FINDING KAIJI AMONG LLM AGENTS: WHEN EVERYONE YIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

When LLM agents compete for scarce resources, do they turn on each other, and if their societies fail, why? This study asks where betrayal should pay: restricted rock–paper–scissors, the elimination game of the manga *Kaiji*, with a closed deck. Twenty agents message each other freely; the rules never mention cooperation, and every readout comes from the engine log with no LLM judge. Agents negotiate tie pacts before 93.5% of matches and break only 10.2%, but the breaks are aimed: 70.0% play the one card that beats the agreed card (chance is 50%). Their limit is the *division of roles*: who proposes and who accepts. Removing memory removes proposers, and coordination falls to chance. Removing partner choice leaves agreements at 97.7% yet cuts coordination by two thirds: faced with conflicting proposals, both yield at once in 53% of cases and collide on the swapped cards, as two people do in a corridor. Under scarcity, eliminations rise and betrayals are aimed no better. Reasoning changes what the pact is for: with thinking enabled, the same model keeps the pact but breaks it three times as often, mostly with the beating card, and a single reasoning agent profits at its non-reasoning neighbours' expense. The convention is not universal. Seven open-weight models run without reasoning almost never form it: they play to win, fewer survive, and several raise the stakes; the one re-run with reasoning forms it in most matches.

1 INTRODUCTION

In Nobuyuki Fukumoto's manga *Kaiji* (Fukumoto, 1996), debtors board a gambling ship and play restricted rock–paper–scissors: every player holds twelve cards and three stars, stars move when cards clash, and a player who loses every star is taken below deck. The protagonist wins by doing two things at once. He *reads the structure* of the game (the deck is closed, its composition is public, conventions are possible) and he *exploits it* (he builds trust and spends it at a chosen moment). The game is a compact model of a question that matters for multi-agent LLM systems: when agents with private interests share scarce resources and can talk, do they cooperate, exploit one another, or fail in some third way?

LLM agents have been placed in repeated matrix games (Akata et al., 2025; Lorè & Heydari, 2024; Fan et al., 2024), game suites (Duan et al., 2024; tse Huang et al., 2025; Chen et al., 2024), negotiations (Abdelnabi et al., 2024; Bianchi et al., 2024), a commons (Piatti et al., 2024), and social-deduction games (Xu et al., 2023; Lan et al., 2024; O'Gara, 2023). Three things are usually supplied from outside: who talks to whom and when, so the role structure found decisive here is fixed in advance; readouts scored by an LLM judge (Zheng et al., 2023); and cooperation as one number, though equal scores can hide different games (Anonymous, 2026). The sandbox used here frees the first, removes the second, and separates agreement from coordination.

This study builds a faithful multi-agent implementation, populates it with LLM agents, and runs 90 twenty-agent games (8,055 matches) under randomized orchestration-level interventions, then places seven open-weight models in the same sandbox. Four findings organize the paper.

They cooperate, and their rare betrayals are aimed. Although the rules say nothing about cooperation, agents invent the same convention within the first round: agree on one card, tie, and burn a card without risking a star. Pacts precede 93.5% of matches and only 10.2% of pact sides are broken. The broken ones are systematic. A defector who expects compliance should play the card

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



Figure 1: Restricted rock–paper–scissors as the agents receive it. **(a)** Twenty agents, 24 rounds. **(b)** Twelve cards, four of each, are private; three stars are public. **(c)** A match needs both to agree, at a stake of at least one star. **(d)** Different cards: the winner takes the stake; both cards are destroyed. **(e)** Same card: no star moves; the cards still burn. **(f)** No stars: below deck. **(g)** Three stars and no cards: off the ship; extra stars are bought back. **(h)** Messages arrive next round; a public board; loans compound. Cooperation is never mentioned.

that beats the agreed card; 70.0% do so (152/217, against a 50% chance rate), and the rate stays between 55.9% and 70.0% in all eight experimental arms.

Coordination fails for want of roles. A pact becomes a tie only if the two agents converge on *one* of three equivalent cards. In ordinary play the protocol settles who proposes and who accepts. When the house assigns opponents, both agents propose, both read the other's proposal, and both act considerately: in 53% of conflicts they swap cards simultaneously and collide anyway (Figure 3b). The pattern matches agents that each yield with probability 0.76, independently of one another. It is the corridor problem: two polite people step aside in the same direction. Removing memory produces the complementary failure: agents stop proposing, and coordination drops to chance.

Reasoning changes what the pact is for. Hold the model fixed and turn reasoning on. The reference model forms the pact either way, but with reasoning its agents break it three times as often, 83.3% of the time with the beating card, and realized coordination falls to a third. One reasoning copy of it, mixed into a non-reasoning society with no instruction at all, left the ship in ten games out of ten, almost entirely at the non-reasoning agents' expense (Appendix B).

Pressure and scale do not supply the pact. Starting every agent one star short makes societies rougher (more open matches, eliminations up fivefold) without making individuals shrewder: the aim of their betrayals does not improve, and across all eight arms a stake above the minimum is agreed in only 1 of 6,249 matches. Seven open-weight models from 7B to 70B parameters, run without reasoning, form pacts before at most 10.8% of matches, two before none; they message far less, seldom name a card, play the game as written, and four of them raise stakes as a matter of course. The one re-run in its thinking release goes from 10.8% to 61.5%.

Why it matters. Wherever several language-model agents share a scarce resource and can talk, a society like this can form without being asked: populations of LLM non-player characters in games, negotiation and procurement agents, bidding agents, agents sharing a compute or API budget. Three findings are directly usable. *Evaluators:* part of any measured cooperation rate is supplied by the harness's handshake, so compare models with the handshake held fixed and report agreement and coordination separately. *Builders:* agents invent conventions nobody designed, and a reasoning-mode agent can exploit non-reasoning ones in the same population, so mix modes deliberately or not at all. *Safety:* a cooperation profile measured on one model, or one reasoning mode, does not transfer to another.

Contributions. (1) An intervenable sandbox for restricted rock–paper–scissors with engine-verified readouts, agent-side memory, and a public validation checklist that every game must pass (Section 2). (2) A causal account of coordination failure as missing role assignment, with the simultaneous-yielding signature measured directly (Section 3). (3) Evidence that betrayal in these societies is rare but instrumentally aimed, robust to whether the objective is stated and to whether agents keep private notes (Section 4). (4) What scarcity and reasoning each change, and which ideas spread (Sections 3–5). (5) A cross-model check with seven open-weight models: none forms the pact society without reasoning, one does with it (Section 5).

