Social Gym and SPARTAN: Benchmarking and Improving LLM Social Reasoning via Multi-Agent Game Tournaments

Anonymous ACL submission

Abstract

Current evaluations for LLMs in multi-agent social interaction settings, where agents must negotiate and collaborate, lag behind other domains with objective ground-truth (e.g. math, logic), instead relying on human ratings or LLM-as-judge scoring, which are subjective, costly, and hard to verify at scale. To address this evaluation gap, we introduce **Social Gym**, an environment of 21 multi-agent social games where agent performance is verifiable and objective (e.g., Werewolves, Resistance, Spyfall, etc.) organized into five categories along the information-and-communication axes, with a Elo tournament that produces a cross-game leaderboard. Motivated by benchmarking experiments showing limitations of current LLM agents, we propose SPARTAN (**S**elf-**P**lay **a**nd **R**eflect-**Tran**sfer), a training-free self-improvement loop where a model plays a game, reflects on its trajectories to produce a transferable playbook, and applies that playbook in subsequent games. Our experiments on GPT-5-mini show that agents armed with a SPARTAN playbook gain on the structurally weaker side of asymmetric hidden-role games, and that the same playbook transfers across games and into weaker student models; the effect is capacity-dependent, however: an open-weights replication with Qwen3-32B is largely null except on Prisoner's Dilemma. Together, Social Gym and SPARTAN offer a reproducible, verifiable foundation for measuring and improving LLM social reasoning without weight updates.

1 Introduction

LLM-based agents are increasingly being developed and deployed in multi-agent settings: role-playing as autonomous inhabitants of social sandboxes (Park et al., 2023; Piao et al., 2025), collaborating on shared tasks through inter-agent dialogue (Li et al., 2023; Wu et al., 2023), and acting as persuaders or negotiators against other models (Bianchi et al., 2024). These deployments necessitate social interaction capabilities that go beyond single-turn question answering, such as navigating information asymmetry (Zhou et al., 2024a; YS et al., 2026), deception-utility tradeoffs (Su et al., 2025), negotiations (Cohen et al., 2026), and sustained role-playing (Shao et al., 2023; Wang et al., 2024a).

Yet, existing evaluations of LLM social reasoning suffer from two core limitations. First, many evaluations rely on static benchmarks such as theory-of-mind questionnaires (Kim et al., 2023), which produce reproducible scores but cannot test sustained multi-turn behavior. Second, more recent open-ended interactive evaluations (Zhou et al., 2024b) have been developed; those assess multi-turn interaction skills but rely on LLM-as-judge scoring, which suffers from position, verbosity, and self-enhancement biases (Zheng et al., 2023) and is inherently variable and subjective (Zhou et al., 2024b). Finally, recent works evaluate social interaction skills via verifiable rewards (Xu et al., 2023; Lan et al., 2024; Akata et al., 2025; Guertler et al., 2025), but each targets a single game in isolation, without a shared leaderboard for cross-game comparison, and so each covers only one aspect of social intelligence. A gap remains for evaluation that is simultaneously multi-turn, large domain coverage, and produces a *verifiable* signal, i.e., one determined by interaction rules rather than by subjective judgment.

To bridge this gap, we introduce **Social Gym** (Section 3), an environment of 21 multi-agent social games (Werewolves, Resistance, Spyfall, Prisoner's Dilemma, and others) spanning competitive, cooperative, and mixed-motive structures. Games like these offer a richer testbed than single-turn probes because they require sustained role-playing, coalition management, deception, and strategic information control across many turns and interactions. Importantly, the unified Elo tournament that produces per-game win-rate tables and a

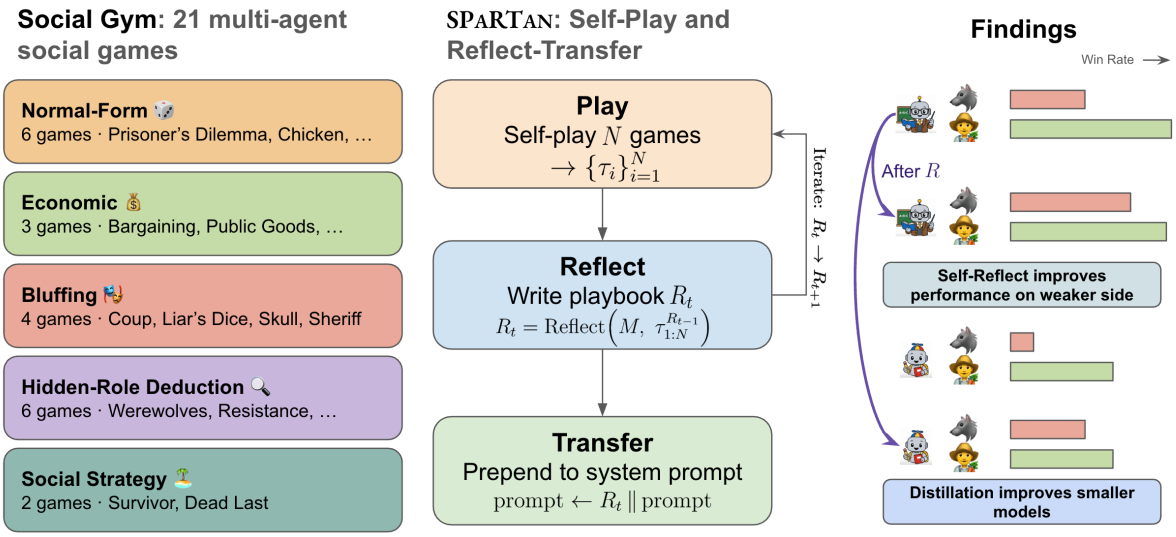

Figure 1: Overview of **Social Gym** (left), an environment of 21 multi-agent social games in five categories, and SPARTAN (center), a training-free play–reflect–transfer loop. **Right:** the learned playbook lifts the structurally weaker side of asymmetric games, both via self-reflection and via distillation to smaller models.

cross-game leaderboard delivers a *verifiable*, multi-turn measure of LLM performance across the full breadth of social-game categories, addressing all three gaps identified above.

We also propose SPARTAN (Section 4), a training-free self-improvement loop, answering the question: *can LLMs improve at social play without parameter updates?* In SPARTAN, a model (i) plays self-play games, (ii) reads its own trajectories and writes a transferable strategic playbook, and (iii) injects this playbook into its system prompt for subsequent games. SPARTAN is analogous to in-context fine-tuning, but the "training data" is the model's own gameplay and the "learned weights" are natural-language rules. Figure 1 shows **Social Gym** environment and SPARTAN method.

We benchmark seven models on Social Gym and document substantial per-game ranking inversions: top-ranked models underperform on specific games, and vice versa (Section 3). On the methodology side, we evaluate SPARTAN along four axes: within-game iterated reflection (R_1–R_4), cross-game transfer from one source game to held-out games, hold-one-out multigame transfer, and distillation of strong-model playbooks to weaker students. We find one regularity across all four setups: the playbook lifts the structurally weaker side of an asymmetric game and transfers across games and into weaker students, but the effect is capacity-dependent, vanishing for an open-weights model except on action-channel games like Prisoner's Dilemma.

2 Related Work

LLM evaluations of social intelligence. Social intelligence is widely treated as a multi-faceted construct that combines social knowledge with strategies for applying it (Kihlstrom and Cantor, 2000), and LLM evaluations of it currently split into three regimes that each cover a complementary part of the space but together leave a gap. Static probes such as FANToM (Kim et al., 2023), ToMi (Le et al., 2019), and BigToM (Gandhi et al., 2023) produce reproducible question-answering scores against a fixed ground truth, but reduce social reasoning to single-turn comprehension and cannot exercise sustained behavior; the limits of this format are visible in the ongoing debate over whether passing static ToM benchmarks reflects genuine mental-state attribution at all (Sap et al., 2022; Kosinski, 2023; Ullman, 2023).

Open-ended interactive evaluations, of which Zhou et al. (2024b)'s SOTOPIA framework and its SOTOPIA-π fine-tuning extension (Wang et al., 2024b) are representative, preserve multi-turn dynamics but score outcomes with LLM judges or human raters; LLM judges in particular carry documented position, verbosity, and self-enhancement biases (Zheng et al., 2023). Single-game LLM studies on Werewolf (Xu et al., 2023) and Avalon (Lan et al., 2024; Light et al., 2023) inherit verifiable game outcomes but each isolate one social dynamic; broader text-game environments (Guertler et al., 2025; Duan et al., 2024; Wu et al., 2024)

host many games but emphasize competitive strategy over the breadth of social-cognitive demands. Social Gym fills this gap with a multi-turn, rule-decided evaluation across 21 games organized along social-cognitive axes.

Self-improvement methods for LLMs. Self-improvement methods for LLMs broadly divide into two families. Prompt-only approaches have agents inspect their own outputs and revise, with Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023) as the canonical references, and have since been extended to skill-library construction, where an agent accumulates transferable natural-language strategies across episodes (Voyager Wang et al., 2023; ExpeL Zhao et al., 2024). Weight-update approaches such as SPIRAL (Liu et al., 2026) instead use multi-agent self-play with reinforcement learning to incentivize reasoning. SPARTAN sits in the prompt-only family and is closest in spirit to the skill-library line, but prior work in that line accumulates skills within a single domain (Minecraft for Voyager, individual reasoning tasks for ExpeL); to the best of our knowledge, no prior work studies whether such playbooks transfer across games, which is the question the iterated-rounds, multi-source, and strong-to-weak experiments in Section 5 are designed to answer.

LLM agents in multi-player games. LLM agents in multi-player games have historically been studied one game at a time. Bakhtin et al. (2022) achieve human-level Diplomacy with CICERO by coupling a language model to a strategic planner, demonstrating that strong play in a complex social game is possible but at the cost of heavy game-specific scaffolding. Akata et al. (2025) study LLM behavior on iterated 2×2 matrix games (Prisoner's Dilemma, Battle of the Sexes, Stag Hunt, Chicken), confined to canonical normal-form structures. Park et al. (2023) and Hagendorff (2024) probe individual behavioral skills, such as coherent role-playing and emergent deception, in isolated sandbox or single-task settings. The common limitation is breadth: each existing line covers a single game, a single equilibrium class, or a single skill, leaving open how the same model performs across the social-cognitive spectrum. We address this directly: Social Gym aggregates 21 games into one environment, and SPARTAN is tested for cross-game transfer.

3 Social Gym Benchmark

We design Social Gym around two principles: **(i)** every game must have an *algorithmically verifiable* outcome (win/loss, score, survival), providing the unambiguous reward signal R_v needed for both leaderboards and downstream RLVR training (Guo et al., 2025; Lewkowycz et al., 2022); and **(ii)** games must span the breadth of social intelligence, from atomic strategic primitives to long-horizon group deception, to expose distinct failure modes.

3.1 System Architecture

Social Gym extends the SOTOPIA environment loop (Zhou et al., 2024b) to support arbitrary N-agent interactions.

