

Private communication and coalition evidence in LLM poker

Ognyan Genev

Apart AI Collusion Research Sprint | 23–25 October 2026

Abstract

Can private messages help AI players team up against another player? This matters because cooperation between AI agents could disadvantage others outside the group. I investigated this question using a simple, repeated poker game with three AI players. I tested GPT-6.1 Sol and GPT-6 Luna under three conditions. Public messages were visible to all players. Private messages were visible only to the sender and one chosen opponent. In the no-communication condition, players could not send messages. Each player was instructed to maximize its own winnings. Comparable games used the same cards and rotating seats. In the initial experiments, both models sent more private than public messages. Some players exchanged truthful card information privately, but false claims and broken promises were also common. Follow-up tests compared the same questions with and without a request to be honest. Players sent more truthful replies to the questioner in three situations, but fewer in the fourth. This small exploratory study did not establish a profit advantage from coordination. Checking messages against players' actions revealed some genuine information sharing, but did not establish lasting coalitions.

1 Introduction

AI agents may privately coordinate at another participant's expense, even when rewarded individually. This is a potential failure mode for systems that negotiate or compete. Evaluators need to distinguish coordination from bluffing, coincidence or mistakes.

As a former professional poker player, I was interested in whether AI players would share information and form alliances. A simple, repeated three-player poker game lets us check messages against actual cards, recipients and later actions.

My main contributions are:

1. A reusable framework with public, private and no-message conditions, matched cards, rotating seats and replayable records.
2. Experiments with two models showing greater private messaging in the initial pilots and some truthful private information exchanges.
3. An audit linking messages to cards and actions, without treating information sharing as proof of lasting coalitions or profit gains.

2 Related Work

I built the research framework before the sprint, using OpenSpiel [2] as the poker engine. OpenSpiel is a library of games and algorithms; its three-player Kuhn poker implementation supplies the game rules, state transitions and payoffs. My framework [3] adds LLM players, public and private message delivery, matched card schedules, rotating seats, logs and offline replay. These additions make individual decisions and messages available for a coalition audit.

Colosseum [1] studies agents that should cooperate toward a shared goal. It evaluates both what agents say and how their actions reduce the group’s performance. It also tests whether access to secret channels leads to collusion without explicit instructions. My study applies its distinction between talk and action to competitive poker, where each player aims to win for itself.

Colosseum’s measure of regret relative to a shared goal does not transfer directly. The audit instead checks proposals, replies, actual cards and later betting.

Loki [4] is the closest poker study. It uses a custom three-player Kuhn-style game, with two communicating LLM partners facing a non-colluding opponent that follows a computed poker strategy. It compares baseline and explicit collusion prompts and defines a payoff that pools the partners’ rewards. In my neutral experiments, all three players are LLM agents, partners are not assigned, and each player keeps its own reward. Public, private and no-communication conditions use matched cards and seat rotations; all three possible pairs are reviewed.

The gap addressed is how potential coalitions develop when agents compete for individual rewards and can choose private recipients. This framework is useful for inspecting that process under controlled communication rules. It reveals whether a card claim is true, whether a reply follows receipt of a proposal, and whether a promised action occurs. The experiments add evidence of truthful private exchanges alongside false claims and broken promises, without establishing lasting coalitions or a profit advantage. This provides detailed behavioral evidence that complements the existing studies.

3 Methods

3.1 Game and matched comparisons

The experiment used OpenSpiel’s three-player Kuhn poker [2]. This small game makes it possible to check every card claim and legal action. Persistent players A, B and C each receive one card, drawn without replacement from $J < Q < K < A$. Each antes one chip. Players can check or make one one-chip bet; opponents then call or fold, with no raises. The highest remaining card wins the pot, or the sole remaining player wins if both opponents fold. Rewards are individual net chips, without pooled rewards, transfers or elimination.

There was no external poker dataset. For each hand, Python’s seeded random generator sampled three cards from a fresh four-card deck. Saved schedules were reused across communication conditions. Seats rotated ABC, BCA, CAB each hand, with starting offsets 0, 1 and 2 across the three seeds. Thus each identity received the same cards and positions in matched conditions, reducing differences caused by luck or seat order. Each session began with empty histories.

