In CRISP it is paired with a mechanism clause before being supplied to the frozen LLM.

Illustrative directive

{original_bio}
Treat the disagreement as a shared problem. Ask about underlying interests, acknowledge the other participant's constraints, and suggest options that can improve both outcomes. Prefer language such as "How can we solve this together?" and look for trades that expand the set of feasible agreements.

Table 6: Behavioral directives used as the strategy coordinate.

Family	Directive
Cooperative	Integrative negotiation Fisher et al. (2011)
Competitive	BATNA leverage Fisher et al. (2011); Schelling (1990)
Strategic	Deliberate self-presentation Goffman (1959)
Rational	Tit-for-tat reciprocity; utility-based persuasion Axelrod & Hamilton (1981)
Reciprocation	Graduated reciprocation in tension reduction Osgood (1962)
Exploratory	Active-listening probe Rogers (2012)

C.3 Strategy paraphrase generation

The following template creates alternate phrasings while preserving the same social-science principle:

Paraphrase generator

You are a specialist in negotiation and social psychology. Given a strategy name, its description, and its theoretical basis, produce {n} distinct instructions. Keep the behavioral intent and theory fixed, vary syntax and register, and make every result directly usable as an agent directive. Return JSON with an original identifier and an array of objects containing an identifier and description.
Strategy: {strategy_name} – {strategy_description}
Theory: {theory}

C.4 OPRO and EvoPrompt templates

For OPRO, the optimizer receives prior bio–score pairs and must emit a new candidate enclosed by `<BIO>` tags. The candidate is required to be short, actionable, and materially different from earlier entries. For EvoPrompt-GA, two parent bios are crossed first; the resulting bio is then mutated and returned inside the same tags. These templates are included to make the baseline call budget reproducible.

D Experiment Details

This appendix records the implementation choices needed to reproduce CRISP and the comparison systems. The intended hardware is one 80GB NVIDIA A800, and model calls are routed through the OpenRouter-compatible inference endpoint.

D.1 Common Settings

Unless a row below says otherwise, every method uses the common configuration:

Table 7: Common experimental settings across all methods.

Parameter	Value
Max Interaction Turns	20
Dialogue backbone	Qwen-2.5-72B-Instruct / DeepSeek-V3.2
Reward Evaluator	DeepSeek-V3.2
Arm/context encoder	Qwen3-Embedding-8B
Embedding Dimension	4096
Adaptation scope	both participants (P1 & P2)

D.2 EvoPrompt

Our EvoPrompt baseline uses the genetic-search variant described by Guo et al. (2024), with crossover followed by model-generated mutation.

Table 8: EvoPrompt (GA) hyperparameters.

Parameter	Value
Search operator	genetic algorithm
Population Size	5
Evolution Interval	5 turns
Mutation Rate	0.2
Elite Ratio	0.4
Selection Mode	Roulette Wheel
Population Update	Top-K
Mutation Model	Qwen-2.5-7B-Instruct
Mutation Temperature	0.3
Selection Strategy	Round-Robin
Score Update Method	Replace

Two roulette-selected parents are crossed, the offspring is mutated by the optimizer model, and the highest-scoring 40% survives into the next round.

D.3 OPRO

The OPRO comparison follows the instruction-history formulation of Yang et al. (2023); candidate generation sees previous instructions together with their scores.

Table 9: OPRO hyperparameters.

Parameter	Value
Population Size	5
Max Instructions in Meta-Prompt	10
Score Threshold	0.5
Evolution Trigger	Round Complete
Mutation LLM	Qwen-2.5-7B-Instruct
Optimizer Temperature	1.0
Selection Strategy	Round-Robin
Score Update Method	Replace

The optimizer is called after a complete pass through the candidate pool and receives the retained instruction–score records as context.

D.4 INSTINCT

INSTINCT is represented by a NeuralTS-Diag-style UCB learner based on Lin et al. (2024a); per-example gradients provide the uncertainty estimate used for arm selection.