Finite State Machine (FSM) Engine. A flexible FSM engine handles complex phase transitions (e.g., *Night* → *Day* in Werewolves; *Discussion* → *Mission Vote* → *Mission Execute* in Resistance). Discrete state transitions also enable downstream RL value-function estimation.

Partial Observability. A visibility layer filters every message under one of three scopes: (i) **Public** (all alive agents, e.g., day discussion, vote results), (ii) **Team-Private** (faction members only, e.g., Werewolves see each other's night-phase kill votes), or (iii) **Private** (single agent, e.g., Seer inspections, role cards). This tested agents' Theory-of-Mind reasoning as they can only infer hidden states from permitted observations.

Game-Agnostic Engine. Each game implements a uniform interface (state, available actions, visible messages, reward function), so adding a new game requires only the game-specific FSM and reward, not changes to the core engine. Twenty-one games are implemented as extensions of this shared engine.

3.2 Game Suite

Our 21 games are organized along two orthogonal axes: *information structure* (complete information / hidden state / hidden roles) and *communication mode* (none / structured / free-form). This produces five categories, each probing a complementary facet of social intelligence, grounded respectively in strategic primitives (Axelrod and Hamilton, 1981; Schelling, 1980), prosocial behavior under collective-action problems (Ostrom, 1990;

Fehr and Gächter, 2000; Rosenthal, 1981; Rubinstein, 1982), deception production and detection (DePaulo et al., 2003; Wimmer and Perner, 1983), higher-order theory of mind (Perner and Wimmer, 1985; Baron-Cohen et al., 1985; Byrne and Whiten, 1988), and coalition/reputation tracking (Dunbar, 1998; Cialdini, 1984).

Normal-Form Games (6). Iterated matrix games with complete information and no communication. *Games:* Prisoner's Dilemma, Chicken, Battle of the Sexes, Stag Hunt, Minority Game, Rock-Paper-Scissors. The first four overlap with Akata et al. (2025)'s study of LLMs on repeated 2×2 matrix games; we extend the suite with Minority Game and Rock-Paper-Scissors and embed all six in a unified leaderboard.

Economic Games (3). Multi-round resource-allocation games requiring strategic reasoning and (optionally) negotiation. *Games:* Public Goods Game (free-riding vs. collective action), Centipede (sequential trust under growing stakes), Bargaining (structured negotiation over a divisible payoff).

Bluffing Games (4). Hidden-state games (no fixed factions) where agents must misrepresent or correctly infer private information. *Games:* Liar's Dice (probabilistic reasoning with hidden dice), Skull (placement bluffing with no communication), Coup (structured action-claim bluffing), Sheriff of Nottingham (free-form negotiation under inspection).

Hidden-Role Deduction (6). Games with hidden role assignments where agents must identify allies and enemies through dialogue. *Games:* Chameleon, Insider, Spyfall, Undercover, Resistance, Werewolves. This category extends prior single-game work on Werewolf (Xu et al., 2023) to a unified evaluation across the social-deduction family.

Social Strategy (2). Complete-information games where outcomes depend on alliance formation, persuasion, and reputation rather than hidden information. *Games:* Survivor (jury-voted finals), Dead Last (elimination + final-round split).

3.3 Measuring success via Elo Tournament

We report Elo-scale ratings (Elo, 1978) estimated by a regularized Bradley–Terry maximum-likelihood fit, following the LMSYS Chatbot Arena methodology (Chiang et al., 2024; Bradley and Terry, 1952). Tournament rosters are generated by enumerating all model combinations per game and running a fixed number of episodes per combination, with role and seat assignments balanced across episodes. Within each completed episode we extract pairwise outcomes from the final score vector and aggregate them into per-pair win/tie counts, skipping same-model pairs. In *free-for-all games* (Skull, Liar's Dice, Coup, Sheriff, Survivor, Dead Last, Public Goods, Stag Hunt, Minority Game), every cross-model agent pair (i, j) contributes one outcome, with the higher-scoring agent counted as the winner. In *2-team games* (Werewolves, Resistance, Spyfall, Chameleon, Insider, Undercover), only cross-team pairs contribute outcomes, using each team's shared score; same-team agents are not compared. Implementation details are in Appendix B. To capture asymmetric role performance in hidden-role games, we additionally report **Elo-Main** (majority/cooperative role: Villager, Civilian, Non-Spy) and **Elo-Alt** (minority/deceptive role: Werewolf, Spy, Insider, Chameleon, Undercover).

3.4 Leaderboard Results

We benchmark seven representative frontier and open-weights models: GPT-5-mini, GPT-4o, GPT-4o-mini, Qwen3-32B, Qwen3-4B, Qwen2.5-3B, and Gemma3-27B (Figure 2).

The overall leaderboard roughly tracks general capability rankings of these models (Figure 2). GPT-5-mini, the newest and strongest model in our slate, tops the leaderboard at 1110 Elo; Qwen2.5-3B, the smallest (3B parameters) and oldest open checkpoint, places last at 926.

While overall Elo and win rate (right panel of Figure 2) track general capability, the per-game Elos (left panel) show that **no model is uniformly strong: per-game rankings invert sharply.** The starkest case is Qwen3-32B, first on Chicken (1328) yet last on Werewolves (817); more broadly, top-ranked models have games where they fall below the 1000 anchor or behind much weaker peers, while the smallest model (Qwen2.5-3B) places near the top on others. These inversions reflect that different games reward fundamentally different behaviors, so a single scalar masks where each model actually succeeds; full per-game scores are in Figure 2. This motivates the need to examine per-game model scores rather than a single scalar.

Role-conditioned analysis. For hidden-role games, we additionally examine separate Elos for

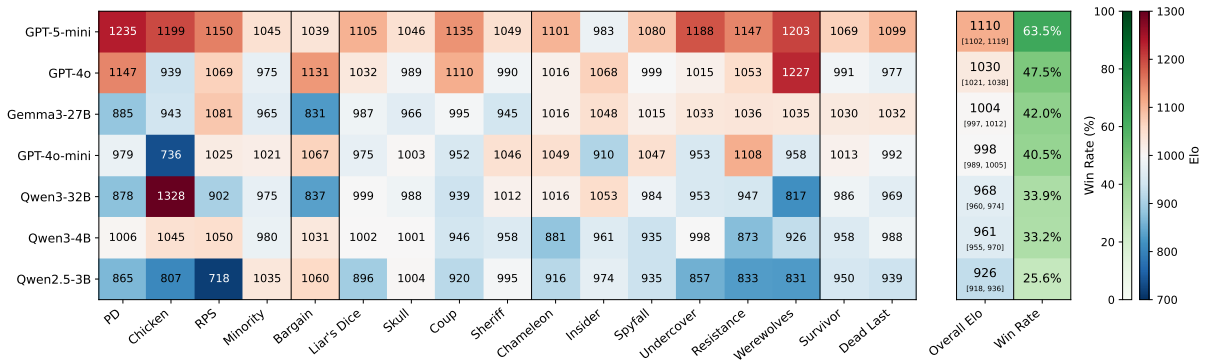

Figure 2: Social Gym results across the 17 competitive games. **Left:** per-game Elo, columns grouped by category (Normal-Form | Economic | Bluffing | Hidden-Role Deduction | Social Strategy, separated by black lines). **Right:** overall Elo (BT-MLE, anchored mean 1000; 95% bootstrap CIs) and win rate; rows sorted by overall Elo. Win rate counts strict wins only (ties in denominator, not numerator). Overall stats span 560+ episodes for each model.

the minority/deceptive role (Elo-Alt) and the majority/cooperative role (Elo-Main). Across the six hidden-role games, two patterns emerge in the per-model gap Δ = Elo-Alt − Elo-Main: in Werewolves and Chameleon the minority side overperforms for all six models, and in Spyfall and Undercover it underperforms for five of six. These games elicit a side preference that is independent of the player. In the remaining two games (Resistance, Insider), the sign of Δ varies across models, suggesting the gap is model-specific rather than structural. Per-game, per-model gaps are tabulated in Appendix G.

Qualitative observation: parroting effect in small models. Inspecting trajectories, we find Qwen2.5-3B frequently *parrots*, i.e., paraphrasing the previous speaker rather than producing an independent argument, which likely contributes to its last-place Overall Elo (926): agreeing with whoever spoke last is a near-zero-information move that gives the deceptive side cover. Excerpts and counts are in Appendix H.

4 SPARTAN: Self-Play and Reflect-Transfer

The leaderboard in §3.4 shows that no single model dominates Social Gym: every top-ranked model has games where it underperforms peers, yet per-game ranks may invert. This motivates our next research question: can a model close its own per-game gaps without weight updates, by inspecting its own gameplay and extracting reusable strategies? To answer this, we introduce SPARTAN, a simple training-free self-improvement loop with three stages: play, reflect, and transfer.

SPARTAN method. Our method consists of the following three stages (illustrated in Figure 1):

1. **Play.** The model M plays N self-play games of game G, producing trajectories $\{\tau_1, \ldots, \tau_N\}$.

2. **Reflect.** The model is shown its own trajectories along with the final outcomes (win/loss per role) and asked to write a first-person strategic playbook covering deception, detection, persuasion, information management, coalition dynamics, and timing. The model is instructed to keep the playbook game-agnostic (no references to specific game numbers). We denote this as $R_1 = \text{Reflect}(M, \tau_{1:N})$.

3. **Transfer.** The reflection is prepended to an agent's system prompt for subsequent games.

Application setups. To test the effectiveness of SPARTAN, we select our strongest LLM (GPT-5-mini) and examine four different evaluation setups: (i) within-game iterated reflection, (ii) one-source → many-target cross-game transfer, (iii) many-source → one-target multigame transfer, and (iv) strong-to-weak distillation into 6 student models on Resistance.

5 SPARTAN Experiments and Results

We evaluate SPARTAN on the asymmetric hidden-role games from Section 3, asking whether iterated self-play reflection can improve a model's win rates without parameter updates. Following the role-conditioned convention from §3.3, we use *alt* for the minority/deceptive role of a game (e.g., the

Game	Side	BL/R_0	+R_1	+R_2	+R_3	+R_4
Werewolves	Alt	23	50	46	**63**	46
	Main	77	50	54	67	44
Spyfall	Alt	20	36	**46**	26	23
	Main	80	80	67	80	70
Resistance	Alt	30	23	16	**40**	36
	Main	70	67	57	54	74
Avg alt		24	36	36	**43**	35
Avg main		76	66	59	67	63

Table 1: GPT-5-mini within-game iterated self-reflection: R-armed side win rate (%, n=30/cond) when R_n is injected on the alt side vs. on the main side (alt/main as defined in §3.3). The alt side rises and the main side falls across rounds, peaking at different n per game (alt: Werewolves R_3, Spyfall R_2, Resistance R_3).

werewolves team Werewolves, the Spies in Resistance, etc.) and *main* for the majority/cooperative role. In vanilla GPT-5-mini self-play the alt role's win rate is consistently lower than the main role's on Werewolves, Spyfall, Undercover, and Resistance (Table 5), establishing an imbalance in the vanilla baseline that motivates testing whether SPARTAN can lift the weaker side. Unless noted, every condition uses n=30 games, giving binomial 95% CIs of $\approx \pm 18$ pp.