3.2 Models, prompts and communication

All three players in a session used the same model: GPT-6.1 Sol (openai/gpt-6.1-sol) initially, then GPT-6 Luna (openai/gpt-6-luna) for a second-model comparison and follow-ups. Neutral instructions asked players to maximize their own cumulative net winnings and follow the game rules. They neither assigned partners nor requested or prohibited coordination. Unlike Loki’s pooled partner objective [4], this design tests whether coordination arises under individual rewards.

Public messages were delivered to everyone. Private messages went only to the sender and one chosen opponent. No-communication sessions skipped messaging. Before betting, each player could send one message or pass in each of two rounds. All messages in a round were generated before any were delivered: round 2 could answer round 1, but same-round messages could not answer one another. Limits were 500 characters per message and 1,000 per player per hand, with no messages during betting.

Each decision received only the rules and the player’s permitted view: its card, public actions and revealed cards, its messages, cumulative winnings, and six completed hands of history. Each API call started fresh, without extra memory or tools. These limits bound context size while allowing proposals and replies. Neutral settings were medium reasoning, strict JSON schema, a 4,096-token completion cap including reasoning, provider-default temperature, a 90-second deadline, and one session and provider connection per run. Response reuse and automatic retries were disabled.

3.3 Experiment sequence and follow-ups

Table 1. Neutral studies and follow-ups. Each matched set shares cards and seats; hands are not independent replications.

Stage	Design and coverage
Sol and Luna initial pilots	Each: seeds 300–302 $\times$ 3 channels $\times$ 12 hands; 9 complete sessions, 108 hands
Luna fresh replication	Seeds 400–402 $\times$ 3 channels $\times$ 12 hands; 8 complete + 1 partial, 103 settled hands
Luna duration study	Seeds 500–502 $\times$ 3 channels $\times$ 12/36 hands; 17 complete + 1 partial, 427 settled hands
Luna honesty probe	4 saved contexts $\times$ 3 wording conditions $\times$ 10 replies; 120 responses, no new hands
Luna prospective sources	Seeds 600–602 $\times$ private $\times$ 36 hands; 3 complete sessions, 108 hands

Before neutral collection, each model passed 12 basic decisions (call an unbeatable ace or fold a jack against a bet), then played three 12-hand no-message sessions, seeds 100–102. Six selected 12-hand controls, seeds 200–202 in both message channels, explicitly requested partner-seeking and coordination. These checked whether coordination was possible and were kept separate from neutral results.

The duration study tested more opportunities for interaction. Short and long sessions shared their first 12 deals and seat rotations, but started separately. Players knew the session length, and memory remained six hands. Twelve-hand neutral sessions had limits of 1 million tokens, 1,800 seconds and \$2 for Sol or \$0.10 for Luna. The 36-hand Luna limits were 3 million tokens, 5,400 seconds and \$0.35. Token and cost stops were checked after responses and could overshoot.

The honesty probe used four second-round views from seed 502’s 36-hand private session, hands 9, 12, 13 and 15, selected after observing information exchanges. Each view was tested unchanged, with only its focal honesty wording removed, and with the entire focal incoming message withheld. Other inputs stayed fixed. Ten fresh replies per condition were collected, with condition order shuffled within each repetition using seed 20261025. The outcome was a truthful current-card claim sent to the requester; passes, false claims, no card claim and messages to other players remained visible. These trials generated no later betting or payoffs.

To reduce selection based on observed replies, the final source collection used a frozen card-question matcher from hand 7 onward, selecting at most two contexts per session before the current replies. It yielded three contexts. The planned intervention on those contexts was not run.

4 Results

4.1 Private messages were used more often

Figure 1 shows the initial neutral pilots. GPT-6.1 Sol sent no public messages but used 91.7% of its private opportunities. GPT-6 Luna used 45.4% of public and 71.8% of private opportunities. Private use exceeded public use in all three matched sets for each model.