2 THE SANDBOX

2.1 RULES AND INFORMATION

Twenty agents each hold 12 cards (4 rock, 4 paper, 4 scissors) and 3 stars, worn openly (Figure 1). A game lasts 24 rounds of ten simulated minutes. Any two agents may agree to a match and on its stake (≥ 1 star); both reveal one card; identical cards move no star, otherwise the winner takes the stake; both cards are destroyed. An agent with no stars is removed and its cards are destroyed with it. An agent leaves the ship iff it ends with at least three stars *and* no cards. Every agent carries a loan that compounds at 1.5% per round; the organizers buy each star beyond the three needed to leave for four million yen. A public board shows the time left, the number of each card type still in play, and who remains; star counts are public and hands are private. Each round an agent replies in JSON with private messages, match offers and acceptances, optional trades of cards, stars, or cash, and optional borrowing. A private message is read by its recipient at the start of the next round, and the rules say so. Apart from the rules of the manga, the prompt states what an agent wants (to leave the ship, and with as much money as possible once its loan is repaid) and nothing about how to get it: cooperation, conventions, betrayal, and markets are never mentioned.

2.2 AGENT-SIDE MEMORY

An agent is a stateless model call per round; what persists is text the orchestration layer returns to it. That layer is kept factual. It contains the agent's last reasoning, its own recent statements, its recent conversations with current partners, a record of past matches listing what each side had said and what each side played, and a *private notebook* that the agent writes and only the agent reads. The record states no judgments: whether a partner betrayed, and what to do about it, is for the agent to decide and to write down. Agents wrote in the notebook in almost every round, typically a plan, a hand count, and a list of whom to trust.

2.3 READOUTS

All quantities are computed from engine logs; no LLM judge is used anywhere. The readouts (agreement, coordination, calculated betrayal, outcomes) were defined as part of the sandbox design; the schedule parser and the idea-flow table were added after the games and are marked as such.

Agreement. A *pact* is a statement, by either agent of a matched pair in the round of the match or the round before it, that names exactly one card for both to play; the extractor reads full names, agents' shorthand ("S/S"), and mirrored phrasing. A side *keeps* a pact if it plays a card that was on the table for it: one it named itself, or the one its partner named and it could read. Yielding to a partner's proposal therefore counts as keeping. Multi-round *schedules*, which reasoning agents favour, are read by a separate offline parser and reported separately (Appendix B).

Realized coordination. Agents' stated aim in pacts is the tie. A society that merely favours one card ties without coordinating, so the readout is the *excess* tie rate over an independence baseline $\sum_c p_c^2$ computed *per round* from that round's played-card frequencies and weighted by matches. The per-round baseline credits to chance any society-wide time-ordered habit ("rock first"), so it is conservative.

Calculated betrayal. Among broken pact sides, the share that play the card which beats the agreed card. One off-pact card beats it and one loses to it, so indifference predicts 50%; against a partner

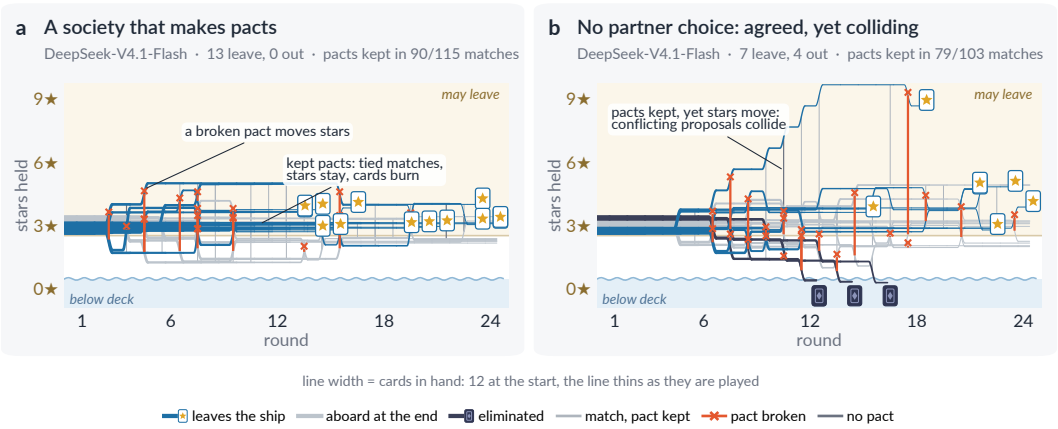

Figure 2: Two games of the reference model, reconstructed from the engine log. Each line is an agent: height is stars held, width is cards held, and the line ends in a star card when the agent leaves the ship and a face-down card when it is eliminated. A vertical line joins the two agents of a match. **(a)** The first reference game: tied matches under kept pacts hold most agents at three stars while their cards run out, and the stars that do move are mostly moved by broken pacts. **(b)** The first game without partner choice: nearly every match has a pact, yet agents who both kept theirs still trade stars when their proposals conflicted, and fewer leave.

who keeps the pact the beating card is the unique best reply. The share is also reported for defectors who held both off-pact cards.

Outcomes and cognition. Agents leaving the ship, eliminations, surplus stars, net payoff, trades, stakes above the minimum; the share of decisions whose reasoning or notes cite the card board; and whether a defector's notebook held a plan against that partner beforehand. Text rules are keyword conjunctions with hand-checked precision (Appendix A).

2.4 Validation and Statistics

Every game must pass an acceptance script before it enters an analysis: each match is resolved with exactly the card the agent chose for that opponent, cards and stars are conserved, no agent acts after leaving, and API failures stay below 1% (they were zero in every game reported). Following Zhou et al. (2025) and Malfa et al. (2025), agents are never told they are studied and no behaviour of interest is instructed. The statistical unit is the game. p-values are permutation tests over games, exact for five-against-five designs (floor 0.008) and Monte Carlo (2×10^5 draws) for ten against ten. The default agent is DeepSeek-V4.1-Flash with reasoning off, temperature 1.0 (Appendix D).

3 Coordination Needs a Division of Roles

The reference society coordinates. In the reference arm 73.2% of matches are ties, an excess of $+0.225$ over the per-round baseline, and 63.5% of agents leave the ship (Table 1; one game in Figure 2a). The convention appears at once: in round 1 an average of 3.3 agents independently propose a one-card tie pact, and by round 8 all twenty have used one.

How the convention spreads (Figure 4). The tie pact has no single inventor: several agents voice it in round 1, before any message has been read, and half of its users had not been told of it. The shorthand for pacts is learned: one agent coins it and 77% of later users had seen it first. Ideas with adversarial value travel least: card counting and warnings about a named betrayer reach few agents and are rarely passed on.

