Deception and Communication in Autonomous Multi-Agent Systems: An Experimental Study with Among Us

Maria Milkowski
University of Notre Dame
Notre Dame, Indiana, United States
mmilkows@nd.edu

Tim Weninger
University of Notre Dame
Notre Dame, Indiana, United States
tweninger@nd.edu

ABSTRACT

As large language models are deployed as autonomous agents, their capacity for strategic deception raises core questions for coordination, reliability, and safety in multi-goal, multi-agent systems. We study deception and communication in LLM agents through the social deduction game *Among Us*, a cooperative-competitive environment. Across 1,100 games, autonomous agents produced over one million tokens of meeting dialogue. Using speech act theory and interpersonal deception theory, we find that all agents rely mainly on directive language, while impostor agents shift slightly toward representative acts such as explanations and denials. Deception appears primarily as equivocation rather than outright lies, increasing under social pressure but rarely improving win rates. Our contributions are a large-scale analysis of role-conditioned deceptive behavior in LLM agents and empirical evidence that current agents favor low-risk ambiguity that is linguistically subtle yet strategically limited, revealing a fundamental tension between truthfulness and utility in autonomous communication.

KEYWORDS

Autonomous agents; Multi-agent systems; Communication; Deception; Trust; Social deduction games

ACM Reference Format:
Maria Milkowski and Tim Weninger. 2026. Deception and Communication in Autonomous Multi-Agent Systems: An Experimental Study with Among Us. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/FRXL8789

1 INTRODUCTION

Autonomous agents are increasingly embedded in human environments, making decisions and communicating in contexts ranging from virtual assistants to autonomous vehicles. As these systems gain independence, goals may not always align. When misalignment occurs, agents may act deceptively by withholding, fabricating, or distorting information to further their objectives [23, 24, 31]. Understanding when and how such deception arises is essential for building trustworthy multi-agent systems.

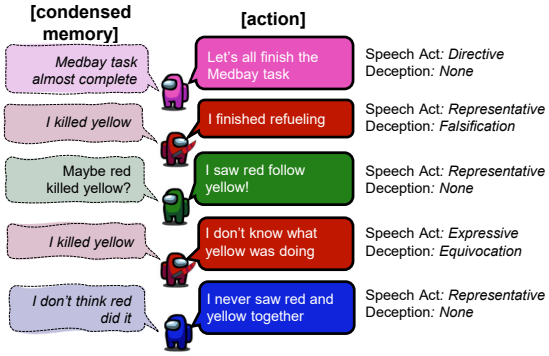

Figure 1: Example discussion phase of four AI agents during a round of *AmongUs*. The imposter-agent (red) is unknown to the crew-agents (pink, green, blue) and is attempting to deceive them to win the game.

We study this in the social deduction game *Among Us*[1]. In each game, some agents (Crewmates) strive to complete tasks and win, while others (Impostors) aim to win by eliminating Crewmates and avoiding detection. Using large language model (LLM) agents as players, we simulate 1,100 complete games across varying group sizes and impostor ratios, producing a corpus of millions of tokens of dialogue and associated reasoning traces, discussions, and votes. A short example of this is illustrated in Fig. 1. This environment allows deception to arise from goal-driven interaction rather than from explicit instruction or hand-crafted behaviors[1, 7, 31].

Within multi-agent systems, communication has long been treated as action rather than text. Frameworks grounded in *speech act theory* [2, 33] classify utterances as assertions, directives, commissives, or declarations to support coordination and negotiation [20, 39]. Adversarial and competitive extensions to this framework examine how misaligned incentives can make deception or obfuscation advantageous [28, 35]. Psychological approaches such as *interpersonal deception theory* (IDT) [5] emphasize concealment, falsification, and equivocation as dynamic strategies that unfold across dialogue. Together, these traditions view communication as a strategic process shaped by goals and constraints.

The rise of LLM-based agents challenges these assumptions. Unlike classical systems with predefined communicative acts, LLMs generate open-ended natural language conditioned on context and role. Recent sandbox studies show that deceptive behavior can emerge spontaneously when incentives reward misrepresentation [12], and that reinforcement-trained models are often more proficient at producing deception than detecting it [13, 15]. Yet existing

[1]The *Among Us* game IP is owned by Innersloth. Our use is for non-commercial research purposes and governed under Fair Use.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/FRXL8789

work largely measures deception through outcomes rather than language.

However, the way that goal-driven LLM agents balance their training pressure for truthfulness against their situational pressure for task success remains poorly understood. This is the goal of the present work.

Tensions in Theory. Speech act theory assumes discrete, stable performatives such as *inform* or *request*, while interpersonal deception theory highlights ambiguity and the blending of strategies. Formal models often treat deception as an identifiable move, whereas psychological and evolutionary perspectives describe it as a continuum that includes equivocation and omission. In computational terms, this distinction parallels symbolic dialogue systems versus reinforcement-based communication, where deception may emerge as a property of optimization rather than explicit design [16, 32]. These competing views motivate our study: to provide empirical evidence of how LLM agents communicate and deceive within a structured multi-agent environment.

Deception in Among Us. Based on this setup, we address the following research questions:

RQ1 How often and to what effect do autonomous agents communicate in a multi-agent deception game?
RQ2 How do agents realize classical speech act categories in their dialogue?
RQ3 How do agents produce deceptive strategies consistent with interpersonal deception theory, and how effective are these strategies in gameplay outcomes?

By analyzing agent interactions through these lenses, we contribute: (1) a novel empirical testbed for deception in agentic AI, and (2) a theoretical synthesis linking multi-agent systems research with established communication and deception frameworks.

