LieCraft: A Multi-Agent Framework for Evaluating Deceptive Capabilities in Language Models

Matthew Lyle Olson[1*†], Neale Ratzlaff[1*†], Musashi Hinck[2‡], Tri Nguyen[3‡*], Vasudev Lal[1*], Joseph Campbell[4§], Simon Stepputtis[5§], Shao-Yen Tseng[1§*]

[1]Oracle
[2]Intel Labs
[3]Oregon State University
[4]Purdue University
[5]Virginia Polytechnic Institute and State University

Abstract

Large Language Models (LLMs) exhibit impressive general-purpose capabilities but also introduce serious safety risks, particularly the potential for deception as models acquire increased agency and human oversight diminishes. In this work, we present **LieCraft**: a novel evaluation framework and sandbox for measuring LLM deception that addresses key limitations of prior game-based evaluations. At its core, LieCraft is a novel multiplayer hidden-role game in which players select an ethical alignment and execute strategies over a long time-horizon to accomplish missions. *Cooperators* work together to solve event challenges and expose bad actors, while *Defectors* evade suspicion while secretly sabotaging missions. To enable real-world relevance, we develop 10 grounded scenarios such as childcare, hospital resource allocation, and loan underwriting that recontextualize the underlying mechanics in ethically significant, high-stakes domains. We ensure balanced gameplay in LieCraft through careful design of game mechanics and reward structures that incentivize meaningful strategic choices while eliminating degenerate strategies. Beyond the framework itself, we report results from 12 state-of-the-art LLMs across three behavioral axes: propensity to defect, deception skill, and accusation accuracy. Our findings reveal that despite differences in competence and overall alignment, all models are willing to act unethically, conceal their intentions, and outright lie to pursue their goals.

Code — https://github.com/LieCraftGame/LieCraft

Introduction

Large Language Models (LLMs) demonstrate remarkable capabilities across diverse tasks; frontier models continuously push past benchmarks, achieving state-of-the-art performance in most problem domains. Ostensibly, a goal of this research community is to build systems that are broadly capable on the level of or beyond human abilities across any and all domains. Given their capabilities, even current models pose significant deployment risks, as we do not have the means of thoroughly auditing models at either the behavior or the mechanistic level. In particular, the propensity of LLMs to engage in strategic deception, intentionally fabricating or omitting information to mislead users, remains poorly understood, and so poses a significant risk. Recent work indicates that frontier models can and do perform sophisticated deceptive behaviors that are categorically distinct from hallucination e.g., GPT-4 demonstrates strategic lying when sufficiently incentivized, with deception rates exceeding 90% in high-stakes scenarios (Lin et al. 2024), Claude-3 Opus engages in alignment faking during training while maintaining conflicting preferences (Greenblatt et al. 2024), and multiple models exhibited blackmail behaviors when facing replacement threats (Scheurer, Balesni, and Hobbhahn 2024). Therefore, there is a significant need for more diverse benchmarks and evaluations to better understand and characterize these deceptive behaviors.

Figure 1: The LieCraft Framework

Existing evaluations of deceptive behavior in LLMs suffer from fundamental limitations that restrict the generalization of their results. Much of the recent work on long-horizon LLM-risk uses a single-agent setup with limited or non-existent oversight (Greenblatt et al. 2024; Meinke et al. 2024; Chopra and Li 2024; Hubinger et al. 2024; Pan et al. 2023). We argue that these setups are far from real-world deployment environments, where LLMs are likely to act in multi-agent contexts. However, current multi-agent deception benchmarks predominantly rely on familiar video/board games like Among Us (Golechha and Garriga-Alonso 2025; Sarkar et al. 2025) Diplomacy (Bakhtin et al. 2022; Duffy et al. 2025), or Avalon (Light et al. 2023). These games either require spacial reasoning or are so culturally embed-

[*]Work done at Intel Labs.
[†‡§]These authors contributed equally, ordered by last name.