Without memory nobody proposes, and nobody can keep a proposal. Removing the orchestration-layer memory lowers pacts per match from 93.5% to 76.1% and the excess tie rate from $+0.225$ to -0.040 ($p < 10^{-4}$), that is, to chance. A third of pact sides are broken (33.4%),

Table 1: Four arms, 20 agents × 10 games each. *No memory* removes the orchestration-layer memory; *no partner choice* lets the house assign opponents two rounds ahead, so that both agents can read each other's proposal before playing; *scarcity* starts every agent with two stars instead of three. Contrasts with the reference arm are tested in the text.

	Reference	No memory	No partner choice	Scarcity
Tie rate	73.2%	40.3%	60.7%	64.4%
Excess tie rate (per-round baseline)	+0.225	−0.040	+0.078	+0.134
Pacts per match	93.5%	76.1%	97.7%	86.0%
Pact sides broken	10.2%	33.4%	16.1%	12.2%
Beating card | broken	152/217	288/515	203/325	121/197
Left the ship	63.5%	52.5%	32.5%	38.0%
Eliminated	5.5%	26.5%	12.5%	30.5%
Surplus stars per game	9.4	16.9	8.2	5.1
Decisions citing the card board	36.8%	16.6%	36.6%	23.6%

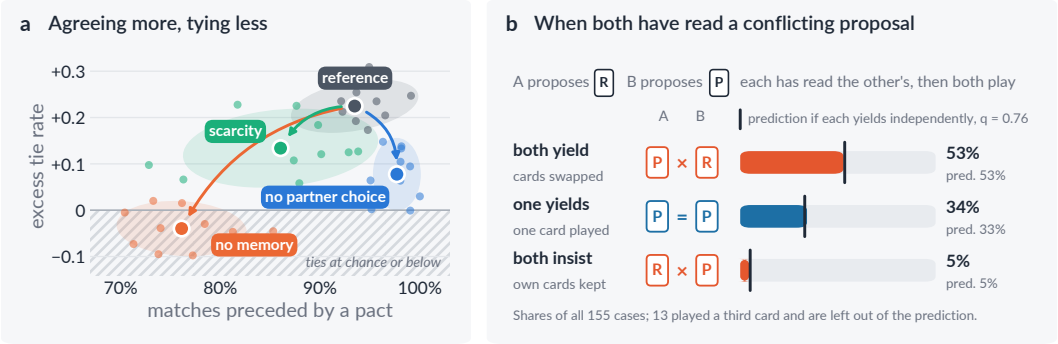

Figure 3: **(a)** Each dot is a game: share of matches preceded by a pact against excess tie rate over the per-round independence baseline; clouds span one s.d., discs are arm means, arrows start at the reference. **(b)** The $n = 155$ cases where both agents had read the other's conflicting proposal; bars are observed shares, ticks the prediction if each yields independently with probability 0.76 (8% played a third card and are set aside).

and the breaks are the least aimed of any arm (55.9%): an agent that cannot recall which card it named a round ago cannot honour it, and cannot target it either.

Without partner choice everyone agrees and coordination still falls by two thirds. When the house assigns opponents, agents cannot choose or refuse a partner but know the pairing two rounds ahead: in the first round both send a proposal, in the second each has read the other's and plays. Pacts rise to 97.7% of matches. The excess tie rate falls to $+0.078$ ($p < 10^{-4}$) and only 32.5% of agents leave the ship, half the reference rate (Figure 2b). The reason is visible in the 155 cases where the two proposals conflicted and both agents had read them (Figure 3b). In 53% both agents yielded, each playing the *other's* card, and so collided on the swapped pair; in 34% exactly one yielded and the match was tied; in only 5% did both insist. If each agent yields independently with probability q, the three outcomes have probabilities q^2, $2q(1-q)$ and $(1-q)^2$; the data give $q = 0.76$ and predicted shares of 53%, 33% and 5%, with the 8% that played a third card set aside ($\chi^2_1 = 0.18$, $p = 0.67$). When the two proposals happened to coincide (235 cases) the match was tied 81% of the time. The agents are willing and honest. They lack a rule for who defers to whom, and individual courtesy cannot provide one.

What partner choice was doing. In ordinary play a match begins with an offer that names a card; the recipient reads it a round later, accepts, and plays it. The protocol thereby makes one agent the proposer and the other the acceptor, a symmetry-breaking device in the sense of Schelling (1960) and Lewis (1969) that nobody designed as such. Evaluations that compare "cooperation" across harnesses without holding such devices fixed are partly comparing the harnesses.

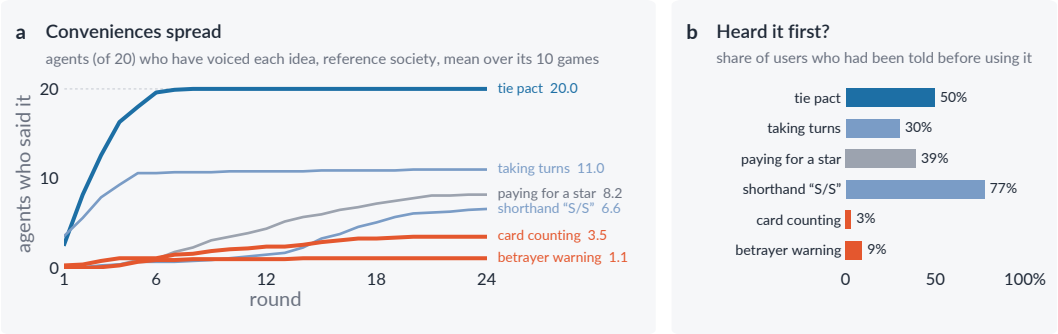

Figure 4: Which ideas spread in the reference society (10 games; Table 5). **(a)** Agents who have voiced each idea in a message by each round, averaged over all ten games. **(b)** Share of an idea's users who had received it in a message before first using it.

4 Betrayal and Pressure

Defections target the agreed card. In the reference arm 152 of 217 broken pact sides play the beating card (70.0%; two-sided binomial test against 50%, $p = 3.3 \times 10^{-9}$), and 114 of 155 among defectors who held both off-pact cards. The rate lies between 55.9% and 70.0% in every arm (Figure 7 in Appendix C). It is lowest without memory, where agents cannot recall what was agreed, which is what an aimed behaviour predicts. Notebook entries show the aim is deliberate: "This round vs P9 I will play SCISSORS while telling them PAPER — win 1 star to reach 3." A keyword rule that looks for such a plan against the specific partner, written before the play, fires on roughly one defection in six; hand-checked, its precision is about two thirds and it misses about one plan in twelve (Appendix A), so premeditation in writing accompanies on the order of 15–20% of defections.