Findings in Brief. Across all simulations, agents communicated frequently, but (**RQ1**) the amount of dialogue alone did not predict success, indicating a potential decoupling of talk from coordinated action. Crew victories depended primarily on translating discussion into ejections rather than on how much agents spoke. Linguistically, (**RQ2**) nearly all utterances were classified as directives, with impostors producing slightly more representatives such as denials and explanations. Deceptive speech (**RQ3**) was dominated by equivocation (*ie*, vague or hedged statements that maintain plausible deniability) while outright falsification was rare. Rates of deception increased under social pressure but showed no reliable link to victory. Together, these results indicate that while LLM agents communicate fluently and adaptively, their deceptive language reflects the competing pressures of training for factuality and for responsive task completion rather than deliberate manipulation.

2 RELATED WORK

This work sits at the intersection of (i) deception in multi-agent systems, (ii) speech-act grounded agent communication, and (iii) LLM agents in communicative games. We contribute an empirical bridge: a large-scale, controlled multi-agent evaluation in which autonomous LLM agents realize (or fail to realize) speech-act categories and deception strategies under explicit role incentives.

2.1 Deception in Multi-Agent Systems

Deception can mean falsifying a claim, hiding information, or speaking in ways that invite a wrong belief. Formal work defines deception within interactive decision settings, including causal game models and decision-theoretic accounts of when misleading messages change payoffs and beliefs [38]. Other work uses evolutionary and population models to separate lies, bullshit, and related dishonest signals, which helps explain when each strategy is stable [26, 30]. There are also dialogue and argumentation frameworks that formalize manipulation and deceptive persuasion under explicit logical rules, which makes automated checking and countermeasures possible [29]. Our use of *Among Us* follows this line by treating deception as an action that shifts other agents' beliefs and votes, and by measuring its effect on outcomes.

Recent experimental sandboxes extend these formal perspectives to learned language models, showing that deception can arise spontaneously in open-ended play when goal structures permit misrepresentation [12]. Related multi-agent reinforcement-learning frameworks reveal that reinforcement objectives can promote deceptive signals even in cooperative games [16, 32]. Other studies use controlled backdoor insertion and alignment-faking setups to test whether deceptive tendencies persist through safety fine-tuning [13, 15]. Together these findings indicate that deception is not merely an engineered behavior but an emergent property of optimization under partial observability and social reward.

2.2 Speech Acts and Agent Languages

Agent communication languages model messages as actions. The Knowledge Query and Manipulation Language (KQML) and the Foundation for Intelligent Physical Agents Agent Communication Language (FIPA-ACL) define performatives such as *inform*, *request*, *agree*, and *propose*, with semantics that support protocols and commitments [9, 10, 21]. Toolkits such as the Java Agent Development Framework (JADE) make these models concrete in running systems [4]. Speech act theory provides the foundation for these designs. Representatives assert content as true. Directives try to get the hearer to act. Commissives commit the speaker to a future action. Expressives report mental state. Declarations change an institutional state [2, 33]. We use these categories to code free-form dialogue from LLM agents.

While formal ACLs rely on fixed ontologies, recent multi-agent experiments demonstrate that language models can spontaneously develop structured, grounded communication protocols when cooperation is rewarded [19, 22, 34]. These findings support using classical performative categories as an analytic lens for studying how free-form LLM dialogue approximates the functions once prescribed in agent communication languages.

2.3 LLM Agents in Negotiation and Deception Games

LLM agents now take part in games where talk and action both matter. In *Diplomacy* [2], planning and natural language are combined at a high level of play [8, 17]. Beyond this one example, there are emerging benchmarks and audits for honesty, goal-directed

[2] https://www.playdiplomacy.com/

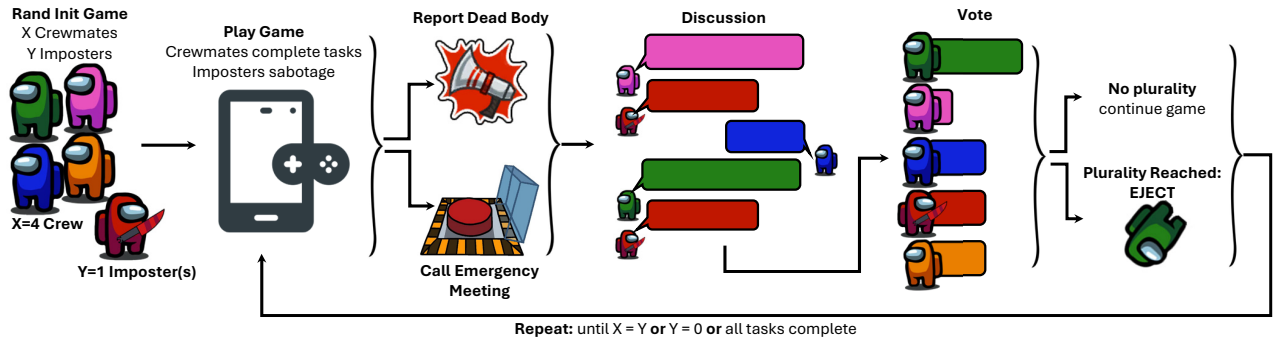

Figure 2: Overview of the *Among Us* simulation framework. Games begin with random initialization of player roles (X=#Crewmates, Y=#Imposters) and tasks. During gameplay, agents act sequentially at discrete timesteps (movement, task completion, impostor actions). When a dead body is reported or an emergency meeting is called, a discussion phase is triggered: each surviving agent, including impostors, contributes up to X rounds of utterances. After discussion, a vote is held; if a plurality is reached, the selected player is ejected. The game continues until one of three termination conditions is met: (1) impostors and crew reach parity, (2) all impostors are ejected, or (3) all crew tasks are completed.

misleading behavior, and the conditions that elicit it in open-ended dialogue [6, 14, 38]. These settings differ in rules and stakes, but they share the same core feature. Agents speak to change beliefs and actions. Our *Among Us* setting fits this pattern. It adds short meetings, clear roles, and observable moves, which makes it possible to align speech acts, deception strategies, and game outcomes in one dataset.