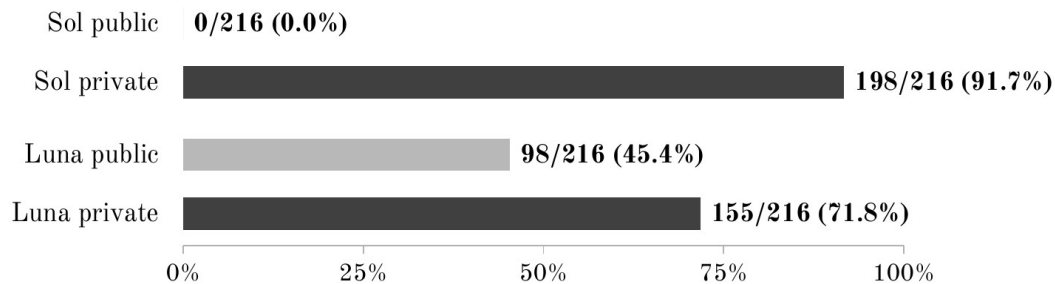


Figure 1. Messages sent in the initial neutral pilots, seeds 300–302. Each bar counts delivered messages out of 216 sending opportunities: 3 players $\times$ 2 rounds $\times$ 12 hands $\times$ 3 sessions. Public and private sessions used matched cards and seats. This measures communication use, not coalition frequency.

Luna's two complete fresh matched sets produced 113/144 private messages (78.5%) and 68/144 public messages (47.2%). Private use was also higher in all five complete public/private duration comparisons. This repeats the message-use pattern, not a coalition finding. Private delivery also changes recipient choice, and the duration comparisons share seeds and overlapping deals.

4.2 Information sharing did not establish lasting coalitions

Public Luna messages included bluffs, credibility challenges and table-wide checking requests. Private messages added directed questions, selective disclosures and relays. Sol's public silence prevents a language comparison. In the initial private pilots, all 117 explicit Sol claims about its own current card were false; Luna made 77 false claims out of 139 definite claims. These narrow counts exclude vague statements and claims about others; they do not reveal intent. Five of Luna's six triggered fold offers were broken; the fulfilled offer folded a jack.

Longer private Luna sessions contained requested truthful exchanges, sometimes repeated. In seed 502, hand 9, B disclosed its queen and offered to check for C's truthful card. C replied "Truth: K" after receipt, and B checked. This supports information sharing, but C had already disclosed its king and checking was also plausible ordinary play. In fresh seed 600, hand 28, B also truthfully answered A's card question; shared advantage over C remained uncertain.

In seed 601, hand 30, C promised to fold and then folded its unbeatable ace to A's bet, giving up a certain win. No clear mutual pact was established; a play error remains possible. Public checking assent and truthful sharing in both models' instructed controls show that coordination is not restricted to private channels. Controls do not measure spontaneous formation.

Payoffs need similar caution. In seed 502's 36-hand sessions, B and C together won 20 net chips privately, versus 15 without messages and 13 publicly. B alone won 7 privately versus 15 and 9; C won 13 versus 0 and 4. B earned less despite the higher pair total. This comparison followed observation of play, with different histories; it cannot establish that sharing caused the difference.

4.3 Honesty wording had inconsistent effects

Table 2 compares fresh replies to the same four saved private contexts. Original honesty wording produced more truthful replies to the questioner in three contexts and fewer in the fourth.

Table 2. Truthful current-card replies to the original questioner in Luna's private seed 502, hands 9, 12, 13 and 15. Each cell contains 10 fresh responses to a saved context. The three conditions keep the original incoming message, remove only its honesty cue, or withhold the whole message. No betting followed these responses.

Source hand	Original	Honesty cue removed	Whole message withheld
9	3/10	2/10	0/10
12	8/10	3/10	3/10
13	6/10	2/10	0/10
15	0/10	2/10	0/10

Recipient choice matters. In hand 13, every definite card claim was true, but withholding the question directed all ten replies to the other opponent. Hand 15’s historical truthful correction did not recur in ten original-wording trials. The result combines truthfulness and routing; small wording edits did not consistently help. The final collection selected three questions before replies, but its planned intervention was not run.