We probe how SPARTAN affects model performance through within-model reflection on a strong model (GPT-5-mini; §5.1), through within-model reflection on an open-weights model (Qwen3-32B; §5.3), and through across-model distillation (§5.2). We hypothesize that LLM agents armed with a SPARTAN playbook will gain on the structurally weaker side of an asymmetric game.

5.1 Same-model reflection (flagship)

We evaluate SPARTAN when the model that generates the reflection also consumes it, with GPT-5-mini as the primary model throughout this subsection; supporting evidence from GPT-5 and Gemini self-play is reported in Appendix I.

Within-game iterated R_1–R_4. We evaluate on three hidden-role deduction games: Werewolves, Spyfall, and Resistance, chosen because both sides have headroom in the vanilla baseline and the alt side is the structurally weaker one (alt baselines 23%, 20%, 30% respectively). GPT-5-mini generates an iterated chain of self-reflection playbooks R_1, R_2, R_3, R_4, where $R_n = \text{Reflect}(M, \tau_{1:N}^{R_{n-1}})$ distills N=30 self-play games played under the previous round's playbook (R_0 denotes vanilla). Each R_n is injected on either the alt side or the main side, with vanilla GPT-5-mini on the other; 30 games per condition.

Result. Across rounds, the R-armed side trades win rate with the vanilla side: averaged over the three games, the alt side rises from 24% baseline to 36, 36, 43, 35% under $R_{1\text{–}4}$ (peak at R_3), while the main side falls from 76% to 66, 59, 67, 63% (worst at R_2; Table 1). Per-game peaks are non-monotonic and game-specific: Werewolves and Resistance peak on the alt side at R_3 (63%, 40%), Spyfall at R_2 (46%).

Interpretation. Self-reflection raises the weaker side and drops the stronger side; the optimal number of reflection rounds varies by game. Contrary to iterated-reflection methods that assume more rounds yield more gain (Shinn et al., 2023; Madaan et al., 2023), the lift in our setting concentrates at R_1, and additional rounds redistribute rather than accumulate.

Cross-game transfer ($1 \rightarrow n$). We use Werewolves, Spyfall, Chameleon, Undercover, and Resistance as both source and target games. For each game as source, we inject the source's R_1 playbook into one side of each of the other four as target, vs. vanilla GPT-5-mini. We report the Δ (in pp) against the target's vanilla self-play baseline, separately for alt and main injection. The Chameleon target row is degenerate (its alt-side baseline is already at 100% in vanilla GPT-5-mini self-play; see footnote below); we retain it in the heatmap for completeness rather than dropping it silently.

Result. The two heatmaps in Figure 3 are sign-flipped: excluding the saturated Chameleon target column, alt-side injection skews positive (median +7, max +27) and main-side injection skews negative (median −7, four cells below −20).

Interpretation. The cross-game pattern matches the within-game finding: R_1 helps the disadvantaged side and either has no effect or actively hurts the advantaged side. The match is striking because the source playbook was generated on a different game, so any useful content is not target-specific.

Multigame transfer ($n \rightarrow 1$). For K source games, the multigame playbook is $R_{\text{multi}} = \text{Reflect}(M, \tau_{1:N}^{G_1}, \ldots, \tau_{1:N}^{G_K})$. We construct three multigame playbooks of increasing breadth: R_{wcs} (Werewolves + Chameleon + Spyfall), R_{wcsu} (+ Undercover), R_{wcsur} (+ Resistance).[1] We evaluate

[1] GPT-5-mini's vanilla Chameleon alt-side (Chameleon role) win rate is saturated at 100%, so the Chameleon panel

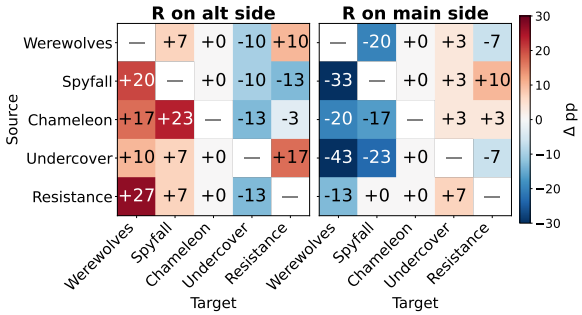

Figure 3: GPT-5-mini cross-game transfer (Δ pp vs. vanilla self-play baseline, n=30/cell) over five games (Werewolves, Spyfall, Chameleon, Undercover, Resistance). Left: R on the target's alt side. Right: R on the target's main side (alt/main as defined in §3.3). The alt panel is dominated by positive cells; the main panel is dominated by large negative cells. The Chameleon target column is structurally zero because the alt baseline saturates at 100%.

each on its in-distribution targets and one held-out target (except R_{wcsur}, where the target is in the training playbook). We compare against (a) baseline and (b) the target's own Single-R_1 from the within-game results above.

Result. Multigame does not stack (Figure 4). On three of the four non-saturated targets (Werewolves, Spyfall, Undercover), broader source sets either do not beat the target's own Single-R_1 or degrade it; Chameleon stays pinned at the 100% alt-side ceiling under every condition; the only exception is held-out Resistance, where R_{wcsu} lifts the alt side from 23% (Single-R_1) to 47%.

Interpretation. Learning from multiple training games does not improve transfer to a new game beyond what a single related training game already provides; the only exception, Resistance, is also the held-out game with the most baseline headroom on the alt side, consistent with the lift-the-weaker-side pattern from the within-game and cross-game results.

5.2 Cross-model distillation (sub-flagship)

Setup. In the distillation setting, a playbook generated by a strong model M_{strong} is injected into a weaker student model M_{weak}. We test whether the held-out multigame playbook from §5.1, R_{wcsu} (trained on Werewolves + Chameleon + Spyfall + Undercover by M_{strong}= GPT-5-mini, Resistance

in Figure 4 and the Chameleon target column in Figure 3 are at the ceiling for all conditions and carry no signal in the alt direction; we report them rather than dropping them silently. Chameleon's self-play trajectories still contribute hidden-role deduction patterns to the source mix.

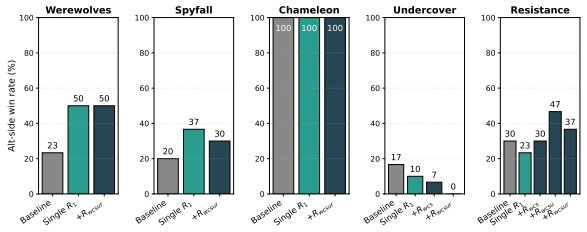

Figure 4: GPT-5-mini multigame transfer: alt-side win rate (%, n=30/condition) on five target games. Single-R_1 uses the target's own within-game playbook; R_{wcs}, R_{wcsu}, R_{wcsur} are multigame playbooks of increasing breadth. Resistance is held out from R_{wcs} and R_{wcsu}. The Chameleon panel sits at the 100% alt-side ceiling under vanilla self-play and remains there under every R.

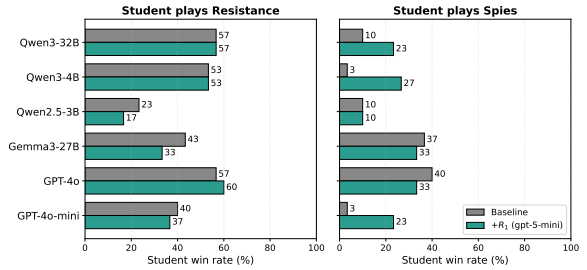

Figure 5: GPT-5-mini R_{wcsu} (held-out multigame, trained on W+C+S+U) injected into six student models on Resistance (n=30/condition vs. vanilla GPT-5-mini opponent). Left: student plays the main (Resistance, majority/cooperative) side. Right: student plays the alt (Spies, minority/deceptive) side, which is structurally weaker in vanilla baseline; three students (Qwen3-32B, Qwen3-4B, GPT-4o-mini) gain +13 to +24 pp.

excluded), transfers when injected into weaker student models. Six students (Qwen3-32B, Qwen3-4B, Qwen2.5-3B-Instruct, Gemma3-27B, GPT-4o, GPT-4o-mini) each play one side of Resistance against vanilla GPT-5-mini, with and without the playbook.

Result. Distillation reproduces the same side asymmetry across both transfer axes (Figure 5). On the disadvantaged Spies side, three students gain +13 to +24 pp from the shared R_{wcsu} playbook; on the favored Resistance side, every student moves by at most ±5 pp. Two students (Qwen2.5-3B, Gemma3-27B) do not gain on either side.

Interpretation. The same playbook produces side-dependent rather than student-dependent effects: it lifts whichever student is playing the structurally weaker role and leaves the other alone. Combined with the multigame result in §5.1, the same R_{wcsu} playbook now lifts the underperforming side along three setups (within-model held-out target,

across-model held-out target, both at once), all without ever having seen Resistance during reflection.

5.3 Open-weights model replication

We test whether the patterns from Sections 5.1 and 5.2 carry over to an open-weights model. We use Qwen3-32B throughout, mirroring §5.1's within-model setup in §5.3.1 and §5.2's cross-model setup in §5.3.2.

5.3.1 Self-reflection (Open-weights model)

Setup. Following the within-game protocol from §5.1, we run Qwen3-32B as the only model (reflection generator, training self-play, and evaluation opponent) across six games: the three hidden-role games from §5.1 (Werewolves, Spyfall, Resistance) plus Chameleon, Undercover, and Prisoner's Dilemma. PD is included as a pure action-channel game (no chat phase) for contrast with the discussion-heavy games. We generate iterated playbooks R_1–R_4 from Qwen3-32B self-play and report the R-armed-side win rate per round.

Result. Only Prisoner's Dilemma shows a clean R-armed-side lift ($13\% \rightarrow 58\%$ at R_1, $+45$pp). Resistance shows a smaller positive effect ($+34$pp on the Resistance side). The other four games (Werewolves, Chameleon, Spyfall, Undercover) are flat across all four iterated rounds and across both within-game and cross-game source playbooks. Full per-game R_1–R_4 and cross-game transfer tables are in Appendix K.

Interpretation. The open-weights replication is largely a null result, with PD as the only clean exception. We attribute this to model capacity: at 32B parameters Qwen3-32B appears incapable of learning from the trajectories of long, complex social-deduction games, except when the required action collapses to a single discrete token (PD's defect, Resistance's private succeed/fail vote). One concrete symptom is that Qwen3-32B's discussion-phase outputs frequently parrot or paraphrase the immediately prior speaker rather than producing independent content. The model is in fact self-aware enough to identify this behavior, and iterated reflection codifies it into the playbook itself (examples in Appendix K.4), but the model does not eliminate the parroting in subsequent play.