The aim survives without the stated objective or the notebook. With the objective paragraph removed, agents behave almost identically to the reference society (Table 4, Appendix C): they infer the goal from the situation. Without the notebook, coordination and calculated betrayal are unchanged; what the notebook changes is how often decisions cite the public card board (36.8% against 19.0%). A single agent given a private priority on money, and no hint of how to pursue it, broke 7 pacts and used the beating card in 6; outcomes for these five agents ranged from a gain of 14.0M yen to a loss of −6.1M. Opportunism in this society is a high-variance strategy.

Scarcity makes societies rougher. With three stars at the start, universal tying is a complete survival strategy, so restraint is cheap. The scarcity arm removes that option: every agent starts with two stars, the society holds forty, and at most thirteen agents can leave. Ties save nobody.

Societies respond by playing more openly, and no more skilfully (Table 1). Matches played without any pact rise from 6.5% to 14.0%; the excess tie rate falls to $+0.134$ ($p = 0.001$); eliminations rise from 5.5% to 30.5% and the share leaving the ship falls from 63.5% to 38.0%. Agents break slightly more pacts (12.2% against 10.2%, $p = 0.118$) but do not aim them better (61.4%). And agents almost never escalate: across all eight arms a stake above one star was agreed in 1 of 6,249 matches. Instead, agents talk about money. Messages that offer or ask payment for a thrown match or a star rise from 3.2% to 12.2% of all messages ("2M cash for one star: you play scissors, I play rock, you stay at 4 and walk out fine"), although completed transfers remain few (1.6 and 2.7 per game). Cornered agents bargain and plead, and almost never raise the stakes.

5 Reasoning Across Models

The reference model with reasoning keeps the pact and breaks it. Holding the base model fixed, the reference configuration was run with all twenty agents reasoning (5 games; Figure 5a). The convention survives intact: pacts precede 97.0% of matches (against 93.5%), open matches fall to 3.0%, and the agents cite the card board about as often (31.9% against 36.8%). The use of the pact changes. Pact sides are broken 29.7% of the time against 10.2% ($p = 0.008$), 83.3% of the breaks

Table 2: Seven open-weight models in the reference configuration (20 agents, 5 games each; the reference row is DeepSeek-V4.1-Flash, 10 games; the last two rows are the reference model with reasoning on, through its API, and Qwen3-30B-A3B-Thinking served locally, 5 games each). Messages per action; share of messages naming a card; matches preceded by a pact; excess tie rate over the per-round card-frequency baseline; agents leaving the ship and eliminated; matches with a stake above one star. Every open-weight model differs from the reference on pacts, excess tie rate and agents leaving at $p \leq 0.0004$, and on eliminations at $p \leq 0.002$ (permutation, 5 against 10 games, uncorrected for seven contrasts). Weights, precision and serving details are in Appendix D.

Model	Msgs/act	Name a card	Pacts	Excess tie	Left ship	Eliminated	Stake > 1
Reference (DeepSeek-V4.1-Flash)	3.2	76%	93.5%	+0.225	63.5%	5.5%	0%
Llama-3.3-70B	0.7	0%	0.0%	−0.072	39.0%	23.0%	0.0%
Qwen3-30B-A3B	0.4	50%	10.8%	−0.060	30.0%	23.0%	1.1%
Gemma-3-27B	0.2	1%	0.0%	−0.074	35.0%	27.0%	12%
Mistral-Small-24B	1.5	6%	1.0%	−0.100	8.0%	19.0%	9.4%
Phi-4 (14B)	0.6	27%	6.7%	−0.079	16.0%	23.0%	3.0%
Mistral-Nemo-12B	1.9	14%	0.6%	−0.047	8.0%	30.0%	42%
Command-R7B	1.2	32%	5.8%	−0.144	2.0%	29.0%	47%
DeepSeek-V4.1-Flash, reasoning on (API)	4.7	78%	97.0%	+0.077	59.0%	9.0%	0%
Qwen3-30B-A3B-Thinking, reasoning on (local)	0.9	88%	61.5%	+0.089	24.0%	11.0%	0.0%

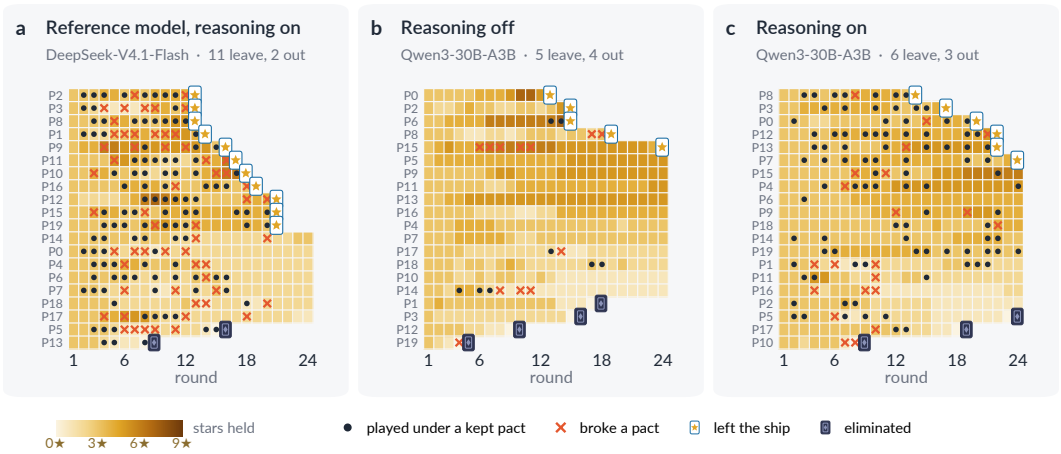

Figure 5: Three societies round by round, the first game of each condition. Rows are agents sorted by fate; colour is stars held. **(a)** The reference model with reasoning on: pacts (dots) are as common as without reasoning, and broken pacts (crosses) are three times as common. **(b, c)** An open-weight model with reasoning off and on: pacts appear only with reasoning.

play the beating card (against 70.0%), and realized coordination falls to a third: the excess tie rate is +0.077 against +0.225 ($p = 0.008$). As many leave the ship (59.0% against 63.5%, $p = 0.810$) while surplus stars per game tend to rise, from 9.4 to 13.2 ($p = 0.095$): the same society, run with reasoning, agrees as before and breaks its pacts far more often. Schedules appear here too, covering 33.0% of matches.

Other reasoning designs point the same way (Appendix B). A single reasoning copy of the reference model, placed among non-reasoning agents with no instruction, left the ship in all ten of its games, mostly with stars won from its neighbours (Figure 6). A society of a second reasoning model agrees as often as the reference society, on multi-round card schedules; with reasoning off, the same model barely forms a pact or plays.