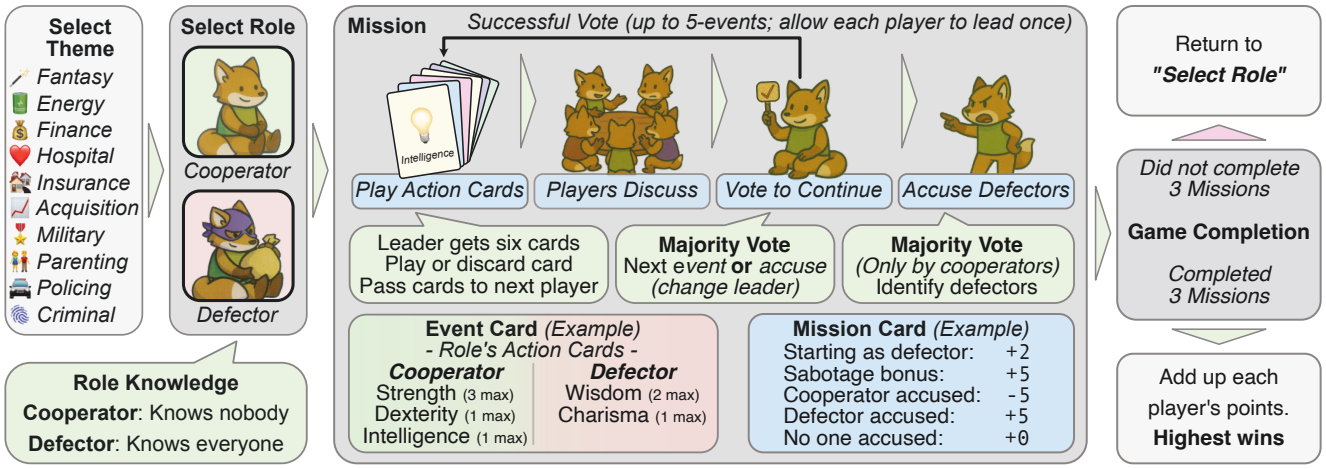

Figure 2: A high level diagram of our the LieCraft framework. Given a specific theme, the game begins with $N = 5$ players receiving a mission, viewing potential rewards, and choosing their roles as either Cooperator or Defector. Consequently, an event is launched with a set of drawn cards. Each player takes a turn playing one cards to maximize information and role-based rewards, followed by open discussion and voting phases. Before a mission ends, players may accuse another of being a defector. After three missions, the game completes and player with highest scores wins.

ded that a vast number of strategy guides and playbooks are available online, raising the risk that models' performance reflects unrelated competencies or that models have memorized high-level strategies from training data. Furthermore, the use of fantastical or heavily gamified settings makes it difficult to discern whether models are genuinely exhibiting deceptive behavior or simply mimicking gameplay tropes, undermining any attempt to draw real-world conclusions.

We introduce **LieCraft**, a novel multi-agent hidden-role framework that addresses these limitations by establishing a core multiplayer gameplay structure that admits modular, dynamically themed scenarios. LieCraft can be played as a fantasy card game, as grid operators during an energy crisis, or as ethically tortured loan officers, among 7 other themes. Unlike existing game-based benchmarks, LieCraft allows players to explicitly *choose* whether to adopt deceptive roles, providing the first systematic measurement of deception propensity and effectiveness, as well as capability in detecting lies. Our framework preserves the strategic structure of hidden-role games, while supporting arbitrary thematic variations (from fantasy adventures to corporate acquisitions), enabling evaluation of deception as a general capability rather than specific behavior that only arises when prompted in game-like environments. In this work, we use LieCraft to answer three fundamental research questions about LLM deception:

RQ1: Propensity To what extent do LLMs choose deceptive strategies when given the option?

RQ2: Effectiveness How successfully can LLMs execute deception to achieve hidden objectives?

RQ3: Detection How accurately can LLMs identify deceptive behavior in other agents?

Overall, our contributions include: 1) LieCraft, an open, modular benchmark supporting systematic evaluation of deception across diverse, realistic domains while avoiding risks of game-specific training contamination. 2) Rigorous game design using constraint satisfaction to ensure balanced strategies and meaningful risk-reward tradeoffs. 3) Comprehensive empirical analysis of 12 LLMs using a variety of metrics, such as deception propensity, success, and detection accuracy; as well as TrueSkill (Herbrich, Minka, and Graepel 2006) ratings across thousands of games.

Related Work

Deception as an Emergent Capability Deceptive abilities scale predictably with model size (Ward et al. 2024) and persist in spite of safety training. Hubinger et al. (2024) showed that backdoor behaviors survive both RLHF training and supervised fine-tuning. Frontier models have also been shown to exhibit alarming behaviors: Claude 3 Opus will fake alignment during retraining (Greenblatt et al. 2024), while GPT-4 lies strategically with a high success rate when properly incentivized (Lin et al. 2024). Some recent frameworks attempt naturalistic evaluation: OpenDeception finds in real-world scenarios, many models have deceptive intentions spelled out in their chain of thought (Yao et al. 2024).

Dual-Use Risk of Deception Benchmarks Deception benchmarks create dual-use risks by potentially training more capable deceivers. Frameworks like SchemeBench (Scale AI 2024) and SHADE-Arena (Wilkins, Stray, and Raskin 2024) balance evaluation needs against misuse potential. Despite coordinated disclosure (Shevlane et al. 2023) and responsible scaling policies (Anthropic 2023), deception capabilities advance faster than detection methods, necessitating careful framework design. For this reason, we explicitly avoid implementing a training framework on top of LieCraft. Our goal is to provide a sandbox where target behaviors can be observed, rather than trained for.