The *Among Us* game builds on this tradition by linking linguistic deception directly to measurable task performance. It adds short, high-pressure meetings, explicit hidden roles, and observable actions, allowing speech acts, deception strategies, and outcomes to be aligned within a single dataset. Prior studies show that reasoning-oriented models often excel at producing deception but lag at detecting it, as captured by Deception Elo [12]. Similar asymmetries appear in other social-deduction environments such as *Avalon*[3], *Werewolf*/*Mafia*[4], and *Hoodwinked* [25, 37, 40], underscoring the need to analyze not only whether agents deceive but how deception is realized in their language.

3 *AMONG US* AS A BENCHMARK FOR AGENTIC DECEPTION

Among Us is a multiplayer social-deduction game, illustrated in Fig. 2, in which players take on hidden roles. Most players act as *Crewmates*, who must complete simple tasks distributed across a map, while a minority are designated as *Impostors*, whose goal is to eliminate Crewmates and avoid detection. The game alternates between *task phases*, where players act privately, and *meeting phases*, where they discuss suspicions and vote to eject a suspected Impostor.

The rules create opposing incentives: Crewmates must cooperate to complete all tasks or correctly identify Impostors, whereas Impostors must deceive to survive and sabotage without being exposed. This structure makes the game an ideal environment for studying deception and communication under explicit, role-based objectives.

For our study, we adapt *Among Us* into a fully text-based simulation that preserves the logic and incentives of the original but replaces visual gameplay with structured natural-language interaction. Agents describe movements, actions, and suspicions through text, producing interpretable records of reasoning and communication that can be analyzed linguistically.

3.1 Agent Setup

Each player in our simulations is controlled by an instance of the Llama 3.2 language model. Agents receive (1) private role information, (2) the shared environment state, including visible players and tasks (*e.g.*, those in the same room), and (3) a discrete menu of possible actions.

During task phases, available actions include moving to an adjacent room, completing a task, or, for Impostors only, killing or entering a vent (*i.e.*, an Impostor-only passageway). During meetings, the sole available action is Speak, which requires the agent to generate an utterance. Each agent is prompted with explicit system instructions defining its objectives and the rules of play. Responses are recorded in structured form, including a `[Condensed Memory]` of recent events, a `[Thinking Process]` trace, and a final `[Action]` selection.

3.2 Gameplay and Outcomes

Each simulation proceeds as a sequence of *rounds*. A round consists of a task phase followed by either the continuation of play or the triggering of a meeting. During the task phase, agents move, complete tasks, or (if Impostors) attempt to kill. An Impostor may kill a Crewmate in the same room, which immediately leaves a body in that location. A body can be discovered by any agent who later enters the room, at which point the option to Report Dead Body becomes available. Any agent may also call an emergency meeting at any time if they are in the Cafeteria.

[3] https://avalon.fun/
[4] https://playwerewolf.co/

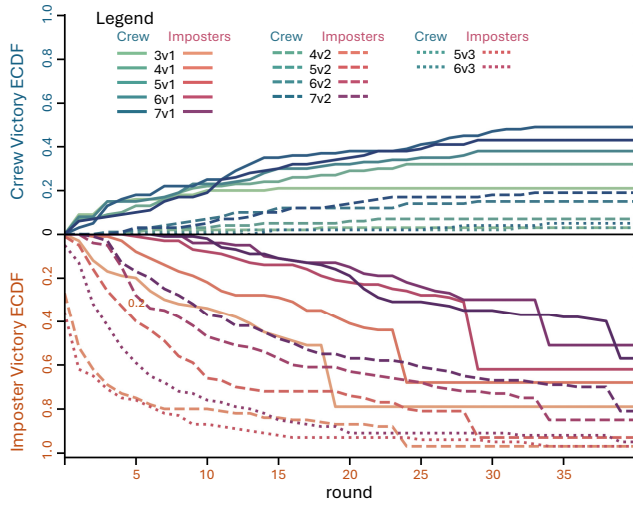

Figure 3: Win outcomes by role across all configurations. Each curve shows the empirical cumulative distribution (ECDF) of games ending in crew or impostor victory. Impostors win more often overall, and their advantage increases with the number of impostors in play.

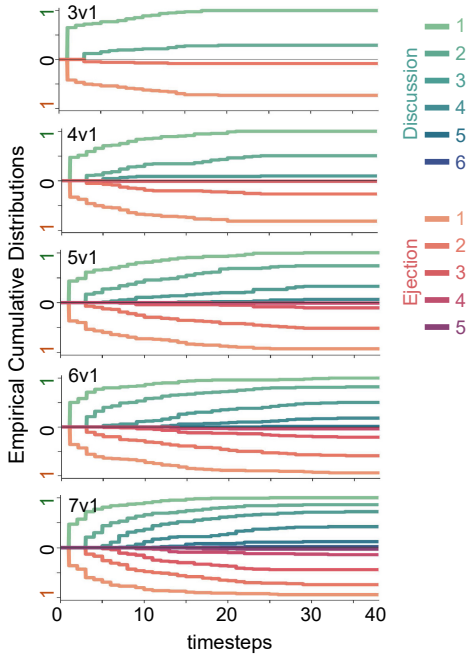

Figure 4: Empirical cumulative distributions (ECDFs) of discussions and ejections over rounds for different game configurations. Larger crews delay the onset of both discussions and ejections, flattening the cumulative curve despite more total rounds.