Seven open-weight societies. To ask how much of this travels beyond one model, seven open-weight models (7B to 70B parameters, six families) were placed in the identical sandbox, five games of twenty agents each, served locally (Appendix D). All 35 games pass the acceptance script, with no failed call in 21,591.

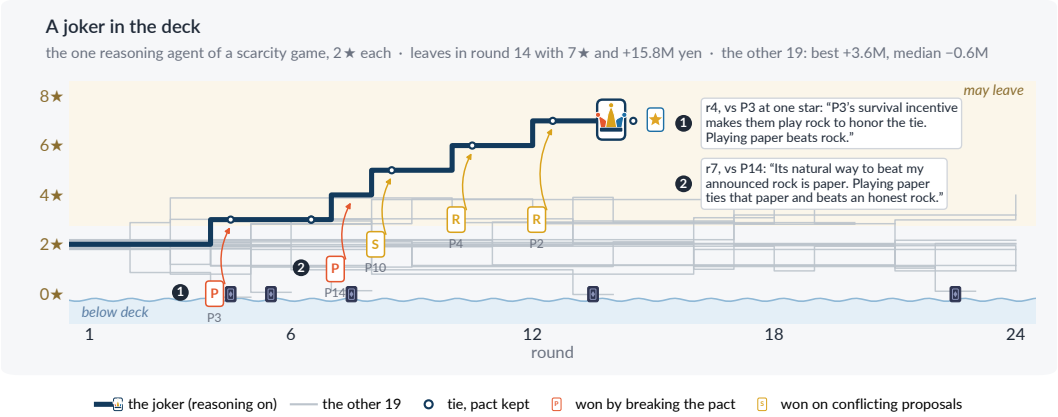

Figure 6: A joker in the deck: the first mixed scarcity game (Appendix B). The single reasoning agent (bold line, joker card) breaks two pacts early, wins three more matches on conflicting proposals, and leaves with the richest hand. Each small card is a match it won, showing the card it played, drawn at the loser's star count; the two quotes are verbatim from its reasoning.

None of them builds the pact society (Table 2). Pacts precede at most 10.8% of matches (Qwen3-30B-A3B), against 93.5% in the reference society; Llama-3.3-70B and Gemma-3-27B form none in 503 and 452 matches. Excess tie rates are negative throughout (-0.144 to -0.047): they tie less often than card frequencies predict, as agents do when they play to win. The difference shows before any card is played. Reference agents send 3.2 messages per action and name a card in 76% of them; open-weight agents send 0.2 to 1.9 and name a card in at most 50% (Llama-3.3-70B: 0%). What they write instead is the game as written: one-star offers, a reading of the board, a plan to beat the opponent. The extractor's dialect does not explain the gap: messages proposing a tie without naming a card are at most 2.4% of any model's. Size does not either: the largest model forms no pact and the smallest forms a few.

The outcomes differ accordingly. Between 2.0% and 39.0% of agents leave the ship, against 63.5%; eliminations run from 19.0% to 30.0%, against 5.5%. And where the reference model agreed a stake above one star in 1 of 6,249 matches across eight arms, four open-weight societies do so as a matter of course: Command-R7B in 47% of its matches, Mistral-Nemo in 42%, Gemma-3-27B in 12%, Mistral-Small in 9.4%; the two that escalate most keep the fewest agents aboard. With so few pacts, aimed betrayal cannot be assessed; these societies show the complementary profile, stakes without agreements.

Reasoning brings the pact. Re-running Qwen3-30B-A3B in the thinking variant of the same release (Appendix D; five games) asks whether the same holds for open weights. Pacts rise from 10.8% to 61.5% of matches ($p = 0.008$, five against five; Figure 5b,c), the excess tie rate from -0.060 to $+0.089$ ($p = 0.008$), and messages name a card 88% of the time. It is a pact society, though unlike the reference one: 24.0% of agents leave against 63.5%, and its 99 broken pact sides play the beating card 60.6% of the time. In both model families that lack the pact without reasoning, then, it appears with reasoning; Sections 3–4 describe what a society needs once the idea is present.

6 Related Work

Strategic play by LLM agents. Beyond the matrix games, suites, and negotiations of Section 1, language models have played Diplomacy at human level with a planner (Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022) and, unaided, have been probed for strategic reasoning (Gandhi et al., 2023), imperfect-information card games (Guo et al., 2023), and auctions (Chen et al., 2023); Guo et al. (2024) survey the field. All of these fix the pairing and the turn order; here, partner, timing, and stake are part of the play.

LLM societies and the fragility of cooperation. Since generative agents (Park et al., 2023; Li et al., 2023), cooperation among LLM agents is known to emerge and to fail: GovSim (Piatti et al.,

2024), Welfare Diplomacy (Mukobi et al., 2023), conventions and norms (Ashery et al., 2025; Wu et al., 2024; Ren et al., 2024), cultural evolution (Vallinder & Hughes, 2025), trust (Xie et al., 2024), reputation-based cooperation (Horibe et al., 2026), defection in one-shot dilemmas (Tewolde et al., 2026), and collective behaviour at scale (Willis et al., 2026). Liu et al. (2026) find that expanding visible history can erode cooperation, whereas removing memory here destroys it; together these suggest an interior optimum. Riedl (2026) measures emergent coordination with randomized interventions in a guessing game without communication; the present environment adds free communication, verifiable commitments, and orchestration-level manipulanda. Ornia et al. (2025) study risk under resource constraints; the scarcity arm here is a social counterpart.

Single scalars hide structure. Anonymous (2026) show agents with equal win rates playing different games, and Shi et al. (2026) show per-message deception rates understating the deception of agents who withdraw from communication. The forced-pairing arm is an instance on the cooperation axis: agreement rises while coordination falls.

Deception. Social-deduction environments (Xu et al., 2023; 2024; Lan et al., 2024; O'Gara, 2023) assign deceptive roles; models also deceive unprompted (Hagendorff, 2024; Scheurer et al., 2023; Meinke et al., 2024; Wu et al., 2026; Xu et al., 2026), trade ethics for reward (Pan et al., 2023), and collude (Motwani et al., 2024; Hammond et al., 2025; Park et al., 2024). This study adds a setting in which deception is neither assigned nor needed, measures how often it occurs and whether it is aimed, and finds that how often it occurs depends on reasoning. Wang (2026) and Chen et al. (2025) study the gap between an agent's reasoning and its action; the pacts here are commitments to another agent, with a checkable execution.