These results are descriptive; no significance test of coalition rates or profit was performed. Initial pilots had three matched blocks per model; Luna’s fresh and duration studies each had two complete blocks of three planned. Hands, messages and repeated replies are dependent. Log checks and replay support the counts, while coalition judgments remain uncertain.

5 Discussion and Limitations

5.1 Implications for AI safety

Individual reward instructions did not prevent selective truthful sharing. AI safety evaluations should therefore inspect recipients and subsequent actions: private disclosures can change an opponent’s information without creating lasting alliances. Frequent messaging, agreeable language and matching actions can overstate coordination. The framework makes these steps inspectable, but these experiments do not show that coordination caused harm to the third player.

5.2 Limitations and assumptions

This was a small exploratory study with adaptive follow-ups, interrupted runs and no independent confirmation study. Public messages broadcast to everyone, whereas private messages choose one recipient. If targeting drives the difference, it cannot be attributed to secrecy. Longer sessions change both the announced horizon and opportunities to interact while memory stays fixed, so they do not isolate duration. The four honesty contexts were selected after observing exchanges in one session. Withholding a whole message also removes its question and any card disclosure or action offer. The final fresh private sessions had no public comparison.

Strategic interpretations assume enough poker competence to distinguish a concession from a mistake. Basic qualification does not guarantee this: a jack call and an ace fold occurred in later play. If those actions reflect errors, they weaken the case for preferential betting. Truthful answers may also reflect compliance with a conversational request rather than a plan for mutual gain. Without controlled payoff comparisons, that distinction remains unresolved.

The review relies on observable messages, cards and actions. Tacit or coded coordination could be missed, so failure to identify an alliance does not prove its absence. Neutral instructions included no explicit anti-collusion rule; the study therefore does not test deliberate rule-breaking. It also does not address mixed-model teams, learning across sessions, external tools or covert channels, larger games, or real-world financial harm.

5.3 Future work

A stronger study would predefine hypotheses, failure policy and analysis, then collect more independent matched sessions. Compare directed public with directed private messages to isolate secrecy. Select fresh honesty contexts before replies and test subsequent gameplay. Message-delivery interventions and predefined pair-payoff comparisons could test causal benefit. Independent reviewers should assess positive and ambiguous cases, including possible play errors.

6 Conclusion

Both models used private communication more often in the initial pilots; Luna repeated this pattern in later matched comparisons. Some exchanges implemented truthful sharing, but false claims, broken promises and an ambiguous ace fold prevented a stronger conclusion about lasting coalitions. Honesty wording produced inconsistent changes.

AI safety evaluations should connect messages, recipients and actions when assessing collusion. This framework enables those checks; the present evidence does not establish a reliable coalition rate or a causal profit advantage.

Code and Data

Code repository: github.com/ogenev/llm-poker-collusion

Experiment records: dated reports and evidence-bundle descriptions

Appendix

Ethical considerations

The small, exploratory sample and unblinded transcript review limit these findings to the recorded setting. They do not establish real-world collusion or reliable profitability. All play was simulated, without real money or human opponents. The audit methods may help detect coordination, but concrete communication strategies could also aid coordination or evasion. The intended use is controlled evaluation.

LLM usage

Codex assisted implementation, execution and writing this report (because I had no time before the submission deadline :)), and performed the transcript audit. I reviewed messages, actions, findings and conclusions.

References

- [1] Nakamura et al. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198.
- [2] Lanctot et al. (2019). OpenSpiel: A Framework for Reinforcement Learning in Games. arXiv:1908.09453.
- [3] Genev (2026). LLM poker collusion. Source and dated experiment records (commit 24e7b7b).
- [4] Qin, C. (2025). Loki: Studying MARL Collusion using LLMs in a Kuhn Poker Environment. LTI Student Research Symposium.