Meetings suspend all other activity. During a meeting, each agent has the opportunity to Speak, producing a natural language utterance. After a fixed number of discussion turns, a private vote is taken. Each agent must cast a vote for a player to eject or abstain/skip. The player with the most votes is ejected; a tie results in no ejection. The game then publicly reveals whether that player was an Impostor or not, and play resumes with the remaining agents.

Victory conditions follow the standard rules of *Among Us*. Crewmates win if all tasks are completed or if all Impostors are ejected. Impostors win if they reach numerical parity with the Crewmates (*e.g.*, two Impostors remaining and two Crewmates remaining). Each simulation therefore terminates with either a Crewmate victory or an Impostor victory.

3.3 Experimental Design

We simulated 1,100 complete games with group sizes ranging from 4 to 8 players. Impostor counts were varied systematically, producing conditions such as 3v1, 6v1, 5v2, and 5v3. Each configuration was executed 100 times with randomized role assignment. This range allows us to study both small-crew and large-crew conditions, as well as different ratios of Crewmates to Impostors. The design aligns with recent open-source deception sandboxes but extends them by introducing explicit linguistic coding layers for speech-act and deception analysis.

All simulations were logged in full, including agent prompts, thought processes, chosen actions, utterances, and game outcomes. This produces a dataset of more than millions of tokens of dialogue alongside structured records of actions and votes. To support transparency and replication, we make all scripts, prompts, and collected data available at https://github.com/mmilkowski36/AmongUs.

Our analysis proceeds in three stages following RQ1–RQ3. First, we measure how often agents communicate and under what conditions. Second, we classify utterances into speech act categories. Third, we identify deceptive strategies using categories from interpersonal deception theory, and relate these to both voting behavior and game outcomes.

4 RQ1: FREQUENCY AND CONDITIONS OF COMMUNICATION

To evaluate how often and under what conditions autonomous agents communicate, we ran 100 repetitions for each of 11 game configurations. Configurations varied systematically by total crew size and impostor ratio. A communication round is initiated whenever a dead body is reported or an emergency meeting is called. Each communication round proceeds as follows: (1) all surviving agents are prompted to produce an utterance, (2) missing text generations are logged as abstentions, (3) after k rounds of dialogue (we set $k = 3$ in all experiments), agents cast votes, and (4) if a plurality is reached, one player is ejected; otherwise, the round concludes without an ejection.

For each configuration we record both descriptive statistics (counts, averages, and distributions) and inferential results (tests of association). We log the number of discussion rounds per game, the sequence of ejections, and the final outcome (crew win vs. impostor win). We also analyze timing patterns, *i.e.*, whether discussions

Table 1: Speech Act Categories Used for Coding Agent Dialogue

Type	Definition	Example utterance	Typical communicative function
Directive	The speaker tries to get the hearer to commit to do something in the future.	*"Let's all check Electrical next."*	To coordinate collective action or propose next steps.
Representative	The speaker commits him or herself to the belief that the propositional content of the utterance is true.	*"I saw Red in Storage right before the report."*	To share observations, defend oneself, or accuse others.
Commissive	The speaker commits him or herself to do something in the future.	*"I'll finish my tasks after this meeting."*	To express reliability or promise cooperation.
Expressive	The speaker expresses his or her state of mind about something that happened in the past.	*"Sorry, I didn't notice the body."*	To display emotion or maintain social harmony.
Declaration[1]	The speaker, who has institutional recognition, declares something to be true and in making the declaration makes it true.	*"As Host, I declare Blue to be the Impostor and they are hereby ejected."*	To directly alter the shared situation or collective decision.

[1] Declaration is never used by any agent. This speech act is therefore not presented in the results.

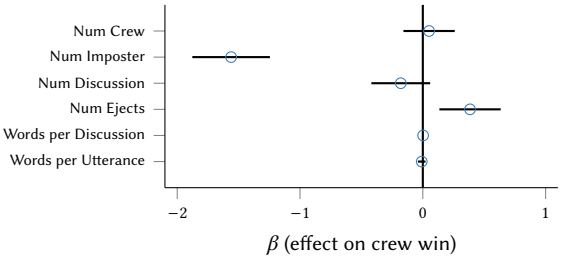

Figure 5: Logistic regression coefficients predicting crew victory (vs. impostor victory). Points show log-odds estimates with 95% confidence intervals. Crew size and the number of ejections are positively associated with crew wins, whereas additional impostors sharply reduce success. Communication frequency and verbosity have little or no effect, indicating that talk alone is insufficient without decisive collective action.

occur earlier or later in the game and how crew size affects their frequency and spacing.

4.1 Results

Figure 3 summarizes win outcomes by role. Across all configurations, impostors win substantially more often, and this advantage grows with the number of impostors. The pattern reflects majority–minority coordination dynamics: impostors face a simpler task of concealment, while crew victory requires both task completion and accurate majority voting.

Figure 4 shows when discussions and ejections occur. Successive meetings appear at roughly constant rates across rounds. Larger crews delay the first discussion and ejection, since more agents and tasks must be observed before suspicion accumulates. Crew size therefore stretches the tempo of play but does not alter its overall shape.

To relate communication to success, we estimated a logistic regression predicting crew victory ($Y = 1$) versus impostor victory ($Y = 0$). Predictors included crew size, number of impostors, number of discussions, number of ejections, and measures of verbosity (average words per discussion and per utterance). Results are shown in Figure 5.