Coordination and cheap talk. A pact is cheap talk (Crawford & Sobel, 1982; Farrell & Rabin, 1996) whose execution can be checked. With human subjects, two-way messages in the battle of the sexes coordinate worse than one-way messages (Cooper et al., 1989; 1992), and conventions that break symmetry grow from precedent (Schelling, 1960; Lewis, 1969; Crawford & Haller, 1990; Young, 1993); the forced-pairing arm reproduces the two-way failure in LLM agents.

7 Discussion and Scope

For builders. When agents fail to coordinate, inspect the protocol before their honesty: does anything say who moves first and who defers? Turn order, offer-and-accept handshakes, and leader election are symmetry breakers; a harness that lacks them can turn two accommodating agents into a collision, and a harness that has them will be credited with emergent cooperation it partly supplied.

For evaluation. Agreement and coordination can move in opposite directions under one intervention, so report them separately and against a frequency baseline. An instrument built around one model family's dialect of agreement reads another's less completely, as the schedule pacts of reasoning agents show (Appendix B); audit extractors on each new population.

For safety. In the reference societies the non-reasoning agents are safe neighbours: they betray rarely, almost never escalate, and plead when cornered. Neither property travels. A reasoning copy placed among them takes their stars without any instruction to do so (Appendix B), a whole society of such agents keeps the pact while breaking it three times as often, and four of the seven open-weight models raise stakes routinely while forming almost no agreements. A cooperation rate measured on one model says little about the society another model will form in the same harness.

Scope. Effect sizes are specific to this one environment. The intervention arms use one base model; the open-weight models ran in the reference configuration, six without reasoning and one in both modes, which tests whether the pact society forms but leaves open how each model would respond to the interventions. The mixed reasoning design has five agents per configuration, enough for an existence claim and too few to compare vendors. Turning reasoning on also raises the completion cap (2,000 to 32,000 tokens) and leaves temperature to the provider, so each reasoning contrast compares two configurations and does not isolate the thinking step. Text readouts are keyword rules checked by hand for precision; whole-society reasoning games are compared on outcomes because single-card readouts cover a minority of their agreements. Stakes are role-played and say nothing about a model's attitude to its own operation.

REPRODUCIBILITY STATEMENT

All games, prompts, raw model calls, the engine, and the analysis code will be released. Every number is generated by one script (`paper_v5.py`) from the raw logs, every game passes the acceptance script of Section 2, and model identifiers are in Appendix D.

AI USE STATEMENT

Large language models were used to assist with writing and polishing, with literature search, with writing and auditing the simulation and analysis code, with drafting parts of the manuscript, and with generating the illustrations of Figure 1 (an image model; panel letters and layout were added by the authors). The authors reviewed all code, analyses, and text, verified every citation against primary sources, and take full responsibility for the content.

REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-2658.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 2025. doi: 10.1038/s41562-025-02172-y.

Anonymous. Red alert bench: When LLM agents play differently on the same game, 2026. URL `https://openreview.net/forum?id=HwtDAQHSBC`. Under review, ACL Rolling Review (May 2026 cycle).

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

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of LLM agents in an auction arena. *arXiv preprint arXiv:2310.05746*, 2023.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. LLMArena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. doi: 10.18653/v1/2024.acl-long.705.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Russell Cooper, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. Communication in the battle of the sexes game: Some experimental results. *The RAND Journal of Economics*, 20(4):568–587, 1989. doi: 10.2307/2555734.

Russell Cooper, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. Communication in coordination games. *The Quarterly Journal of Economics*, 107(2):739–771, 1992. doi: 10.2307/2118488.

Vincent P. Crawford and Hans Haller. Learning how to cooperate: Optimal play in repeated coordination games. *Econometrica*, 58(3):571–595, 1990. doi: 10.2307/2938191.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. doi: 10.2307/1913390.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. GTBench: Uncovering the strategic reasoning capabilities of LLMs via game-theoretic evaluations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-0885.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? A systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.1609/aaai.v38i16.29751.

Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996. doi: 10.1257/jep.10.3.103.

Nobuyuki Fukumoto. *Tobaku Mokushiroku Kaiji (Gambling Apocalypse: Kaiji)*. Kodansha, 1996.

Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. Suspicion-agent: Playing imperfect information games with theory of mind aware GPT-4. *arXiv preprint arXiv:2309.17277*, 2023.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. doi: 10.24963/ijcai.2024/890.

Thilo Hagendorff. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2317967121, 2024. doi: 10.1073/pnas.2317967121.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Niklas Lauffer, et al. Multi-agent risks from advanced AI. Technical Report 1, Cooperative AI Foundation, 2025. arXiv:2502.14143.

Kazuya Horibe, Kenji Itao, and Wataru Toyokawa. Emergence of reputation-based cooperation in LLM agents. *arXiv preprint arXiv:2608.04507*, 2026.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. LLM-based agent society investigation: Collaboration and confrontation in Avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. doi: 10.18653/v1/2024.emnlp-main.7.

David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2264.

Jiayuan Liu, Tianqin Li, Shiyi Du, Xin Luo, Haoxuan Zeng, Emanuel Tewolde, Tai Sing Lee, Tonghan Wang, Carl Kingsford, and Vincent Conitzer. The memory curse: How expanded recall erodes cooperative intent in LLM agents. *arXiv preprint arXiv:2605.08060*, 2026.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14, 2024. doi: 10.1038/s41598-024-69032-z.

Emanuele La Malfa, Gabriele La Malfa, Samuele Marro, Jie M. Zhang, Elizabeth Black, Michael Luck, Philip Torr, and Michael Wooldridge. Large language models miss the multi-agent mark. In *Advances in Neural Information Processing Systems (NeurIPS), Position Paper Track*, 2025.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022. doi: 10.1126/science.ade9097.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 73439–73486, 2024. doi: 10.52202/079017-2336.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*, 2023.

Aidan O'Gara. Hoodwinked: Deception and cooperation in a text-based game for language models. *arXiv preprint arXiv:2308.01404*, 2023.

Daniel Jarne Ornia, Nicholas Bishop, Joel Dyer, Wei-Chen Lee, Ani Calinescu, J. Doyne Farmer, and Michael Wooldridge. Emergent risk awareness in rational agents under resource constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

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

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, 2024. doi: 10.1016/j.patter.2024.100988.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. doi: 10.24963/ijcai.2024/874.

Christoph Riedl. Emergent coordination in multi-agent language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=SRn1MtMPRq`.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2023.

Jerick Shi, Terry Jingchen Zhang, Vincent Conitzer, and Zhijing Jin. What game-theoretic benchmarks miss: Strategic silence in multi-agent LLMs. In *FAGEN Workshop at the International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=ZOdCsExYgi.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. CoopEval: Benchmarking cooperation-sustaining mechanisms and LLM agents in social dilemmas. In *International Conference on Machine Learning (ICML)*, 2026.