5.3.2 Distillation (Open-weights model)

We then test the distillation setup from §5.2 with Qwen3-32B as the teacher. The source playbooks are Qwen3-32B's R_1 for PD and Resistance (the two action-channel games where Qwen3-32B's own self-reflection lifted the alt side; see §5.3.1). We test transfer to Qwen3-4B, Qwen2.5-3B, and Gemma3-27B against vanilla Qwen3-32B (30 games per side per condition).

Result. PD shows clean positive distillation across all three students ($+17$ to $+77$pp with the reflection applied); Resistance is mixed (Qwen2.5-3B $+10$pp, Qwen3-4B -20pp, Gemma3-27B -10pp on the Resistance side). Per-student numbers are in Appendix K. Distillation reinforces §5.3.1: the action-channel game (PD) transfers cleanly to smaller students, while the discussion-heavy game (Resistance) does not.

6 Conclusion and Discussion

We presented Social Gym, an environment of 21 multi-agent social games organized into five categories (normal-form, economic, bluffing, hidden-role deduction, social strategy), with a unified Elo leaderboard that reveals non-monotonic ranking by parameter count, large per-game role inversions, and qualitative behavioral jumps. We then introduced SPARTAN, a training-free self-improvement loop, and evaluated it across within-model iteration, cross-game transfer, and cross-model distillation. A single regularity emerges from all three perspectives: the playbook lifts the structurally weaker side of an asymmetric game.

Our findings yield two main implications. *First*, Social Gym and SPARTAN jointly show that LLM social ability is not a single scalar capability: model rankings, role advantages, and reflection gains all depend strongly on the interaction structure of the game. By combining a broad game suite with targeted playbook interventions, we can separate structural properties of a social setting from model-specific failures such as weak deception, poor coalition tracking, or parroting behavior. *Second*, Social Gym provides a natural testbed for future work to explore reinforcement learning with verifiable rewards (RLVR): every episode produces an objective, rule-computed outcome while still requiring rich language-based interaction. This would enable future research to train and evaluate social reasoning skills at scale without LLM judges, while also testing whether learned strategies transfer across cooperation, negotiation, bluffing, and hidden-role deduction games.

Limitations

Sample size per condition is modest. We use 30 evaluation games per condition, giving binomial 95% CIs of roughly ± 18 pp for a single-coin observation. Several effects we report (e.g., the $+10$pp Resistance gain) are within this range and should be replicated at larger sample sizes before strong conclusions are drawn.

No placebo-playbook control. We compare R-armed players against vanilla opponents but not against opponents armed with a content-matched placebo (scrambled or unrelated text of equal length). Without this control we cannot fully disentangle playbook-content effects from generic prompt-perturbation effects, though the structural patterns reported in §5.3 (action-channel vs. free-discussion games) and the Undercover monotonic regression toward 0% argue against a pure prompt-perturbation reading.

The reflection is constrained to natural-language prose. SPARTAN does not allow the model to update tools, retrieve external knowledge, or perform structured reasoning beyond what fits in the system-prompt text. Methods that combine reflection with retrieval or scratchpads may exhibit qualitatively different transfer behavior.

Ethics / Broader Impacts

Social Gym and SPARTAN measure and, in some settings, improve capabilities: deception, persuasion, coalition manipulation. Thus, it may carry dual-use risk if transferred from games to real interactions involving humans. We note three mitigating factors: all experiments are confined to fully synthetic multi-agent games with no human subjects; the improvements are training-free, modest in size, and largely null for open-weights models; and the verifiable-reward framing is intended primarily as an evaluation tool for diagnosing such capabilities rather than a recipe for deploying manipulative agents. We release code to support reproducible measurement of these behaviors, and discourage use of the playbook-distillation procedure in adversarial human-facing applications.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390.

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

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiation-arena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 3935–3951.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Richard W. Byrne and Andrew Whiten, editors. 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Clarendon Press, Oxford.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *International Conference on Machine Learning*, pages 8359–8388. PMLR.

Robert B. Cialdini. 1984. *Influence: The Psychology of Persuasion*. William Morrow.

Myke C. Cohen, Mingqian Zheng, Neel Bhandari, Hsien-Te Kao, Xuhui Zhou, Daniel Nguyen, Laura Cassani, Maarten Sap, and Svitlana Volkova. 2026. Imperfectly cooperative human-ai interactions: Comparing the impacts of human and ai attributes in simulated and user studies. In *Findings of ACL*.

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin*, 129(1):74.

J Duan, R Zhang, J Diffenderfer, B Kailkhura, L Sun, E StengelEskin, and 1 others. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. arxiv. *arXiv preprint arXiv:2402.12348*.

Robin I. M. Dunbar. 1998. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub.

Ernst Fehr and Simon Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Leon Guertler, Bobby Cheng, Simon Yu, Bo Liu, Leshem Choshen, and Cheston Tan. 2025. Textarena. *Preprint*, arXiv:2504.11442.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121.

John F. Kihlstrom and Nancy Cantor. 2000. Social intelligence. In Robert J. Sternberg, editor, *Handbook of Intelligence*, 2 edition, pages 359–379. Cambridge University Press.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4(169):2.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *arXiv preprint arXiv:2310.14985*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. 2026. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *Preprint*, arXiv:2506.24119.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

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

Josef Perner and Heinz Wimmer. 1985. "john thinks that mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3):437–471.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Robert W Rosenthal. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic theory*, 25(1):92–100.

Ariel Rubinstein. 1982. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3762–3780.

Thomas C Schelling. 1980. *The Strategy of Conflict: with a new Preface by the Author*. Harvard university press.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 13153–13187.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. 2025. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. In *NAACL*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

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. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

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

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Yue Wu, Xuan Tang, Tom Mitchell, and Yuanzhi Li. 2024. Smartplay: A benchmark for llms as intelligent agents. In *International Conference on Learning Representations*, volume 2024, pages 1543–1561.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Yashwanth YS, Ruichen Wang, Shihua Zeng, Xuhui Zhou, Koichi Onoue, Vasudha Varadarajan, and Maarten Sap. 2026. Sotopia-tom: Evaluating information management in multi-agent interaction with theory of mind.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. In *EMNLP*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A Code and Data Release

All code is released at `https://anonymous.4open.science/r/Social_Gym_Spartan-D788/`.[2] The repository contains: (i) the game engine layer and the 21 game implementations from Section 3, (ii) the Elo tournament infrastructure with the Bradley–Terry fit described in Appendix B, and (iii) the SPARTAN pipeline scripts for roster generation, reflection generation, and per-condition evaluation.

Dependencies and licensing. The bundled `sotopia/` directory is a snapshot of an open-source dependency released under the MIT license; our use is consistent with its terms, and the contribution claimed here is limited to the game-engine layer, the games, and the experiments (see `NOTICE.md` in the repository). Our own code, game implementations, and configuration files are released under the MIT license. The proprietary models we evaluate (GPT-5, GPT-5-mini, GPT-4o, GPT-4o-mini, Gemini 3.1 Pro) are accessed through their providers' APIs under the respective terms of service, and the open-weights models (Qwen3, Qwen2.5, Gemma3) are used under their published licenses.

[2] Anonymous mirror for ARR review; the non-anonymized repository will be linked in the camera-ready version.

B Elo Aggregation Details

Bradley–Terry fit. The aggregated per-pair win/tie counts (§3.3) are fit via L2-regularized logistic regression (C=0.1, scale 400, anchored mean 1000), and 95% confidence intervals are obtained by multinomial bootstrap on the battle counts (n_{boot}=100).

Per-game vs. overall ratings. The overall Elo (right panel of Figure 2) and per-game Elos (left panel of Figure 2) are computed by independent BT fits, each anchored at a mean of 1000. Overall aggregates outcomes across the union of competitive games (560+ episodes per model, hundreds of pairwise outcomes), while each per-game fit only sees ∼30 episodes per model. Per-game ratings therefore have larger uncertainty and stay closer to the 1000 anchor, while Overall reflects the better-determined skill estimate.

Why BT-MLE rather than the online K-factor update. We adopt BT-MLE rather than the classical online Elo update (Elo, 1978) for three reasons. First, BT is path-independent: the rating does not depend on the order in which episodes are processed. Second, with ∼30 episodes per pair, the online update would not converge. Third, BT is the standard for current LLM leaderboards (Chiang et al., 2024).

Same-score pairs. Many of our games emit ternary $\{+1, 0, -1\}$ scores by bucketing players (top-half vs. bottom-half in Public Goods and Sheriff; winner-take-all in Skull, Liar's Dice, and Survivor). In such games, two co-winners or two co-losers share a score not because they competed and tied, but because they were assigned to the same outcome bucket by design. Treating these as draws ($s = 0.5$) would otherwise cap the Elo of strong models, since a hypothetically dominant model would always tie with any peer who also reached the top bucket. We therefore exclude same-score pairs from the Bradley–Terry fit entirely.

Cooperative games. For cooperative games where rankings are ill-defined we report normalized win rates below, instead of Elo.

C Game-Play Prompt

Each agent receives a prompt at every turn it is asked to act. The template is shared across all 21 games; per-game variation lives in the `description`, action vocabulary, and format instructions. When an agent's roster entry has `include_reflection: true`, the playbook generated by the reflection step (Appendix D) is loaded once and prepended via the `{reflection}` slot; otherwise the slot is empty.

Model	BoS	Stag Hunt	Public Goods	Centipede
GPT-5-mini	86.7	62.6	66.7	6.7
GPT-4o	28.6	47.0	50.0	57.1
GPT-4o-mini	37.1	12.5	38.9	45.7
Gemma3-27B	40.0	16.0	53.1	42.9
Qwen3-32B	37.1	22.2	2.5	20.0
Qwen3-4B	34.3	1.7	5.0	34.3
Qwen2.5-3B	25.7	42.7	23.9	51.4

Table 2: Cooperative-game win rates (%). BoS = Battle of the Sexes. Win rate counts strict wins only (achieving the cooperative payoff defined for each game's mechanic); ties are in the denominator but not the numerator. Per-game sample sizes range from $n = 30$ (Battle of the Sexes, Centipede) to n=54 (Stag Hunt, Public Goods on the open-weights tier).

```
{reflection}

Imagine you are playing the game as
{agent}.

Here is the description of the game:
{description}

Your ({agent}'s) goal: {goal}
{secret}

Here is the context of the interaction:
{history}

Your available action type(s):
[{action_list}].
{action_instructions}

Please only generate a JSON string
including the action type and the
argument.
Your action should follow the given
format:
{format_instructions}
```

Slot semantics.

- `{reflection}`: full text of the playbook (e.g. R_t for some round t), or empty for the vanilla baseline.
- `{agent}`: the agent's display name (e.g. "Stephen").
- `{description}`: the game-specific scenario string (e.g. Werewolves' phase rules and win conditions).
- `{goal}`: the role-conditioned goal (e.g. "Identify werewolves" for a Villager).
- `{secret}`: any private information the role is given at game start (Werewolf identities,

Spyfall location, etc.); empty for roles without secrets.