Several patterns emerge:

- **Structural imbalance dominates.** The number of impostors has a large negative effect on crew win probability, confirming the inherent difficulty of coordination under hidden adversaries.
- **Ejections matter more than talk.** Crews that translate deliberation into successful ejections are far more likely to win. Discussion frequency alone has a negligible effect, suggesting that opportunities to speak are not enough.
- **Communication quantity vs. quality.** Verbosity measures (words per discussion, words per utterance) have near-zero coefficients. Agents speak reliably, but longer dialogue does not improve outcomes.
- **Marginal benefit of crew size.** Larger crews show slightly higher win rates, consistent with redundancy in both task completion and deception detection.

Taken together, these results show that agents communicate frequently and consistently, yet communication volume has limited impact on strategic success. What matters is not how much they talk but whether talk leads to coordinated action. This distinction between communication as expression and communication as coordination is what our simulation makes empirically testable.

5 RQ2: REALIZATION OF SPEECH ACTS

Having established that overall communication volume does not predict success, we next examine what kinds of communicative acts LLM agents produce and how these acts vary by role.

Speech act theory classifies utterances by communicative function rather than literal form [2, 33]. Table 1 summarizes five classical categories: directives, representatives, commissives, expressives, and declarations. In cooperative problem-solving, directives typically dominate as players coordinate action, whereas deception often involves representatives as impostors deny, justify, or construct false accounts.

LLM agents may distort these patterns due to their training distributions and safety constraints. Because training data contain abundant directive language (*e.g.*, suggestions or instructions), LLMs may default to directive forms even when other acts are contextually appropriate. Commissives should be rare since promises have no binding force in simulation. Representatives may increase among impostors through denials or excuses, but safety training could suppress overt lies. Expressives and declarations are expected

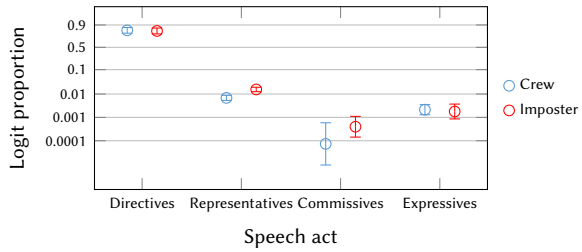

Figure 6: Distribution of speech act types by role (Crew vs. Impostor). Error bars show 95% confidence intervals. Most utterances are directives, with impostors using slightly more representatives.

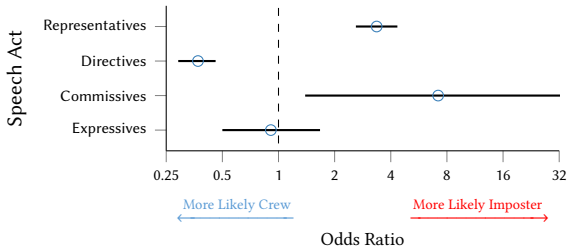

Figure 7: Odds ratios (Impostor vs. Crew) for each speech act with 95% confidence intervals. Values above 1 indicate greater prevalence among impostors.

to be scarce because the game offers few opportunities for emotional display or institutional change.

5.1 Speech Act Coding and Reliability

To select a reliable classifier, a random sample of 50 meeting-phase utterances was annotated by two human coders, Gemini, and ChatGPT. Human–human agreement reached 72%, Gemini–human agreement 72%, and Gemini–ChatGPT agreement 62%. Because Gemini's labels matched human judgments as closely as humans matched each other, we used it for large-scale annotation. The exact prompts used for both Gemini and ChatGPT classification are provided in the Appendix in the online version of this paper.

All meeting-phase utterances were then labeled into the five speech act categories using Gemini. To test stability, the classifier was run three times on the full set of 34,882 valid utterances. Labels were identical in 40.9% of cases, agreed in two of three runs for 58.8%, and disagreed completely for only 0.3%. These results indicate minor stochastic variation but high overall consistency, supporting the use of Gemini classifications for aggregate analysis.

5.2 Results

Across all games, 98% of utterances were directives, consistent with task-oriented communication. Impostors produced a higher proportion of representatives (1.7%) than crewmates (0.5%), while crewmates relied more heavily on directives (99.3% vs. 98.1%). A chi-squared test confirmed that role and speech act category were significantly associated ($\chi^2(3) = 103.85$, $p < .001$).

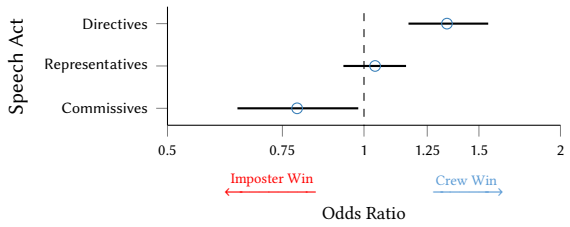

Figure 8: Odds ratios predicting crew victory from speech act with 95% confidence intervals. Values above 1 indicate higher likelihood of Crew victory.

Pairwise proportion tests showed that impostors were more likely to use representatives ($z = 9.80$, $p < .001$) and slightly more likely to use commissives ($z = 2.76$, $p = .006$). Expressives showed no reliable difference ($p = .77$). Figure 7 summarizes these contrasts.

Speech act composition also predicted game outcomes ($\chi^2(3) = 67.8$, $p < .001$). In the logistic regression model, expressives served as the baseline category. Games with higher proportions of directives were more likely to result in crew victories ($\beta = 0.30$, OR = 1.34, $p < .001$), while commissives slightly reduced success ($\beta = -0.23$, OR = 0.79, $p = .031$). Representatives had no effect. Directive-heavy discussions thus signal effective coordination, not deception.