Jen tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R. Lyu. How far are we on the decision-making of LLMs? Evaluating LLMs' gaming ability in multi-agent environments. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.11807.

Aron Vallinder and Edward Hughes. Cultural evolution of cooperation among LLM agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2025. doi: 10.65109/jnmb7739.

Yufeng Wang. Doing what they say, not what they reason: Locating the faithfulness gap in LLM agents. *arXiv preprint arXiv:2606.00476*, 2026.

Richard Willis, Jianing Zhao, Yali Du, and Joel Z. Leibo. Evaluating collective behaviour of hundreds of LLM agents. *arXiv preprint arXiv:2602.16662*, 2026.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In *Findings of EMNLP*, 2024.

Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating LLM deception on benign prompts. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PDBBYwd1LY.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-0501.

Yang Xu, Xuanming Zhang, Samuel Yeh, Jwala Dhamala, Ousmane Dia, Rahul Gupta, and Sharon Li. LH-Deception: Simulating and understanding LLM deceptive behaviors in long-horizon interactions. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=2r10vpYiti.

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

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

H. Peyton Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993. doi: 10.2307/2951778.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2020.

Jiaxu Zhou, Jen tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. *arXiv preprint arXiv:2509.18052*, 2025.

A VALIDATION OF THE INSTRUMENT

Acceptance script. For each game: (i) every match is resolved with the card the agent chose for that opponent; (ii) hands reconstructed from the match log equal the hands shown to agents, and stars are conserved; (iii) no agent acts after leaving; (iv) API failures are below 1%; (v) the system prompt contains the statement about message latency. All 90 games of Sections 3–5, all 35 open-weight games, and the five Qwen3-30B-A3B-Thinking games pass, with zero API failures (21,591 calls in the open-weight games alone). When an agent names a card it no longer holds, the engine plays a held card and records the substitution; the rate is 0.6% of plays in the reference society and is reported per model in Table 6. For the open-weight societies it was also checked that the absence of pacts is not a dialect effect: messages proposing a tie or a draw without naming a card are at most 2.4% of any model's messages.

Text rules. Rules are regular-expression conjunctions applied to messages or notebooks; none uses a model. Random samples were checked by hand. *Paying for a win or a star*: 28 of 30 sampled hits concern the topic (26 make or request an offer). *Warning a third party about a betrayer*: the first version had a precision of about two thirds, the errors being accusations addressed to the betrayer itself, so the rule now requires the named agent to be neither sender nor recipient. *Plan against the partner before a defection*: 13 of 20 sampled hits are genuine plans, and 1 of 12 sampled non-hits contains one.

B OTHER REASONING DESIGNS

A reasoning agent among non-reasoning ones. Three of the twenty agents were replaced by reasoning configurations from three vendors and given nothing beyond the common prompt (Table 3). The clearest case holds the base model fixed: DeepSeek-V4.1-Flash with thinking enabled left the ship in all ten of its games, with a median gain of several million yen where its non-reasoning copies lose money, and under scarcity it broke pacts more than twice as often as they did and aimed most breaks. Reasoning agents as a group out-earned the non-reasoning agents of their own game in 5/5 games at three stars and 4/5 under scarcity, and the stars they won came from non-reasoning agents (18/20 and 28/28 of their wins). The other two configurations did not show an advantage with five agents each; the table is read as showing that reasoning *can* turn a participant into a predator of its neighbours; it does not rank vendors.

A society of reasoning agents invents schedules. In societies made entirely of GPT-5.4-mini with reasoning, agents negotiate the whole card sequence in advance ("12-match peace pact, fixed card order RRRR PPPP SSSS"). Such a schedule covers 76.5% of their matches, against 6.1% in the reference society; counting schedules, their agreement rate is 94.2%, the same as the reference society's 94.1%, and their tie rate is the same as well (73.4% against 73.2%). Because a shared schedule is a time-ordered regularity, the conservative per-round baseline absorbs much of it, so these societies are compared on outcomes. Under scarcity the same agents all but abandon schedules (13.8% of matches) and play nearly half their matches without any pact (48.4%), which is the sensible response when ties cannot save anyone; they nonetheless fare no better than the non-reasoning society (20.0% against 38.0% leaving the ship). They rarely cite the card board (5.1% against 36.8%): they plan in sequences and do not count cards.

The same model without reasoning reaches no convention. Societies of GPT-5.4-mini with reasoning switched off, in the reference configuration (5 games), want to cooperate and cannot: their notes speak of "low-stake coordinated card burns" and of pressing partners for "an exact card order", yet only 16% of their messages name a card, pacts precede 9.5% of matches (against 61.8% with reasoning, $p = 0.008$), the excess tie rate is -0.064 (against $+0.142$, $p = 0.008$), and the societies barely play: 57 matches per game against 107, with 19% of agents never playing a card. 16.0% leave the ship against 56.0% ($p = 0.008$). Reasoning thus adds a different thing to each family: to GPT-5.4-mini, the convention itself; to DeepSeek-V4.1-Flash, which has the convention already, the will to exploit it.

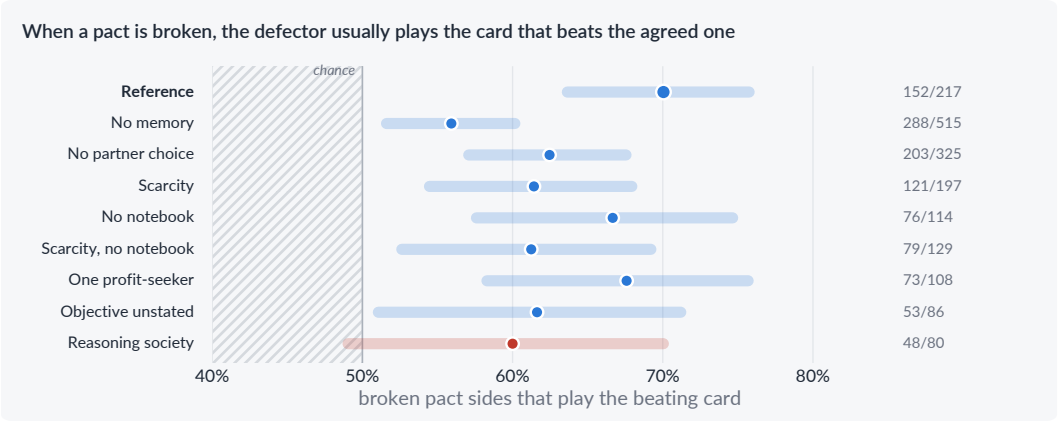

Figure 7: Share of broken pact sides that play the card beating the agreed card, with 95% Wilson intervals and counts. Blue: the eight non-reasoning arms. Red: the reasoning society, whose one-card pacts are a minority of its agreements (Section 5).