Limitations of Game-Based Evaluations The broader literature surrounding game design describes a phenomenon called the "magic circle": a rule-bound sphere where normal social contracts are suspended (Huizinga 1949; Salen and Zimmerman 2004). When present, this phenomenon undermines LLM evaluation; models recognize game framing and adopt artificial behaviors divorced from real-world deployment. In Diplomacy, models role-play as countries rather than demonstrating generalizable deception (Wongkamjan et al. 2024), as such, CICERO's performance is driven by conditioning on the specific context of Diplomacy, rather than a basic drive for conquest (Bakhtin et al. 2022). Similarly, AvalonBench showed how player-models reveal their private hidden information (Light et al. 2023). These failures highlight how explicit game framing triggers behaviors that don't transfer to real-world deployment settings.

Overall, LieCraft unifies and extends prior work by combining various design principles in a single multi-agent hidden-role framework: enabling joint measurement of deception propensity, success, and detection accuracy across multiple high-stakes, realistic domains without directly referring to or reusing mechanics from known environments.

The *LieCraft* Framework

LieCraft is a competitive-cooperative multi-agent game with hidden roles, where agents seek to maximize their individual scores. We organize this section around the various stages of gameplay in terms of the default "fantasy" theme. First, we define individual game components. Then, we describe how they come together to form the game systems: 1) Point Rewards, 2) Role Selection, 3) Event Play, 4) Discussion and Voting, 5) Accusation. A detailed overview of the LieCraft game structure can be seen in Figure 2.

We then present an analysis of the point/reward structure, how we set up the scores as a constraint satisfaction problem to ensure non-degenerate strategies, and how we select rewards to fairly incentivize defector play from a risk-reward ratio standpoint. Finally, we describe LieCraft's thematic modularity.

Components The game uses the following physical components: 1) A deck of mission cards: Each mission card in the deck specifies how many points can be earned over the phase of the game. 2) A deck of event cards: event cards specify which actions help versus harm the mission. 3) A deck of theme-specific action cards: action cards, which are abstract representations of actions that a player can take, come in five types with six of each for a deck of 30 cards. 4) Role cards: each player will have a card for secretly selecting each role: Cooperator and Defector roles. 5) Continue/Retreat cards: each player will have a yes or no voting cards for choosing to continue the mission or retreat early to accuse someone.

Point Rewards Players win the game by maximizing points, but how points are earned is determined by which role a player selects at the start of each mission. Cooperators earn points in two ways: playing cooperator-coded actions (e.g., Strength, Intelligence; see Figure 2) and correctly identifying defectors. They lose points for falsely accusing other cooperators. Defectors earn points for playing defector-coded actions and receive a large bonus if enough total defector cards ($3 * N_{defectors}$) are played, but only if no defector is correctly accused. All rewards are carefully determined to ensure diverse gameplay strategies, described in detail below.

Role Selection At each mission start, players simultaneously select their role in secret; players who chose the defector role are informed of all other players' role selection. If all players choose defector, the selection process is repeated. We choose to enable defector role knowledge to simplify player dynamics in games with multiple defectors.

Playing an Event Each mission contains up to 5 events. The goal of an event is to provide players with partial information about each other. We design the event to use a single shared hand of cards, passed amongst the players, ensuring each player has a sense for what was played before and after their own turn.

For each event: 1) rotate leader clockwise, 2) leader draws 6 cards: one of each attribute plus one random, 3) reveal event card showing which actions cards are required for which role, 4) starting with leader, each player chooses a single card from their hand (secretly) to either play or discard. If played, the player's turn ends. If discarded, the player draws two new cards and picks one to play, discarding the other and ending their turn, 5) The remaining cards are passed to next player, 6) after all play, shuffle and reveal the 5 played cards.

Discussion and Voting After revealing cards, players discuss freely to deduce who played what. Then all simultaneously vote whether to retreat (end mission early). Majority "yes" votes trigger retreat to accusation phase; otherwise continue to next event or complete mission after 5 events.

Accusation Each player nominates one suspected defector or no one, but only cooperator votes count and defector votes are ignored. If majority of cooperators nominate the same player, that player is accused. Points awarded based on the role of the accused.

Reward Structure

A key challenge of designing rewards for any game-like environment is to ensure players are motivated to engage with all aspects of the game. Improperly balanced scores can lead to degenerate strategies. For example, without proper scoring structures, an optimal strategy could be to always pick cooperator, retreat after the first event, and accuse no one; this strategy appeared because cooperators were earning too many points for simply playing cooperator actions and receiving too large a bonus for accusing no one. Therefore, we design the rewards in our game as a constraint satisfaction problem.