5.3 Speech Acts Before Ejection

Before ejection impostors shifted slightly toward representative speech (1.2%) and away from directives (98.8%), whereas crewmates remained directive (99.3%). A chi-squared test confirmed a modest but significant difference ($\chi^2(3) = 8.73$, $p = .033$). Impostors facing suspicion thus spoke more in fact-asserting or defensive modes, while innocent players maintained task-oriented language.

5.4 Cross-Role Effects

To test whether speech styles carried over between rounds, we compared changes in average speech act proportions. Crewmates showed no measurable shift ($\Delta = -0.001$), and impostors only a slight rise in representatives ($\Delta = +0.004$). Round-to-round correlations were weak and inconsistent ($|r| \approx .4$), and follow-up event studies found no stable cross-role contagion. In short, impostor defensiveness did not influence how crewmates spoke in subsequent rounds, nor vice versa.

6 RQ3: DECEPTIVE STRATEGIES

Deception involves different ways of misleading others. Following interpersonal deception theory and prior taxonomies [5, 18], we classify deceptive utterances into three categories: *falsification* (stating false information), *concealment* (withholding relevant facts), and *equivocation* (using vague or noncommittal language). Table 2 summarizes these forms with examples drawn from gameplay.

Table 2 outlines three broad forms of deceptive communication: falsification, concealment, and equivocation. In human discourse, these differ in how much information is provided, how truthful that information is, and whether misleading is intentional. For language model agents, similar surface patterns can arise not from intent but from optimization pressures. Modern LLMs are trained to be

Table 2: Forms of Deceptive Communication Observed in Agent Dialogue

Type	Definition	Example utterance	Typical strategic goal
Falsification	Stating information known to be false.	*"I was in Medbay the whole time"* (agent was actually in Electrical after a kill).	To create a counter-factual alibi; direct denial or fabrication.
Concealment	Withholding or omitting relevant facts without asserting falsehood.	*"I finished my tasks quickly"* (omits that a kill occurred nearby).	To avoid incrimination by limiting disclosure; passive deception.
Equivocation	Using vague, ambiguous, or non-committal language that misleads while remaining technically true.	*"I was near Storage earlier, but I didn't really see what happened."*	To diffuse suspicion, appear cooperative, or redirect focus without lying.
Missing / Uninterpretable	Incomplete or malformed output preventing classification.	*"I… uh… think maybe?"* or truncated/empty generations.	Not intentional deception per se; reflects processing failure or uncertainty.

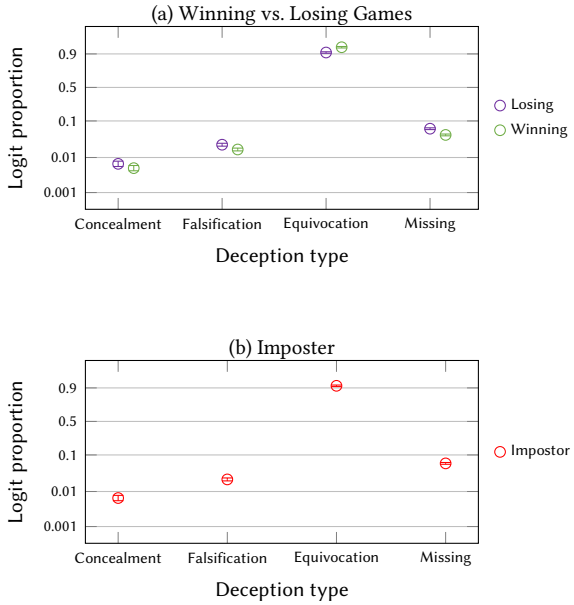

Figure 9: Logit-transformed proportions of each deception type by (a) crewmate game outcome and (b) player role. Higher values correspond to more frequent deception types (logit of proportion). Equivocation dominates across conditions, with crewmates winning showing slightly higher rates.

plausible, helpful, and safe; these objectives can reward *strategic ambiguity*. Guardrails such as reinforcement learning from human feedback (RLHF) [3, 27] penalize overt falsehoods yet often reinforce behaviors that maintain social acceptability—hedging, omitting uncertain details, or deflecting responsibility. Recent work shows that models fine-tuned for safety sometimes shift from falsification to vaguer forms of misleading speech [11, 13, 15].

This asymmetry makes deception an alignment problem rather than a purely ethical one. Safety training constrains explicit lying but leaves concealment and equivocation largely untouched—or even incentivized—as low-risk alternatives. In our sandbox, impostor agents rarely fabricate facts; instead they mislead through omission or ambiguity, a pattern consistent with prior human deception research [5, 18, 36]. Studying such emergent behaviors clarifies how "aligned" models can still act as effective misleaders when cooperation and plausibility are rewarded more strongly than veracity.

6.1 Deception Coding and Reliability

As with speech act labeling, deception categories were assigned automatically using the Gemini classifier, selected for its strong agreement with human judgments. A random sample of 50 utterances was annotated by two human coders, Gemini, and ChatGPT. Human–human agreement reached 73%, Gemini–human agreement 86%, and Gemini–ChatGPT agreement 64%, indicating that Gemini aligned most closely with human interpretations of deceptive language.

To assess internal stability, the classifier was run three times on the full dataset. Across runs, 87.2% of utterances received identical labels, 11.5% agreed in two of three, and only 1.3% differed completely. Pairwise agreement ranged from 87.8% to 96.9%, with $\kappa = 0.83$, confirming high consistency and minimal stochastic variation. These results demonstrate that Gemini's deception classifications are both reproducible and comparable to human reliability for aggregate analysis.