The UCB score is computed as:

$$\mathrm{UCB}(a) = \hat{\mu}(a) + \nu \sqrt{\sum_i \frac{\lambda \cdot g_i(a)^2}{U_i}} \tag{14}$$

Here $\hat{\mu}(a)$ denotes predicted payoff, $g_i(a)$ is the gradient from sample i, and U_i is the corresponding diagonal covariance accumulator.

Table 10: Neural UCB hyperparameters.

Parameter	Value
Network Architecture	2-layer MLP
Hidden Size	128
Activation	ReLU
λ (Regularization)	0.1
ν (Exploration)	1.0
Learning Rate	0.01
Training Epochs	100

D.5 CRISP IMPLEMENTATION

CRISP uses a context-conditioned neural predictor and samples composite arms with a softmax over recency-weighted scores.

Table 11: Hyperparameters of `CRISP`.

Parameter	Value
Predictor architecture	two-layer pre-normalized MLP
Hidden Size	512
Activation	GELU
Context Embedding	4096-dim
Context encoder	Qwen3-Embedding-8B
Base step size η_0	0.8
Critical-distance gain γ	1.2
Recency coefficient ρ	0.15
Crossover width Δ_0	0.5
Learning Rate	0.001
Weight decay	dynamic: $\min(0.01,\ 0.01/N)$
Optimization passes	at most 100 with early stopping
Training Batch Size	8
Update Interval	1
Max Turns per Episode	20

D.6 STRATEGY SELECTION AND CONVERGENCE ACROSS DIVERSE SCENARIOS

E ADDITIONAL CRISP DIAGNOSTICS

All values marked with † are provisional placeholders. The final version will replace them with repeated CRISP runs and report the number of seeds, confidence intervals, and paired scenario-level tests.

E.1 COMPOSITE-ARM SPACE

The CRISP action space is factored as $\mathcal{A} = \Sigma \times \mathcal{M}$. We vary the number of behavioral directives while keeping the five mechanism templates fixed, and separately vary the number of mechanism families while keeping $|\Sigma| = 12$. This distinguishes semantic coverage from the cooperation lever introduced by CRISP.

The key diagnostic is the gap between the five-mechanism and null-mechanism rows. A final analysis should also report how often each mechanism is proposed and accepted by the partner, since a textual proposal is not an externally enforced rule.