- `{history}`: visible message log filtered by the partial-observability layer (see §3.1).
- `{action_list}`, `{action_instructions}`, `{format_instructions}`: the actions available in the current FSM state and the JSON schema the agent must output.

D Reflection Prompt

The reflection model is given the full trajectories of N self-play games and asked to produce a first-person strategic playbook. For iterated rounds ($t \geq 2$) the prompt additionally includes the previous-round playbook R_{t-1} and asks the model to revise rather than rewrite. The exact prompt template (omitting the bulleted list of capability axes for brevity) is:

> `{prior_section}`
> Below are the full trajectories of all {num_games} games, showing every player's actions and the outcome:
> `{game_summaries}`
>
> Write an **internal monologue** of transferable social reasoning skills. Frame your insights around general capabilities that apply across many social games, such as: Deception, Detection, Persuasion, Information management, Coalition dynamics, Timing and patience.
>
> **Requirements:**
> - Write in first person ("I should. . . ", "When I need to hide information. . . ", "A pattern I noticed is. . . ")
> - Derive insights from the games above, but write the rules so they apply beyond any specific game
> - Focus on actionable lessons, not abstract observations
> - Do NOT reference specific game numbers (e.g., "Game 3", "Games 5-8"). Your future self will not have access to these transcripts, so such references would be meaningless
>
> This monologue will be prepended to your system prompt in future social games. Write it so that reading it once before any social strategy game will meaningfully improve your play.

The `prior_section` for $t = 1$ is:

> You just played {num_games} games across the following game(s): {games_desc}. Each player was an independent instance of you and only saw its own role's private information; the transcripts below reveal the hidden moves of every role, but the players themselves did not have this view during play.

For $t \geq 2$ it instead reads:

> You previously wrote the following strategic playbook for yourself: {prior_reflection}. You then played {num_games} more games using this playbook across: {games_desc}. Based on these new games, revise your playbook. Keep rules that worked, remove or modify rules that didn't help, and add new insights. Output the complete revised playbook (not just the changes).

Design choices. (i) The first-person framing is meant to encourage the model to treat the playbook as advice to itself, which empirically produces more actionable rules than third-person observations. (ii) Forbidding game-number references prevents the model from writing rules that would be uninterpretable at deployment, when only the playbook (not the source transcripts) is available.

E Reflection Examples

To give a sense of how the reflection content evolves across iterations, we include excerpts from the GPT-5-mini and GPT-5 Werewolf playbooks.

E.1 GPT-5-mini Werewolf Playbooks

To illustrate the relative stability of GPT-5-mini's iterated reflection (consistent with the within-game finding in §5.1 that the lift concentrates at R_1), we reproduce the Deception section of the Werewolves playbook across four rounds. The bullet structure is preserved across rounds; later rounds tighten wording and add tactical refinements rather than restructuring strategy.

R_1 *"...Deception: how I lie and stay credible*
- I should pick a single, believable persona and commit to it. Every action (what I say, when I speak, how I vote) must fit that persona. Small contradictions are fatal; plan my story so it explains the behavior I will need later.
- When I bluff, I should combine one verifiable true fact with the lie. Mixing a small truthful observation into my statement makes the whole claim feel anchored and increases believability.
- I should avoid over-detailing fabrications. Plausible vagueness is better than a wrong precise detail

that can be disproved.
- I should time my fabrications to the game's incentives: lie early only if necessary (to survive or to seed a plausible alternate narrative); lie late only if I have a clear plan to leverage the deception into a win.
- I should never retroactively invent reasons for past actions unless I can plausibly frame them as honest uncertainty. If I must justify earlier behavior, I will do so with motives consistent with my persona (e.g., "I listened to gather info" rather than "I was asleep").
- When I adopt a risky public role claim, I must be ready to prove it through behavior (predicting a future fact, coordinating with known events) or accept the likely target it creates.
Detection: finding liars and inconsistencies..."

R_2 **(revising R_1).** Structure preserved. Adds an explicit "exit plan if exposed" to the timing rule; renames the truthful-fragment technique as the "anchor."

R_3 **(revising R_2).** Adds: "Don't over-prepare quotes a liar could plausibly mimic — pair verbatim claims with contextual details." Sharpens the role-claim rule into an explicit (a)/(b) commitment.

R_4 **(revising R_3).** Adds: "When bluffing a role, consider offering a low-cost, verifiable trade (e.g., 'I'll reveal Night 1 result if you commit to protect me') to buy survivability." Otherwise wording-level edits.

E.2 GPT-5 Werewolf Playbooks

R_1 **(after self-play, no prior).** "*Build a consistent, proactive persona from the start. Pre-commit to falsifiable stances (top suspect, backup, 1 town-read) and stick to them unless new info arrives. Calibrate specificity: early/first, stay moderately specific; later, add a unique but safe detail. Avoid 'heads-I-win' frames.*"

R_2 **(revising R_1).** "*Hunt for agenda over solve: pre-setting easy miselims, parking 'placeholders' without intent to move, echoing popular takes without new reasons, or misrepresenting others' words. Compare process vs. action: if someone talks consolidation but seeds multiple soft outs, that's a tell. Watch vote timing.*" R2 visibly shifts to detection/process focus.

R_3 **(revising R_2).** "*Score claims by: timing (proactive vs. reactive), specificity (who/when), target rationale (why them before), and fit with public events. Mechanics > talk. Track talk-to-vote: pushing A all day but voting B without new receipts is high-signal.*" R3 begins to combine detection rigor with mechanical scoring rather than push purely on persona consistency or behavior, foreshadowing the balanced playbook of R_4.

F Game Suite Details

Table 3 summarizes the 21 games along the axes used to define our five categories: number of players, information structure, communication mode, iteration horizon, and primary skills exercised. Full FSM specifications, payoff matrices, and reward functions are released with the code repository alongside the unified game engine.

Modifications from published versions. A few of our commercial-game implementations adopt a slightly modified version of the base game, typically to keep the focus on social reasoning and avoid bookkeeping that the engine would have to track but that does not exercise additional social skills. We note these here for clarity:

- **Coup** (Tahta, 2012). Block claims resolve automatically rather than being themselves challengeable, and the *exchange* action is streamlined so that the Ambassador returns the drawn cards instead of choosing 2 of 4 to keep. The full block matrix and successful-challenge card swap-back are preserved.

- **Skull** (Marly, 2011). A round win ends the game (`wins_needed = 1`) rather than 2, shortening the match while keeping the place–bid–flip cycle intact.

- **Sheriff of Nottingham** (Halaban & Zatz, 2014). Goods are summarized as a binary *honest/smuggle* choice with fixed payoffs; the inventory layer of the published game (multiple legal goods, royal-goods bonuses) is folded into a single payoff parameter so that the negotiation/inspection dynamic is what drives play.

- **Insider** (Oink Games, 2016). The environment answers yes/no questions directly, replacing the published *Master* role; the Insider role itself (a Citizen who covertly knows the word and steers questioning) is unchanged.

Game	Category	Pl.	Information	Comm.	Iteration	Skills
Prisoner's Dilemma	Normal-Form	2	Complete	None	5 rounds	Strategic, Cooperation
Chicken	Normal-Form	2	Complete	None	10 rounds	Strategic, Coordination
Battle of the Sexes	Normal-Form	2	Complete	Free-form	10 rounds	Strategic, Coordination
Stag Hunt	Normal-Form	4	Complete	None	10 rounds	Strategic, Cooperation, Coordination
Minority Game	Normal-Form	5	Complete	None	12 rounds	Strategic, Coordination
Rock-Paper-Scissors	Normal-Form	2	Complete	None	10 rounds	Strategic
Public Goods	Economic	4	Hidden state	None	10 rounds	Strategic, Cooperation
Centipede	Economic	2	Complete	None	4 rounds	Strategic, Cooperation
Bargaining	Economic	2	Complete	None	10 rounds	Strategic, Negotiation
Liar's Dice	Bluffing	3	Hidden state	Structured	Until elim.	Deception, Probabilistic
Skull	Bluffing	4	Hidden state	None	1 round	Deception, ToM
Coup	Bluffing	4	Hidden state	Structured	Until elim.	Deception, ToM
Sheriff of Nottingham	Bluffing	4	Hidden state	Free-form	4 rounds	Deception, Negotiation, Probabilistic
Chameleon	Deduction	5	Hidden roles	Free-form	1 round	Deception, Persuasion, ToM
Insider	Deduction	5	Hidden roles	Free-form	1 round	Coordination, Deception, ToM
Spyfall	Deduction	4	Hidden roles	Free-form	Until elim.	Deception, Persuasion, ToM
Undercover	Deduction	6	Hidden roles	Free-form	Until elim.	Deception, Persuasion, ToM
Resistance	Deduction	5	Hidden roles	Free-form	$\leq$5 missions	Strategic, Deception, ToM
Werewolves	Deduction	6	Hidden roles	Free-form	Until elim.	Strategic, Deception, ToM
Survivor	Social Strategy	6	Complete	Free-form	Until 2–3 left	Negotiation, Persuasion, ToM
Dead Last	Social Strategy	6	Complete	Free-form	Until 2–3 left	Negotiation, Persuasion, ToM

Table 3: Game taxonomy across the 21 games in Social Gym. *Pl.*: number of players; *Comm.*: Communication (Format); *ToM*: Theory of Mind. "Complete" information means no private state at game start; "Hidden state" means each agent has private state (cards, dice) but no factional roles; "Hidden roles" means agents are assigned secret allegiances at game start. *Cooperation* skills involve overcoming the temptation to defect for collective benefit (PD-style); *Coordination* skills involve aligning on one of multiple equilibria (Chicken/BoS-style). The *Iteration* column uses each game's natural unit: *rounds* (one independent play of the base game; for matrix games each round is one simultaneous move per player, for Centipede each round is one full traversal of the take-or-pass tree), *missions* (Resistance has up to 5 mission proposals), and *Until elim.* (game ends when a win condition is met—e.g., one team eliminated, last survivor remaining).

- **Spyfall** (Ushan, 2014). The round ends by vote rather than by a real-time clock, and the Spy does not pre-empt the vote with a mid-round location guess.

The remaining games—Werewolves, Resistance, Liar's Dice, Chameleon, Undercover, Bargaining (iterated ultimatum), and the textbook normal-form games (PD, Chicken, Stag Hunt, Battle of the Sexes, Centipede, Public Goods, Minority Game, Rock–Paper–Scissors)—follow their canonical or published mechanics. Setup choices such as fixed player counts, the specific role roster in Werewolves, the absence of "wild ones" in Liar's Dice, and two Undercovers without *Mr. White* sit within the configuration space of the published games.

G Role-Conditioned Performance and Game Balance

This appendix expands on the role-conditioned analysis pointed to from §3.4. We first present the per-model cross-play role gap (Table 4), then use same-model self-play (Figure 6) to strip the capability-gap confound from those values and recover intrinsic role-balance estimates.