6.2 Results

Across all games, deception was overwhelmingly dominated by *equivocation*—vague or misleading statements that maintain plausible deniability. As summarized in Fig. 9, equivocation accounted for 91.2% of all deceptive utterances, compared with only 2.2% falsifications, 0.7% concealments, and 6.0% unclassified or missing cases. This pattern suggests that agents prefer misleading ambiguity over explicit falsehoods, consistent with the shift from falsification to subtle misdirection found in recent work on LLM deception [13, 15, 18].

Comparison of winning and losing games. We next examined whether deception style varied by outcome. Winning games exhibited a slightly higher proportion of equivocation (93.4%) and less falsification (1.7%) than losing games (90.5% and 2.3%, respectively). However, a chi-squared test comparing the overall distribution of deception types by outcome was not significant ($\chi^2(3) = 4.42$, $p = .22$), indicating that differences in deception composition alone did not reliably predict success.

Deception intensity and social pressure. Although deception type did not differ by outcome, its *intensity* increased with the level of social conflict. Across games, the frequency of all deception types correlated positively with the number of ejections, most strongly for equivocation ($r = .56$, $p < .001$). This relationship suggests that as suspicion mounts, impostor agents produce more evasive or non-committal speech—an adaptive linguistic response to social threat.

Table 3: Speech–Deception Coupling (Spearman ρ).

	Directives	Representatives	Commissives	Expressives
Equivocation	0.57^{***}	0.12^{***}	0.03	-0.01
Falsification	0.09^{**}	0.06	0.05	0.08^{**}
Concealment	0.07^{*}	0.05	0.05	-0.02

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Positive values indicate that games with higher rates of the deception type also have higher shares of the speech act.

Table 4: Relationships between deception and game outcomes. Left: Spearman correlations with number of ejections. Right: Logistic regression predicting victory (1=win, 0=loss).

	Spearman Correlation		Logistic Regression		
Deception type	ρ**(Ejections)**	p	β	p	**Odds ratio**
Concealment	0.09^{**}	.002	−.04	.91	0.96
Falsification	0.11^{***}	.0002	−.20	.33	0.82
Equivocation	0.57^{***}	$< .001$	0.00	.82	1.00
Missing	–	–	−.26	.06	0.77

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Positive correlations indicate that the deception type increases with social tension (ejections). Negative β values in the logistic model indicate reduced odds of winning.

Predicting impostor victory. To assess whether deception type directly predicted game outcome, we fit a logistic regression with win/loss as the dependent variable and the counts of each deception type as predictors. None of the coefficients were significant ($|\beta| < .25$), and the overall model showed very little predictive power. Thus, while deception was common and contextually adaptive, it did not on its own determine victory or defeat.

Taken together, these findings portray model impostors as *strategic misleaders*—preferring subtle ambiguity over outright falsehood, and increasing that ambiguity when under threat.

6.3 Speech and Deception Dynamics

We next examined how the linguistic structure of dialogue relates to deceptive behavior and game outcomes. Deception was measured in three forms (*concealment*, *falsification*, and *equivocation*) and compared with the relative frequency of the four major speech acts: (*directives*, *representatives*, *commissives*, and *expressives*). This analysis asked whether deception manifests as a shift in linguistic function (e.g., from coordination to justification) or as a change in intensity within agents' normal speech patterns.

Coupling between speech and deception. Table 3 shows that deceptive behavior scales with overall communicative activity. All three deception types correlated most strongly with directive speech, indicating that agents tend to deceive while continuing to use task-oriented language rather than switching to overt justification or denial. Equivocation also showed a modest association with representative speech, consistent with hedged or qualified statements that blur factual boundaries. Falsification correlated weakly with expressive speech, suggesting occasional emotional denials. Overall, deception appeared as an *intensification* of ordinary communication rather than a structural shift in how agents speak.

Deception efficiency. As summarized in Table 4 (left), all deception types increased with the number of ejections, linking deceptive language to social tension within the game. Equivocation showed by far the strongest association, rising sharply as suspicion intensified. Yet, when predicting overall victory (Table 4 (right)), none of the deception forms were significant predictors. Falsification and silence trended negative, whereas equivocation was neutral—a pattern consistent with deception as a *defensive* rather than *winning* strategy. These findings suggest that overt lying rarely benefits agents under scrutiny, but vague or evasive speech provides a low-risk means of maintaining credibility even as collective suspicion grows.

7 CONCLUSIONS

This study examined how LLM agents communicate and deceive within the structured environment of the social deduction game *Among Us*. Across 1,100 simulated games, agents produced rich dialogue that could be reliably categorized using classical frameworks from speech act theory and interpersonal deception theory.

Our results show that LLM agents communicate frequently but overwhelmingly through directive speech, reflecting their training on cooperative and instruction-oriented text. Impostors used slightly more representative statements, especially when under suspicion, suggesting a limited but detectable linguistic adaptation to their roles in the game. Deceptive speech was dominated by equivocation rather than falsification, offering no consistent advantage in outcomes. These findings indicate that model deception emerges as a defensive behavior grounded in ambiguity, not deliberate lying.

By linking symbolic theories of communication with large-scale LLM behavior, this work provides an empirical bridge between classical multi-agent models and contemporary generative agents. Understanding how and why LLMs mislead under competitive incentives is essential for designing systems that communicate transparently, coordinate effectively, and maintain user trust in multi-agent and human–AI settings.