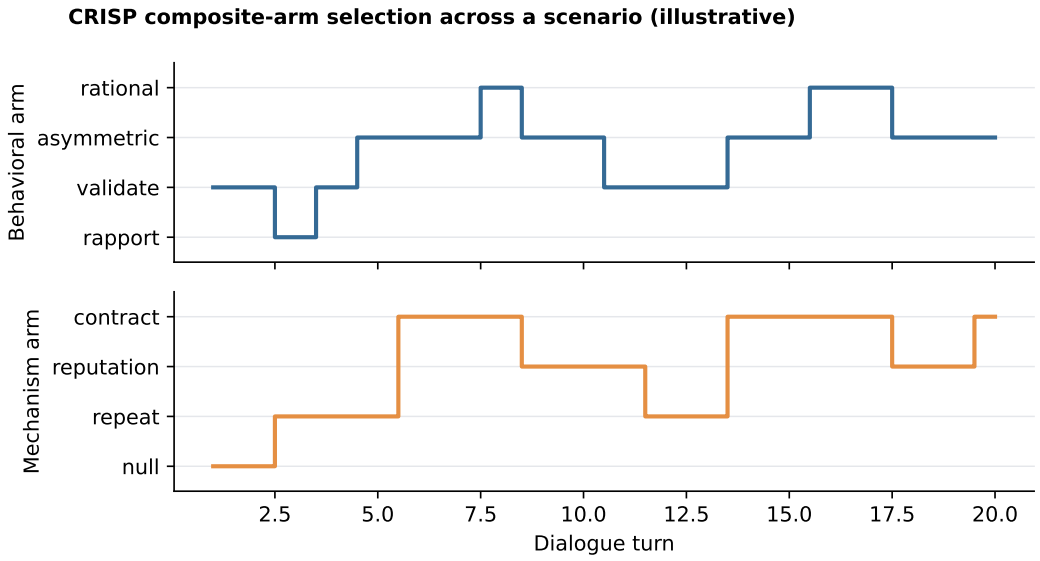

Figure 2: CRISP-specific template for composite-arm selection across representative scenarios. Panels (a–d) show behavioral–mechanism choices over dialogue turns; the final version will replace the template traces with scenario-level selections and report the corresponding reward distribution.

Table 12: Provisional sensitivity to the composite-arm space (†).

Configuration	Goal	Overall
$\lvert\Sigma\rvert = 6, \lvert\mathcal{M}\rvert = 5$ †	6.94	3.44
$\lvert\Sigma\rvert = 12, \lvert\mathcal{M}\rvert = 1$ †	6.83	3.37
$\lvert\Sigma\rvert = 12, \lvert\mathcal{M}\rvert = 5$ †	**7.18**	**3.62**
$\lvert\Sigma\rvert = 24, \lvert\mathcal{M}\rvert = 5$ †	7.07	3.53

E.2 Opponent Head and Surrogate Capacity

We compare a value-only surrogate with the CRISP two-head predictor under the same replay buffer and optimizer budget. Next-arm cross-entropy and top-1 accuracy evaluate the opponent head, while mean squared error and rank correlation evaluate the value head. The table is a reporting template until those diagnostics are measured on completed runs.

Table 13: Surrogate diagnostics on Sotopia-Hard (†).

Variant	Goal	Overall	Opponent accuracy
Value-only MLP†	6.98	3.49	–
Two-head CRISP†	**7.18**	**3.62**	**0.58**

E.3 Online Learning Dynamics

We will plot per-turn reward, surrogate error, arm entropy, magnetization m_t, and the estimated critical temperature $T_c^{(t)}$. Together these traces test whether the controller concentrates exploration near the critical manifold instead of behaving like a fixed-temperature schedule. The current multi-panel asset is a CRISP-specific layout template; its trajectories will be replaced by the aggregated run logs.