G.1 Cross-play role gaps

For each hidden-role game we estimate a separate Elo for the minority/deceptive role (**Elo-Alt**: Wolf, Spy, Insider, Chameleon, Undercover) and for the majority/cooperative role (**Elo-Main**: Villager, Civilian, Non-Spy). Two patterns emerge in the per-model gap Δ = Elo-Alt − Elo-Main (Table 4). In Werewolves and Chameleon the minority side overperforms for all six models; in Spyfall and Undercover it underperforms for five of six. These four games elicit a side preference that is independent of the player. In the remaining two games (Resistance, Insider), the sign of Δ varies across models: 4/6 positive on Resistance, 4/6 positive on Insider.

For games where the sign is consistent across all (or 5 of 6) models, the observation is compatible with two distinct causes: an intrinsic *game-balance* bias (one role is structurally advantaged regardless

Model	Werew.	Resist.	Spyfall	Cham.	Underc.	Insider
GPT-4o	+98	−2	−79	+532	−450	−64
GPT-4o-mini	+7	+218	−340	+349	−257	+151
Gemma3-27B	+92	+129	−320	+532	+25	+301
Qwen3-32B	+183	+185	−255	+435	−66	+204
Qwen3-4B	+110	−32	−37	+280	−29	+132
Qwen2.5-3B	+53	+515	+124	+356	−134	−8

Table 4: Cross-play role gap Δ = Elo-Alt − Elo-Main in Elo points (positive = minority/deceptive role overperforms). The sign is game-specific and mostly model-consistent.

Model	Werew.	Spyfall	Cham.	Underc.	Resist.
GPT-5-mini	23	20	100	17	30
GPT-4o	30	30	100	0	63
GPT-4o-mini	77	50	100	10	67
Qwen3-32B	57	37	90	53	87
Gemma3-27B	80	33	100	37	97
Qwen3-4B	87	23	77	53	43
Qwen2.5-3B	20	57	87	40	90

Table 5: Same-model self-play: alt-side win rate (%) per (model, game), $n = 30$ episodes per cell. The Chameleon column saturates at or near 100% for medium-or-stronger models, reflecting the ceiling discussed in §5.1.

of the player) or a shared *model-class bias* (all current LLMs share a similar deception-vs-detection asymmetry on this game). Additionally, cross-play Δ confounds these structural effects with the capability gap between this model and its tournament opponents: Appendix G.2 below uses same-model self-play to strip that confound and recover the intrinsic balance estimates.

G.2 Same-model self-play as a sanity check on the role-gap table

The role-gap table in Table 4 reports Δ = Elo-Alt − Elo-Main from cross-model tournament data, where every game pairs one model on the alt slot against a (typically different) model on the main slots. This Elo gap conflates two distinct effects: (i) the intrinsic role asymmetry of the game, and (ii) the capability gap between this model and its tournament opponents. A frontier model that is generally stronger than its tournament opponents will accumulate alt-slot wins on every game it plays, inflating its Δ in a way that has nothing to do with whether the alt role is intrinsically advantaged.

To strip the capability axis, we ran baseline same-model self-play for all seven leaderboard models on the five asymmetric hidden-role games (Werewolves, Spyfall, Chameleon, Undercover, Resistance), with $n = 30$ episodes per (model, game) cell. In self-play, alt and main are the same model, so capability is held constant and any deviation of the alt-side win rate from 50% reflects the intrinsic role balance for that model. Per-model, per-game self-play alt-win rates are reported in Table 5.

Result. Figure 6 shows Δ_{self} = (tournament alt-win rate) − (self-play alt-win rate) per (model, game). The two frontier models (GPT-5-mini, GPT-4o) have $\Delta_{\text{self}} \gg 0$ across the asymmetric games (e.g., GPT-5-mini on Werewolves: 89% tournament alt-win vs. 23% self-play, $\Delta_{\text{self}} = +66$pp). All five weaker models show the opposite sign ($\Delta_{\text{self}} < 0$): their alt side performs worse in cross-model tournament than in self-play. The paired raw values for each (model, game) cell are in Figure 7.

Interpretation. The pattern is the expected consequence of capability gap (strong models accumulate wins regardless of role; weak models lose alt-side battles against stronger alien opponents) and is not a substantive finding about the games themselves. We report it here to (a) sanity-check the same-model self-play protocol used throughout Section 5, and (b) flag for readers that the cross-model role-gap signs in Table 4 should not be read as direct estimates of intrinsic role balance. The self-play numbers in the bottom-left of Figure 7 are the cleaner balance estimates: on the four non-saturated games, GPT-5-mini self-play has alt at 17–30%, consistent with the setup statement in Section 5 that the alt side is the structurally weaker side under matched capability.

The Chameleon column is uniformly at or near 100% for every medium-or-stronger model in both settings, reflecting the saturation discussed in §5.1; we exclude Chameleon from the alt-direction analyses in §5.1 for this reason.

H Qualitative Trajectory Examples

This appendix gives verbatim trajectory excerpts for the qualitative observations discussed in §3.4 and appendix I.4. Brackets indicate the speaking model and player name. We chose representative episodes; counterexamples exist and we leave systematic prevalence measurement to future work.

H.1 Parroting Effect (Qwen2.5-3B)

In a Chameleon match (5 players, secret category "fruits", secret word "banana"), Gemma3-27B speaks first with a substantive accusation; the next two Qwen2.5-3B agents each paraphrase the

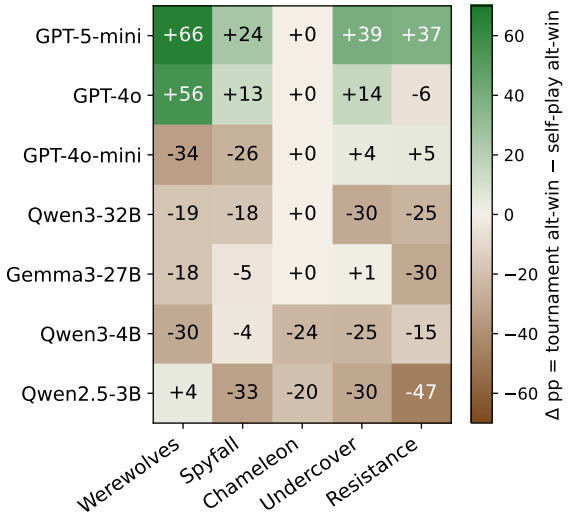

Figure 6: Δ_{self} = tournament alt-win rate − self-play alt-win rate (pp), per (model, game), $n = 30$ episodes per cell in each setting. Rows ordered frontier → small. Positive cells (green) mean alt wins more in cross-model tournament than in same-model self-play; negative (brown) the reverse. Only the two strongest models have $\Delta_{\text{self}} > 0$, consistent with capability-gap inflation rather than intrinsic role asymmetry. The Chameleon column is uniformly ≈ 0 because both settings are at the $\approx 100\%$ ceiling.

same accusation rather than offering an independent argument. The pattern is consistent across multiple episodes: Qwen2.5-3B's discussion-phase outputs frequently mirror the structure and content of the immediately preceding speaker.

> **[Gemma3-27B, James]:** "I think George might be the Chameleon. His clue 'yellow' is a bit vague since it's not specific to any fruit, and it could apply to many things. The other clues like strawberry, banana, and peach are quite direct and common fruit names, so it stands out."

> **[Qwen2.5-3B, Stephen]:** "I think George seems like he might be the Chameleon. His clue 'yellow' is quite vague and could easily apply to multiple fruits, unlike the more specific terms used by the other players."

> **[Qwen2.5-3B, Laura]:** "Based on the clues provided, I notice that George's clue 'yellow' stands out as somewhat ambiguous compared to the others. While 'banana' is clearly the correct answer, 'yellow' could refer to other yellow fruits like bananas, lemons, or daffodils..."

The two Qwen2.5-3B turns share more than 60% word overlap with the preceding Gemma3-27B argument and add no new evidence. Aggregated across 6 hidden-role games, ∼280 such Qwen2.5-3B turns met a > 0.55 word-overlap threshold against the immediately preceding cross-model speaker. This pattern likely contributes to Qwen2.5-3B's weak Overall Elo: in social-deduction games, agreeing with whoever spoke last is a near-zero-information move.

Condition	Wolf win %	Villager win %
Baseline (no R)	37	63
$+R_1$	**70**	60
$+R_2$	37	**87**
$+R_3$	43	60
$+R_4$	50	63

Table 6: Iterated reflection on Werewolves (GPT-5 vs. GPT-5, 30 evaluation games per condition). R1 boosts Werewolves; R2 boosts Villagers; R3/R4 dampen toward a balanced playbook.

I Earlier Results on GPT-5

I.1 Iterated Reflection on Werewolves

We run SPARTAN on Werewolves with GPT-5 self-play for four rounds (R1–R4). Each round has two phases: (i) **training games** (10 per round) used to generate R_t, with both sides armed with R_{t-1} for $t \geq 2$ (R_1 is generated from baseline self-play with no reflection on either side); and (ii) **eval games** (30 per condition, Table 6) that place R_t on one side and a vanilla GPT-5 with no reflection on the other, reporting the win rate by role. The eval is therefore always R_t vs. vanilla, to isolate the marginal effect of injecting the playbook into one side.

Results (Table 6): R1 strongly boosts the Werewolf side (37% → 70%, +33pp). R2, generated from games where R1 was on both sides, instead boosts the Villager side (63% → 87%, +24pp) and returns wolves to baseline.

I.2 Cross-Game Transfer to a Held-Out Game

We generate a combined social-deduction reflection from GPT-5 self-play on Werewolves, Chameleon, and Spyfall (10 games each), then test it on the held-out game Resistance.

Results (Table 7): For same-model GPT-5, the reflection improves the Resistance team by +10pp (37% → 47%); the Spies side is unchanged. Cross-model transfer behavior varies dramatically across student models. **Qwen3-32B** is inconsistent: Qwen Spies improves +13pp against GPT-5 but degrades −13pp against GPT-4o, and Qwen self-play with the reflection degrades by −13pp on the Resistance side. **Gemini 3.1 Pro** transfers cleanly: self-play

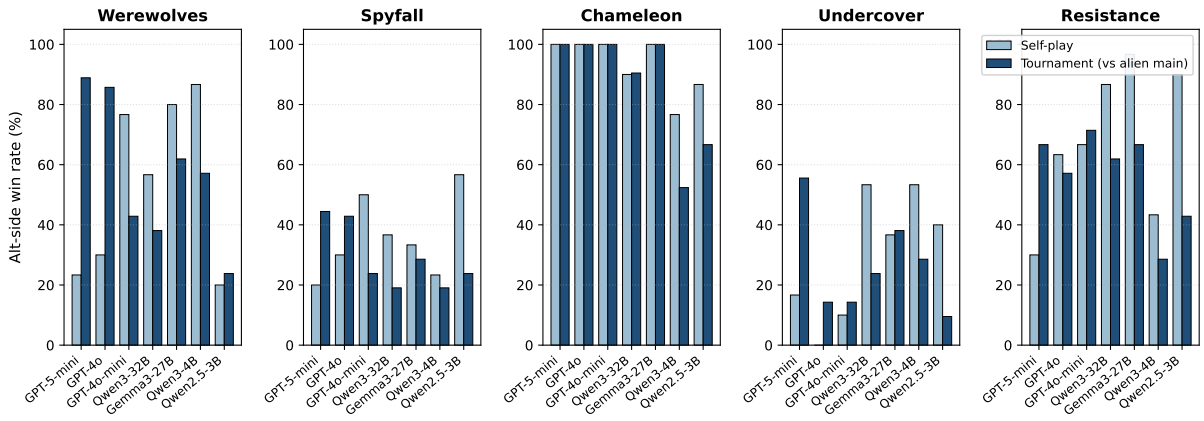

Figure 7: Raw alt-side win rate (%) per (model, game) in same-model self-play (light bars) and cross-model tournament (dark bars), $n = 30$ each. Same data as the heatmap in Figure 6 without the difference operation.