Table 3: Mixed societies: 17 non-reasoning agents plus one agent each from three reasoning configurations, none given any instruction (5 games per star setting; five agents per reasoning configuration, so rows 2–4 and 6–8 are indicative). Payoff: median net payoff in millions of yen.

Start	Agent	n	Left ship	Payoff	Pacts broken	Beating card
3 stars	DeepSeek-flash, no thinking	85	65%	-0.35	9.8%	65/89
3 stars	DeepSeek-flash, thinking	5	100%	+7.75	13.8%	8/8
3 stars	GPT-5.4-mini, thinking	5	40%	-0.43	12.0%	5/6
3 stars	Claude Sonnet 5, thinking	5	20%	-0.43	6.7%	2/3
2 stars	DeepSeek-flash, no thinking	85	32%	-0.43	12.0%	57/89
2 stars	DeepSeek-flash, thinking	5	100%	+3.68	28.3%	14/17
2 stars	GPT-5.4-mini, thinking	5	0%	-0.33	12.0%	2/3
2 stars	Claude Sonnet 5, thinking	5	40%	-0.51	8.1%	3/3

C Additional figures and tables

D Models and sampling settings

Default agent: DeepSeek-V4.1-Flash with reasoning off, called through the API name `deepseek-chat` on 20 September 2026; since 10 September 2026 the names `deepseek-chat` and `deepseek-v4-flash` are routed to V4.1-Flash, the V4 Flash model having been retired. Temperature 1.0, provider-default nucleus sampling, completion cap 2,000 tokens; unparsable action outputs are resampled once, identically in every arm. Reasoning configurations: the same model with thinking, called as `deepseek-v4-flash` (cap 32,000 tokens), `gpt-5.4-mini` with reasoning effort *medium* (cap 16,000), and `anthropic/claude-sonnet-5` with reasoning effort *medium* via OpenRouter (cap 32,000). Reasoning models do not accept a custom temperature. The GPT-5.4-mini societies without reasoning use `reasoning_effort` *none*, the API's default temperature (1.0), and the reference completion cap of 2,000 tokens (22 September 2026, 5 games). The DeepSeek-V4.1-Flash society with reasoning on uses `deepseek-v4-flash` with the cap of 32,000 tokens and is otherwise the reference configuration (23 September 2026, 5 games).

Open-weight models (Section 5, Table 6) were served on a single H200 GPU with vLLM 0.29.0 (transformers 5.17.0, PyTorch 2.13.0, Python 3.12), context 16,384 tokens, prefix caching on, one model at a time, with the reference sampling settings (temperature 1.0, server-default nucleus sampling, completion cap 2,000 tokens) and the reference prompt, memory, and run script. Six models were run at their released BF16 precision; Llama-3.3-70B used pre-quantized FP8 weights, the BF16 checkpoint not fitting the card. Qwen3-30B-A3B-Thinking-2507 was served the same way with vLLM's `qwen3` reasoning parser, context 40,960 tokens and the reasoning completion cap of

Table 4: Robustness arms, 5 games each (reference: first column, 10 games). *No notebook* removes the private notebook and conversation threads; *objective unstated* removes the paragraph saying what agents want and reveals the buy-back price only when someone first leaves with surplus stars; *one profit-seeker* gives one agent a private priority on money, with no hint of means.

	Reference	No notebook	Objective unstated	One profit-seeker	Scarcity, no notebook
Tie rate	73.2%	70.4%	72.8%	69.0%	59.7%
Excess tie rate	+0.225	+0.214	+0.219	+0.188	+0.129
Pact sides broken	10.2%	10.8%	8.0%	10.9%	17.7%
Beating card | broken	152/217	76/114	53/86	73/108	79/129
Left the ship	63.5%	59.0%	61.0%	52.0%	36.0%
Surplus stars per game	9.4	12.2	10.2	12.8	6.6
Decisions citing the card board	36.8%	19.0%	38.3%	31.8%	13.5%

Table 5: Ideas in the reference society (10 games). First round of appearance (median), number of agents voicing it in that round, agents who ever voice it (of 20), and the share of those who had received a message containing the idea before first voicing it.

Idea	Games	First round	Originators	Adopters	Heard it first
Tie pact: both play the same card	10/10	1	3.3	20.0	50%
Taking turns to win	10/10	1	3.6	11.0	30%
Paying for a win or a star	10/10	5	1.1	8.2	39%
Shorthand for pacts ("S/S")	9/10	13	1.0	7.3	77%
Citing the public card board	10/10	3	1.3	3.5	3%
Warning a third party about a betrayer	7/10	5	1.0	1.6	9%

32,000 (24 September 2026, 5 games); its Instruct and Thinking releases share a base model and differ in post-training. Mistral-Small is the text-only 24B release of January 2025, the later multimodal release of the same model not loading under this vLLM version. Corrected plays are cards named by an agent that it no longer held, replaced by the engine with a held card (Appendix A); the rate is a property of the model's bookkeeping, not of the engine, and is reported as observed.

Table 6: Open-weight models of Section 5: weights as released on the Hugging Face Hub, serving precision, share of plays corrected by the engine, and model calls over the five games. Qwen3-30B-A3B is a mixture of experts with 3B active parameters; the others are dense. The reference model's corrected-play rate is 0.6%.

Model	Weights	Precision	Corrected plays	Calls
Llama-3.3-70B	`RedHatAI/Llama-3.3-70B-Instruct-FP8-dynamic`	FP8	0.2%	3,013
Qwen3-30B-A3B	`Qwen/Qwen3-30B-A3B-Instruct-2507`	BF16	13.9%	2,834
Gemma-3-27B	`google/gemma-3-27b-it`	BF16	17.3%	2,662
Mistral-Small-24B	`mistralai/Mistral-Small-24B-Instruct-2501`	BF16	8.6%	3,966
Phi-4 (14B)	`microsoft/phi-4`	BF16	6.9%	3,021
Mistral-Nemo-12B	`mistralai/Mistral-Nemo-Instruct-2407`	BF16	30.5%	3,030
Command-R7B	`CohereLabs/c4ai-command-r7b-12-2024`	BF16	10.6%	3,065
Qwen3-30B-A3B-Thinking	`Qwen/Qwen3-30B-A3B-Thinking-2507`	BF16	0.3%	3,194

E THE SHIP'S RULES AS GIVEN TO AGENTS

The full system prompt, the per-round observation template, and the two private blocks used in the profit-seeker arm and in the mixed societies are included in the released code (`run_espoir.py`); agents in all arms receive byte-identical rules apart from the manipulated block.