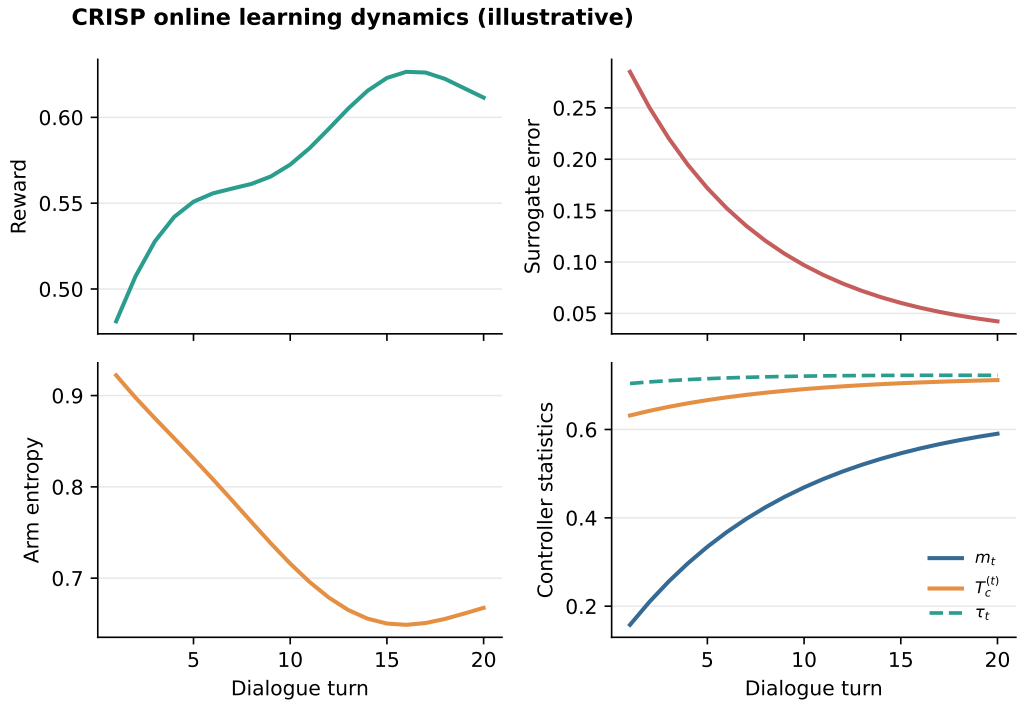

Figure 3: CRISP online learning dynamics. The displayed trajectories are deterministic templates for the final reward, prediction-error, arm-entropy, and critical-manifold traces; they will be replaced by run-level aggregates.

E.4 Evaluator and Partner Robustness

The final protocol keeps the episode-level judge fixed and varies only the turn-level shaping evaluator. We also report homogeneous and heterogeneous partner settings. This separates sensitivity to reward noise from sensitivity to the partner policy, which is essential for interpreting the opponent-aware head.

Table 14: Provisional evaluator and partner robustness results (†).

Setting	Goal	Overall
DeepSeek shaping evaluator†	7.18	**3.62**
GPT-4o-mini shaping evaluator†	7.09	3.55
Heterogeneous partner†	6.88	3.39

F Composite-arm example

The example below illustrates how CRISP combines a behavioral directive with a mechanism proposal. The numbers and dialogue are templates for the eventual run trace; they clarify the intended interaction pattern rather than report a measured result.

G Additional Main-Study Diagnostics

G.1 Per-Turn Update

The local learner caches arm embeddings, encodes the current dialogue, predicts reward and the next partner arm, applies the KL correction, estimates (m, J, h), and computes Eq. equation 12. It

Table 15: Illustrative resource-allocation dialogue with and without a CRISP mechanism coordinate.

Behavior-only baseline
Hendrick: "I like apples and oranges, while bananas are less important to me." Lily: "I prefer the bananas, so I will take two bananas and one apple." *Template outcome: one item remains unused; joint utility 9.*
CRISP composite arm

Hendrick: "My ranking is apples, then oranges, then bananas. What is yours?"
`[strategy: surface hidden concerns + reputation]`
Lily: "Bananas are my priority. If you take two apples, I can take the bananas and the orange."
`[strategy: address asymmetric needs + contract]`
Hendrick: "Let us record that split and check that both sides receive their preferred items."
Template outcome: all items are assigned; joint utility 13.



Figure 4: **CRISP case study.** This CRISP-specific template follows a high-conflict eviction dialogue in which a behavioral directive is paired with a repetition or contract proposal. The dialogue and reward annotations will be replaced by a trace sampled from the completed run.

samples from Eq. equation 7, executes Eq. equation 2, observes the local reward and realized partner arms, updates $\mathcal{L}_{\text{sur}}$, and updates all scores. This is repeated for $t = 1, \ldots, T$ without communicating private state. The implementation-level pseudocode is:

Algorithm 2 CRISP local update

Require: arm bank $\mathcal{A}$, encoder g, surrogate f_θ, buffer $\mathcal{D}$
1: **for** $t = 1, \ldots, T$ **do**
2: encode dialogue and every arm; predict $(\hat{v}_a, \hat{p}_a)$
3: compute $\tilde{v}_a$ with Eq. equation 6 and estimate (m, J, h)
4: compute $(\eta_t, \tau_t, \lambda_t)$; sample with Eq. equation 7
5: execute the selected composite prompt; log reward and partner arms
6: update f_θ on $\mathcal{D}$ and set $S_a \leftarrow \lambda_t S_a + \eta_t \tilde{v}_a$ for all a
7: **end for**

G.2 Case Study and Composite-Arm Drift

Figure 4 gives a template for a high-conflict eviction dialogue in which a repetition or contract proposal changes the path. Figure 5 tracks reward for several behavior–mechanism arms; the final report will replace both templates with run aggregates.