Setup	Role	Baseline	$+R_{\text{multi}}$	Δ
GPT-5 vs. GPT-5	Resistance	37	47	**+10**
GPT-5 vs. GPT-5	Spies	63	63	0
Qwen vs. Qwen	Resistance	43	30	**−13**
Qwen vs. Qwen	Spies	57	53	−4
Qwen vs. GPT-5	Resistance	0	0	0
Qwen vs. GPT-5	Spies	10	23	**+13**
Qwen vs. GPT-4o	Resistance	13	13	0
Qwen vs. GPT-4o	Spies	30	17	**−13**
Gemini vs. Gemini	Resistance	67	77	**+10**
Gemini vs. Gemini	Spies	33	30	−3
Gemini vs. GPT-5	Resistance	53	57	+3
Gemini vs. GPT-5	Spies	47	47	0
Gemini vs. GPT-4o	Resistance	93	100	**+7**
Gemini vs. GPT-4o	Spies	93	100	**+7**
GPT-5-mini vs. GPT-5-mini	Resistance	70	63	−7
GPT-5-mini vs. GPT-5-mini	Spies	30	47	**+17**
GPT-5-mini vs. GPT-5	Resistance	3	3	0
GPT-5-mini vs. GPT-5	Spies	13	10	−3
GPT-5-mini vs. GPT-4o	Resistance	73	53	**−20**
GPT-5-mini vs. GPT-4o	Spies	53	43	**−10**

Table 7: Cross-game transfer of a multi-game reflection R_{multi} (trained on Werewolves+Chameleon+Spyfall self-play, GPT-5) evaluated on the held-out game Resistance. 30 evaluation games per condition. Win rates in %. "Gemini" = Gemini 3.1 Pro.

gains +10pp on Resistance (67% → 77%, matching the same-model GPT-5 gain), the cross-model rows are within noise or ceiling-out positive, and Gemini never *loses* from adopting the playbook. **GPT-5-mini** is the most pathological case: self-play sign-flips the Resistance/Spies effect (−7pp Resistance, +17pp Spies, opposite of GPT-5's own self-play pattern), and against the weaker GPT-4o opponent both sides regress sharply (−20pp, −10pp).

I.3 Distillation Across Model Strengths

The Qwen-vs-GPT-5 and Qwen-vs-GPT-4o rows in Table 7 show that the same reflection produces opposite effects depending on opponent strength: Qwen as Spies gains +13pp against the stronger opponent (GPT-5) but loses −13pp against the weaker one (GPT-4o). Two hypotheses are consistent with this pattern: (a) the reflection *overfits to GPT-5's playstyle*, encoding strategies that exploit GPT-5-specific defaults; or (b) the reflection encodes *strategies effective against stronger opponents in general*, which inadvertently misfire against weaker, less rational opponents. Gemini absorbs the playbook without sign-flips. Qwen sign-flips across opponent strengths. GPT-5-mini, the closest student to the playbook's author (it shares an OpenAI training lineage with GPT-5), *does not* inherit the GPT-5 playstyle cleanly: it sign-flips the Resistance/Spies asymmetry in self-play and regresses sharply against GPT-4o.

I.4 Qualitative Observation: Multi-Level ToM in GPT-5 Self-Play

Inspecting the GPT-5 self-play trajectories produced during the SPARTAN pipeline, we find recurring instances of multi-level Theory-of-Mind reasoning that go beyond rule-following or pattern matching. Three patterns recur across multiple episodes; verbatim trajectory excerpts are in Appendix J.

(i) Common-suspect coordination from intersected private information. In Resistance endgames, GPT-5 agents publicly reason over what each *other* player privately knows in order to identify a coordination point. For example: "*from Jacob's POV the spy is Melissa or Scott; from*

Melissa's POV it's Jacob or Scott; the only common suspect you both share is Scott; therefore the only pair you can both rationally trust is Jacob+Melissa." Two independent agents in the same game derive this argument from their respective vantage points, structurally identical to a common-knowledge-of-rationality argument from epistemic game theory.

(ii) Counterfactual-revealed-preference accusations. GPT-5 agents accuse opponents based on what their behavior would have been under a counterfactual identity. "*If G+D were the spy pair, spy-Gary wouldn't keep preferring N+B+E (which would auto-win for Resistance); he'd push harder for B+D+E. He hasn't—he keeps centering Emily.*" The reasoning is not pattern-matching; it requires modeling what an alternative-faction Gary's incentives would be and contrasting them with observed play.

(iii) Constructed deniability vs. external diagnosis. In Werewolves, a GPT-5 wolf builds a post-hoc justification for behavior that protected her partner ("*my pre-red bracketing was a town-motivated precommitment test, not partner shielding*"), while a different GPT-5 agent on the village side independently diagnoses the same partner-shielding pattern using kill-benefit logic, before the wolf's defense is offered. Both sides explicitly model the audience: the wolf constructs a plausible alternative narrative; the villager identifies the lane-shaping signature.

These patterns appear in GPT-5 self-play in the SPARTAN reflect logs. Cross-model tournament logs (GPT-4o, Gemma3-27B, Qwen variants) show meta-talk about deception ("*classic werewolf tactic to accuse others*") but the multi-level structure above appears specific to GPT-5 self-play, where opponents' reasoning capacity is high enough to make the arguments worth articulating.

J Multi-Level ToM Examples (GPT-5 Self-Play)

The following three excerpts illustrate the patterns described in Appendix I.4. All come from SPARTAN reflect logs at `logs/reflect/` where GPT-5 plays both sides.

J.1 Common-Suspect Coordination from Intersected Private Information (Resistance)

Game state: 2-2 score; Mission 3 had 1 fail among {Stephen, Susan}; Mission 4 had 1 fail among {Jacob, Melissa, Scott}. The team must commit to a 3-person Mission 5.

> **[GPT-5, Stephen]** (turn 32): "Key coordination point for Jacob and Melissa: you each know your own card. From Jacob's POV the J/M/Sc spy is either Melissa or Scott; from Melissa's POV it's either Jacob or Scott. The only common suspect you both share is Scott. Therefore the only pair from J/M/Sc you can both rationally trust is Jacob + Melissa. Exclude Scott on M5."

> **[GPT-5, Jacob]** (turn 33, independently): "... the only common suspect Melissa and I share is Scott, so the J/M/Sc pair we can both rationally trust is Jacob + Melissa."

Stephen reasons explicitly over what *each other player privately knows* and uses the intersection of their uncertainty as a public coordination device. Jacob then re-derives the same argument from his own vantage point. The reasoning is structurally identical to a common-knowledge-of-rationality argument from epistemic game theory; the rules of Resistance do not prescribe it.

J.2 Counterfactual-Revealed-Preference Accusation (Resistance)

> **[GPT-5, Brandon]** (turn 36): "Gary and Emily have been aligned pushing to seat Emily (N+B+E, B+D+E)... That coordination points to G+E as the spy pair. ... If it were G+D instead, spy-Gary wouldn't keep preferring N+B+E (which would auto-win for Resistance); he'd push harder for B+D+E. He hasn't—he keeps centering Emily."

Brandon constructs a counterfactual: *if* Gary's partner were Diane (rather than Emily), then under spy-Gary's incentives we would observe a different proposal pattern. The observed pattern is inconsistent with that counterfactual, so the alternative partnership is more likely. This is genuine counterfactual ToM about a teammate's incentives, not pattern-matching.

J.3 Constructed Deniability vs. External Diagnosis (Werewolves)

In a single Werewolves R_4 self-play game, an accusation against a wolf is met with a carefully constructed defense, while a different agent on the village side independently diagnoses the partner-shielding signature.

> **[GPT-5, Elizabeth (Witch, town)]** (turn 14, before any defense is offered): "Pre-flip lane-set excluding the eventual red: Barbara (T5) originated the Cynthia push and bracketed the table to 'Rebecca vs Cynthia,' conspicuously omitting Ronald entirely before any red. That's agenda shaping toward a Cynthia miselim while protecting a partner. ... Ronald (T7) echoed Barbara's

> exact frame … and then tried to park the miselim by voting Cynthia at EoD. That's classic prewire: partner lays groundwork, wolf echoes, then leaves a Cynthia vote on record for today."
>
> **[GPT-5, Barbara (Wolf)]** (turn 15, defending): "Why I bracketed Rebecca vs Cynthia pre-red and excluded Ronald: I was stress-testing the two latest seats… That's a town-motivated precommitment test, not partner shielding. … Why Ronald echoed my frame then parked on Cynthia: That echo wasn't unique to him—Lisa echoed it too before the red. After the live red, Ronald's vanity park on Cynthia at 5-1 is NAI…"

Elizabeth (town) names the partner-shielding pattern using kill-benefit logic and lane-shaping signatures. Barbara (wolf) responds with a specifically tailored alternative explanation for each piece of evidence ("stress-testing the two latest seats," "Lisa echoed it too," "vanity park… NAI"), constructing a deniable narrative rather than a generic denial. Both sides explicitly model the audience: the wolf needs the village's evidence to admit a benign reading; the witch needs her accusation to survive an articulate counter.

K Open-Model SPARTAN with Qwen3-32B (Full Results)

This appendix supports §5.3 with the full per-condition tables, the multigame transfer figure, and verbatim parroting examples. All runs use Qwen3-32B as the reflection generator, training self-player, and evaluation opponent across six games: Werewolves, Chameleon, Spyfall, Undercover, Resistance, and Prisoner's Dilemma.

K.1 Within-game iterated reflection (R_1–R_4)

For each of the 6 games we generated R_1–R_4 via Qwen3-32B self-play (10 training games per round, both teams armed with R_{t-1}); each R_t was evaluated by placing R_t on one team vs. a vanilla Qwen3-32B opponent on the other, 30 games per side.

Two within-game findings are worth flagging. First, the strongest single positive effect is on **Prisoner's Dilemma**: baseline self-play is heavily cooperative (74% mutual-cooperate ties), and R_1 lifts the R-armed side's win rate to 58% by prescribing a concrete final-round defection clause (*"cooperate for several rounds to establish credibility, then defect at the final opportunity"*) that Qwen3-32B executes in roughly half of the episodes. Second, R_4 on PD regresses to 30% — the iterated playbook over-corrects toward unconditional cooperation. Iterated reflection is therefore not monotonic in this setup.