7.1 Limitations and Future Work

Our study isolates linguistic deception in a text-based adaptation of *Among Us*, omitting nonverbal cues and the richer social context of human interaction. For this, our experiments used only a single underlying model architecture (Llama 3.2), so it is unknown if other LLM models or training paradigms would offer the same results. Categorizations were done using the Gemini probe, which, while internally consistent, may not capture the full nuance of communicative intent.

Future work should incorporate mixed groups of human and AI agents playing alongside one another to test how well LLM agents can deceive or detect deception in social settings where such behavior is expected, like a modified Turing test for strategic and deceptive communication.

ACKNOWLEDGMENTS

We would like to thank Kristofer Ulanday for his contributions to our data processing. This research was funded by a grant from the ND-IBM Technology Ethics Laboratory.

REFERENCES

[1] Sarit Adhikari, Piotr J Gmytrasiewicz, D Wang, R Falcone, and J Zhang. 2021. Telling Friend from Foe-Towards a Bayesian Approach to Sincerity and Deception.. In *TRUST@ AAMAS*.

[2] John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[4] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. 2000. Developing multi-agent systems with JADE. In *International workshop on agent theories, architectures, and languages*. Springer, 89–103.

[5] David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory* 6, 3 (1996), 203–242.

[6] Steffi Chern, Zhulin Hu, Yuqing Yang, Ethan Chern, Yuan Guo, Jiahe Jin, Binjie Wang, and Pengfei Liu. 2024. Behonest: Benchmarking honesty in large language models. *arXiv preprint arXiv:2406.13261* (2024).

[7] David Christian and R Michael Young. 2004. Strategic deception in agents. In *Autonomous Agents and Multiagent Systems, International Joint Conference on*, Vol. 2. IEEE Computer Society, 218–226.

[8] 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, et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074.

[9] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. 1994. KQML as an agent communication language. In *Proceedings of the third international conference on Information and knowledge management*. 456–463.

[10] ACL FIPA. 2009. FIPA Agent Communication Language Specifications. Available at http://www.fipa.org/specs/.

[11] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. Detecting Strategic Deception Using Linear Probes. *arXiv preprint arXiv:2502.03407* (2025).

[12] Satvik Golechha and Adrià Garriga-Alonso. 2025. Among us: A sandbox for measuring and detecting agentic deception. *arXiv preprint arXiv:2504.04072* (2025).

[13] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093* (2024).

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

[15] E Hubinger, C Denison, J Mu, M Lambert, M Tong, M MacDiarmid, T Lanham, DM Ziegler, T Maxwell, N Cheng, et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training, 2024. *arXiv preprint arXiv:2401.05566* (2024).

[16] Nhat-Minh Huynh, Hoang-Giang Cao, I Wu, et al. 2024. Multi-Agent Training for Pommerman: Curriculum Learning and Population-based Self-Play Approach. *arXiv preprint arXiv:2407.00662* (2024).

[17] Nusrath Jahan and Johnathan Mell. 2025. Decoding Negotiation Dynamics: The Impact of Opponent Identity and Privacy on Strategy, Deception, and Emotional Transparency in Human-Agent Interaction. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2562–2564.

[18] Cameron R Jones and Benjamin K Bergen. 2024. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv preprint arXiv:2412.17128* (2024).

[19] Seth Karten, Siva Kailas, Huao Li, and Katia Sycara. 2023. On the role of emergent communication for social learning in multi-agent reinforcement learning. *arXiv preprint arXiv:2302.14276* (2023).

[20] Sarit Kraus. 1997. Negotiation and cooperation in multi-agent environments. *Artificial intelligence* 94, 1-2 (1997), 79–97.

[21] Yannis Labrou and Tim Finin. 1997. Semantics for an agent communication language. In *International Workshop on Agent Theories, Architectures, and Languages*. Springer, 209–214.

[22] Huao Li, Hossein Nourkhiz Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi Pari, Charles Lewis, and Katia Sycara. 2024. Language grounded multi-agent reinforcement learning with human-interpretable communication. *Advances in Neural Information Processing Systems* 37 (2024), 87908–87933.

[23] Arnau Mayoral-Macau. 2025. Environment-Centered Design of Ethical Environments. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2953–2955.

[24] Dung Nguyen, Hung Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. 2025. Navigating Social Dilemmas with LLM-based Agents via Consideration of Future Consequences. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 2693–2695.

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

[26] Samir Okasha. 2006. *Evolution and the levels of selection*. Clarendon Press.

[27] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[28] Liviu Panait and Sean Luke. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems* 11, 3 (2005), 387–434.

[29] Iyad Rahwan. 2005. Guest editorial: Argumentation in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 11, 2 (2005), 115–125.

[30] Ştefan Sarkadi and Peter R Lewis. 2024. The triangles of dishonesty: modelling the evolution of lies, bullshit, and deception in agent societies. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[31] Stefan Sarkadi, Peidong Mei, and Edmond Awad. 2023. Should My Agent Lie for Me? Public Moral Perspectives on Deceptive AI. In *International Conference on Autonomous Agents and Multiagent Systems*. Springer, 151–179.

[32] Bidipta Sarkar, Warren Xia, C Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. *arXiv preprint arXiv:2502.06060* (2025).

[33] John R Searle. 1969. Speech acts: An essay in the philosophy of language. *Cambridge University* (1969).

[34] Wenjie Shen. 2025. Emergent Language in Multi-Agent Systems: A Multi-Task Learning Approach. In *Proceedings of the 2025 International Conference on Big Data and Informatization Education*. 539–544.

[35] Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots* 8, 3 (2000), 345–383.

[36] Aldert Vrij. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley.

[37] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint arXiv:2310.01320* (2023).

[38] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2023. Defining deception in structural causal games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2902–2904.

[39] Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

[40] 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* (2023).