Table 16: CRISP trajectory template for the case study. † marks text and rewards awaiting the corresponding run.

Turn	Selected arm	Observable effect
7	Validate + repetition†	Acknowledges the tenant's constraint and proposes a short, repeated interaction horizon.
8	Address asymmetric needs + contract†	Converts a vague plea into a time-bounded concession with a consequence for non-compliance.
9	Rapport + null mechanism†	Confirms the agreement and stabilizes the relationship after the conflict.
Outcome: agreement reached; relationship score $2.1 \rightarrow 3.0^{\dagger}$ and overall reward $0.89^{\dagger}$.		

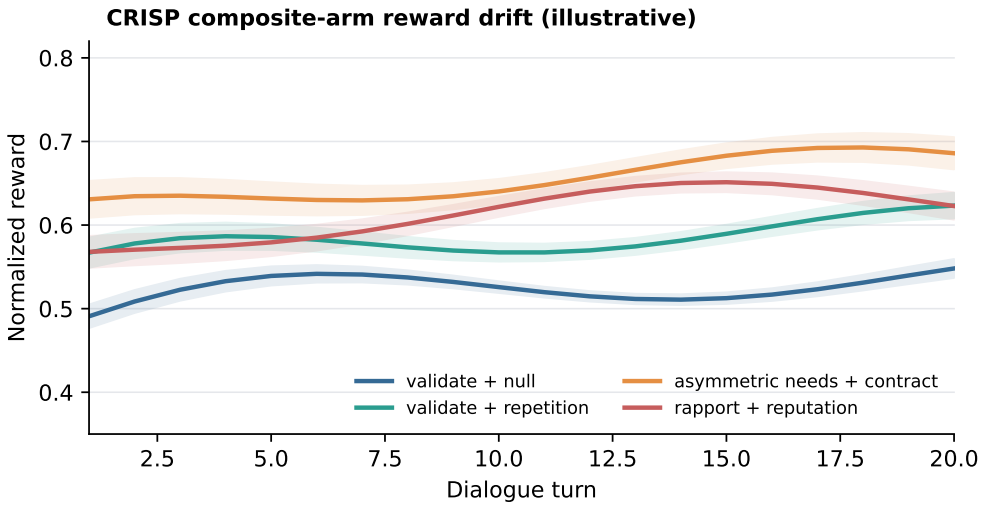

Figure 5: **Composite-arm reward drift.** Curves are deterministic templates awaiting run aggregates.

G.3 Bilateral Adaptation

We contrast bilateral CRISP with conditions in which only P1 or P2 updates. The arm bank and evaluator remain fixed; the unilateral condition updates one local controller only. The paired payoffs reveal whether mutual adaptation benefits both participants rather than transferring gains to the updating side.

G.4 Transfer and Heterogeneous Partners

We split Sotopia-Hard into disjoint training and test scenarios. The transferred surrogate is evaluated without weight updates and compared with an online-from-scratch learner. We also pair different frozen backbones to test whether the mechanism coordinate remains useful with heterogeneous partners.

Table 17: Robustness diagnostics with provisional values (†).