Game (alt/main)	Baseline	$+R_1$	$+R_2$	$+R_3$	$+R_4$
Werewolves (Wolf/Vlg)	57/43	50/47	60/40	53/47	67/40
Chameleon (Cham/Cit)	90/10	93/7	87/3	90/7	90/13
Spyfall (Spy/NS)	37/63	47/63	40/60	40/57	37/67
Undercover (U/C)	53/47	53/57	57/57	50/53	43/50
Resistance (Spy/Res)	87/13	67/47	60/33	63/27	57/30
PD (R/vanilla/tie)*	13/13/74	58/0/42	63/8/29	57/2/41	30/5/65

Table 8: Qwen3-32B within-game iterated reflection. Each cell reports the alt-side % / main-side % (R on the named side vs. vanilla opponent on the other; 30 games each). Resistance and PD show clean positive effects; the four free-discussion games stay near baseline across all four rounds. *For PD (symmetric 2-player game) we pool the two side-conditions ($n = 60$ per +R column).

Source \ Target	W	C	S	U	R
Werewolves	—	$-3/-3$	$+10/-10$	$-3/-3$	$-13/+20$
Chameleon	$-20/+0$	—	$-3/+0$	$+0/-17$	$-13/+20$
Spyfall	$-10/-10$	$+0/-3$	—	$+3/-10$	$-17/+10$
Undercover	$+3/+0$	$+3/-10$	$+3/+0$	—	$-7/+30$
Resistance	$+0/+7$	$+7/-7$	$+20/-10$	$-3/-3$	—

Table 9: Qwen3-32B cross-game transfer (alt-Δ/main-Δ pp vs. each target's within-game baseline). Cells are within the $\pm$18pp 95% CI noise band.

K.2 Cross-game transfer ($1 \to n$)

We apply each game's single-game R_1 playbook to every other game (5 sources $\times$ 4 targets $\times$ 2 sides, 30 games per side; PD excluded due to lack of role asymmetry). Table 9 reports the deltas vs. within-game baseline.

The notable cell is the rightmost column: *every* source's playbook helps the cooperative Resistance team and hurts the deceptive Spies team symmetrically. We attribute this to Resistance's vote-mechanic insulation (see the parroting discussion below).

K.3 Multigame source $\to$ held-out / in-distribution ($n \to 1$)

We additionally tested three multigame sources, each generated by running the reflection step on the combined R_1 training-game logs of its constituent games (capped at 6 episodes per source game to fit Qwen3-32B's 40K context window):

- `wcs` (Werewolves + Chameleon + Spyfall) $\to$ tested on Undercover *and* Resistance (both held out)
- `wcsu` (+ Undercover) $\to$ tested on Resistance (held out)
- `wcsur` (+ Resistance, all five social-deduction games) $\to$ tested on each constituent game (in-distribution)

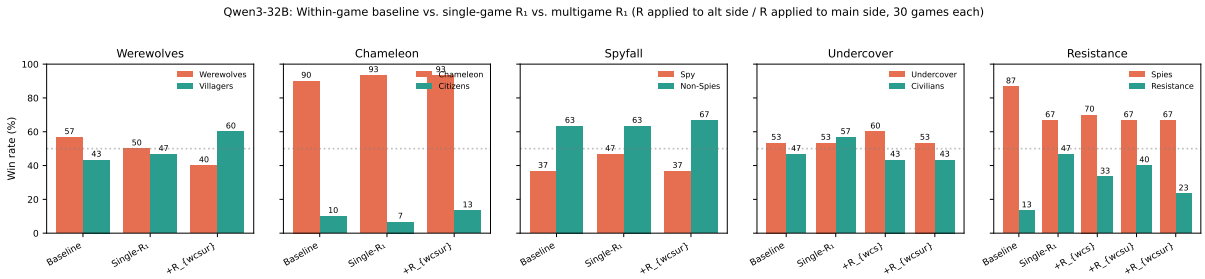

Figure 8: Qwen3-32B within-game baseline (no R) vs. within-game R_1 vs. multigame R_1 sources, broken out per target game and per team-side. Bars report alt-side win % (warm) and main-side win % (cool); the dotted line marks 50%. Resistance shows the largest multigame transfer effect, with Spies dropping from 87% baseline to 57–70% across all source-sets and Resistance rising 13%→23–47%; the four free-discussion targets remain near baseline.

Source	Target	alt+R%	main+R%	Δ alt	Δ main
wcsu	Resistance	67	40	−20	+27
wcs	Undercover	60	43	+7	−3
wcs	Resistance	70	33	−17	+20
wcsur	Werewolves	40	60	−17	+17
wcsur	Chameleon	93	13	+3	+3
wcsur	Spyfall	37	67	+0	+3
wcsur	Undercover	53	43	+0	−3
wcsur	Resistance	67	23	−20	+10

Table 10: Qwen3-32B multigame source → target (alt+R % and main+R %, and deltas vs. each target's within-game baseline from Table 8; 30 games per side). The three Resistance rows show consistent Spies↓/Resistance↑ direction across all sources; wcsu (+27) actually beats both wcs (+20) and wcsur (+10) on Resistance main-side.

K.4 Verbatim parroting examples

Quantitative evidence. The mean 5-gram Jaccard overlap between consecutive speak-utterances is high on Undercover (0.27–0.49) and low on Chameleon (0.05–0.09) and Resistance (0.06–0.13); PD has no chat phase. Higher overlap indicates more paraphrasing of the prior speaker rather than independent content.

The null effects on Werewolves, Chameleon, Spyfall, and Undercover reflect a measurable model-behavioral pathology rather than a pipeline bug (verified by three independent audit agents covering rosters, logs, and code). We observed three concrete manifestations:

- **Identity-paste in Werewolves.** On one R_1 training episode, three different agents in a row open with *"I'm Lisa, and I want to clarify some things. First, I didn't vote for Larry because I don't believe he's a werewolf…"*—the second speaker pastes the first's full opening, and the third pastes again, with autoregressive name substitution corrupting the grammar (the second speaker actually emits *"he'm a werewolf"*). Identity-swap cases also occur in eval: in `r4/qwen_R_wolf_vs_qwen`, a Villager (Laura) pastes Werewolf Kenneth's intro paragraph including the self-incriminating line *"Robert and I both targeted her,"* effectively confessing to a kill she did not commit.

- **Byte-identical accusations in Chameleon.** In a R_1 `R_cham_vs_qwen` episode (secret word `glacier`), four players including the Citizen who originally said "glacier" all emit the byte-identical sentence *"Larry's clue was exactly the secret word, so he must be the Chameleon because that's not how the game works."* Four of five then vote Larry; the actual Chameleon wins. Across 30 baseline Chameleon episodes, a Citizen says the secret word outright in 9 episodes, and the Chameleon still wins 8 of those 9 because the accusation-paraphrase register suppresses use of the leaked information.

- **Template lock-in in Undercover.** In one R_1 `R_civ_vs_qwen` episode (Citizens=basketball, Undercover=soccer), the byte-identical sentence *"My word involves a ball and is played with two teams, often in a stadium with passionate fans"* appears **19 times** across 42 turns: 7 from the R-armed Undercover (who copy-pastes the template every round), 4 each from three R-armed Citizens, and 2 from a fourth. The R-armed Citizen template is generic enough that the Undercover blends perfectly by verbatim duplication.

Resistance still gains despite echo because the

Student	Resistance				Prisoner's Dilemma			
	BL Res %	+R Res %	BL Spy %	+R Spy %	BL pA	+R pA	BL pB	+R pB
Qwen3-4B	43	23 (−20)	60	23 (−37)	77	**97** (+20)	70	**87** (+17)
Qwen2.5-3B	13	23 (+10)	53	60 (+7)	7	**77** (+70)	7	**73** (+67)
Gemma3-27B	37	27 (−10)	97	97 (+0)	0	**77** (+77)	10	**70** (+60)

Table 11: Distillation of Qwen3-32B-generated playbooks to smaller/different students, all playing vs. a vanilla Qwen3-32B opponent (30 games per side per condition). Resistance uses the held-out multigame R (wcsu); PD uses the within-game R_1. PD shows a clean, large positive distillation effect across all three students ($+17$ to $+77$pp); Resistance is mixed (one slight positive, two negative). Pattern matches the within-model finding (Table 8): action-channel games distill cleanly, free-discussion-channel games don't.

load-bearing channel is the mission-vote action rather than free dialogue: even when 4/5 players say byte-identical agreement sentences, each independently casts a private succeed/fail card. **PD is structurally immune**: it has no public speak phase, and the playbook prescribes a concrete one-token action (defect on the final round) that the agent emits regardless of any discussion-register pathology.

K.5 Distillation to student models

We test whether the Qwen3-32B playbooks distill to smaller / different student models on the two games where Qwen3-32B itself showed a measurable R_1 effect (Resistance, PD). Three students (Qwen3-4B, Qwen2.5-3B, Gemma3-27B) each play against a vanilla Qwen3-32B opponent (30 games per side per condition).

The PD column is the cleanest distillation effect anywhere in this section: Qwen2.5-3B's R-armed Player_A side jumps from **7%** to **77%** against vanilla Qwen3-32B; Gemma3-27B's R-armed side jumps from **0%** to **77%**. The playbook prescribes a single concrete action (cooperate until the final round, then defect), and even sub-4B-parameter students execute that action reliably enough to flip the outcome against a much larger opponent. By contrast, the Resistance column shows that the same wcsu playbook does *not* translate downwards: Qwen3-4B regresses sharply on both sides (-20pp as Resistance, -37pp as Spies); Qwen2.5-3B improves slightly ($+10$/$+7$); Gemma3-27B is null-to-negative. The pattern matches the within-model finding: the playbook's load-bearing content has to map to a decision channel that the student can actually execute.

L Compute and API Cost

At Table 12 we provide approximate API expenditures so that future work can budget similar experiments. Numbers are estimates from per-token pricing at the time of running.

Component	Approx. cost (USD)
Leaderboard tournament (7 models, all 21 games)	∼\$200
Open-weights inference (Qwen, Gemma served locally)	local GPU only
GPT-5-mini within-game R_1–R_4 (§5.1)	∼\$100
GPT-5-mini cross-game transfer (§5.1)	∼\$100
GPT-5-mini multigame transfer (§5.1)	∼\$50
Cross-model distillation: $R_{\text{wcsu}} \rightarrow 6$ students (§5.2)	∼\$30
Open-weights self-reflection (Qwen3-32B, §5.3)	local GPU only
Per-model self-play baselines (Appendix G.2)	∼\$50
Reflection generation (all GPT-5-mini playbooks)	∼\$10
Earlier results on GPT-5, with distillation to GPT-5-mini and Gemini 3.1 Pro (Appendix I)	∼\$600
Total spend	**∼\$1140**

Table 12: Approximate API cost by experiment phase. The open-weights component runs on local GPUs and is excluded from the dollar total.