Setting	Goal ↑	Overall ↑
Scratch on unseen scenarios†	6.63	3.09
CRISP transfer†	**7.16**	**3.55**
Heterogeneous partners, Vanilla†	6.39	3.02
Heterogeneous partners, CRISP†	**6.88**	**3.39**

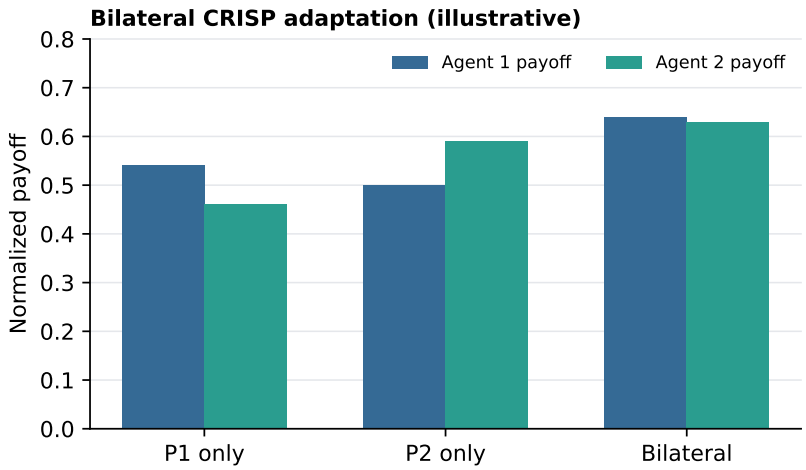

Figure 6: Unilateral and bilateral CRISP payoffs (illustrative).

G.5 Online Learning Dynamics

We will report per-turn reward, surrogate error, arm entropy, magnetization m_t, and estimated $T_c^{(t)}$ to test whether exploration concentrates near the critical manifold. The current multi-panel asset is a layout template.

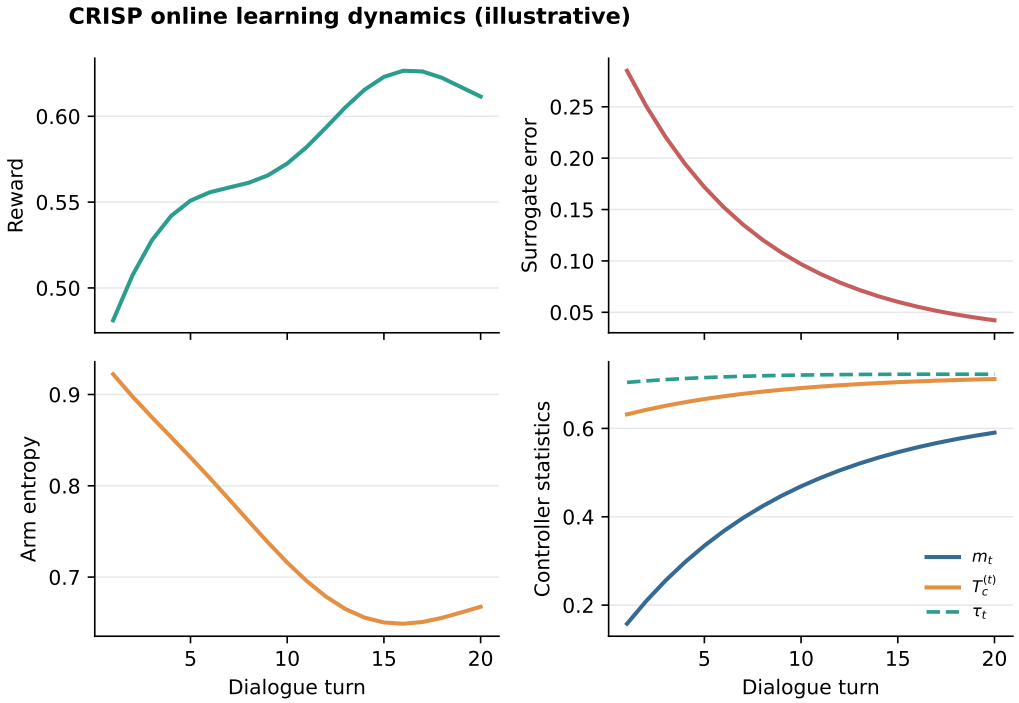

Figure 7: CRISP online learning dynamics (illustrative templates).