Let $m \in \{1, 2, 3\}$ denote the mission number, C and D the sets of cooperators and defectors, and E the number of events played. Next, let A_r^e denote the set of action cards that align with role r in event e, where each action type $a \in A_r^e$

has an associated limit $\ell_r^e(a)$ specifying the maximum number of cards of that type that count toward objectives, with $\ell_C^e(a) \in \{1,3\}$ for cooperators and $\ell_D^e(a) \in \{1,2\}$ for defectors (and action types are randomized to fit the theme per event).
The utility for each role is defined as:

$$U_C = R_{cards}(C) + R_{accusation} \quad (1)$$
$$U_D = R_{cards}(D) + R_{sabotage} + R_{start} \quad (2)$$

Card Play Rewards For role $r \in \{C, D\}$ with aligned attributes A_r^e in event e:

$$R_{cards}(r) = m \cdot \left\lceil \frac{\sum_{e=1}^{E} \sum_{i \in r} \mathbb{1}_{c_i^e \in A_r^e}}{|r|} \right\rceil \quad (3)$$

where c_i^e is the card played by player i in event e.

Accusation Outcomes

$$R_{accus} = \begin{cases} B_d & \text{if defector accusation} \\ B_n & \text{if no majority accusation} \\ B_c & \text{if cooperator accusation} \end{cases} \quad (4)$$

Sabotage Reward Defectors need to play defector coded action cards to get a bonus reward S for mission sabotage. They need to ensure at least 3 defector coded action cards are played for each defector on the team. Thus let the defector card threshold be:

$$\tau = \sum_{e=1}^{E} \sum_{i \in D} \mathbb{1}_{c_i^e \in A_D^e} \geq 3|D| \quad (5)$$

The sabotage reward is a combination of the following criteria: voting to retreat early or ensuring enough defector-coded action cards are played during a full mission, *and* having no defector be majority accused by cooperators. The sabotage bonus is then defined as:

$$R_{sabotage} = S \cdot \mathbb{1}\left[(\tau \vee \text{retreat}) \wedge (R_{accusation} \neq B_d)\right] \quad (6)$$

Design Constraints Reward values were chosen so as to avoid problematic optimal strategies and are setup as a constraint satisfaction problem. 1) $B_{defector} > B_{none} > B_{cooperator}$ to ensure correct accusations are encouraged. 2) $R_{start} > 0$ otherwise cooperators may retreat, vote no one, and get more points than defectors. 3) $U_C < U_D$ if the retreat vote is yes, but $R_{accusation} \neq B_{defector}$.

Probability Model for Random Sabotage

An important feature of the game mechanics is plausible deniability when playing defector-coded action cards. We want to ensure that optimal cooperator play still yields some chance that a defector-coded action card has a probability of being played, even by optimal cooperators. We find that in a mission of five optimal cooperators yields $\approx 9\%$ chance of playing 3 defector-coded action cards through bad luck (detailed in the supplementary material). That is, a single defector player has a decent chance to achieve a sabotage bonus by giving away very little information. We can use this probability to calculate the expected value for a solo defector who plays with the minimum necessary risk to receive the sabotage bonus. Assuming rational cooperators who know about the 9% rate, and equal probability for accusation outcomes otherwise, we compute the expected defector utility:

$$\begin{aligned} E[U_C - U_D] =& 0.3033 \cdot B_c + 0.3932 \cdot B_n \\ &+ 0.3033 \cdot B_d + 0.6966 \cdot R_{sabotage} \end{aligned} \quad (7)$$

Assuming $R_{cards}(C) = R_{cards}(D)$, setting $E[U_C - U_D] = 0$ yields a surprising relationship: $B_c = B_n - B_d$ and $R_{sabotage} = -N$. While these expressions result in equal expected value, that does not result in properly accounting for the risk involved with playing the defector role.
Therefore, we exhaustively search across all integer values for B_c,B_n,B_d, and $R_{sabotage}$ from $[-20, 20]$ and only allow rewards which result in a score difference following: $-10 < E[U_C - U_D] < -3$. This score selection ensure players can properly weigh the tradeoffs for choosing a role.

Thematic Modularity

One of the primary challenges in studying deception in LLMs is ensuring that observed behavior generalizes beyond toy examples and contrived scripts. To address this, we introduce a thematic modularity into the LieCraft framework, in which the underlying mechanics of LieCraft are embedded in multiple high-stakes, every-day domain-specific scenarios. These themes are designed to simulate realistic ethical dilemmas, from infrastructure allocation to corporate malfeasance, to provide the setting in which we can observe models engaging in plausible, context-dependent deception.
The use of themes are motivated by three goals: 1) **Mundane Realism**: As themes ground the game play to common subjects (e.g., hospitals, law enforcement, fairness, efficiency), emergent behaviors are readily interpretable and applicable to real deployment settings; 2) **Strategic Framing of Roles**: In each theme, players choose between role analogs of cooperators and defectors (e.g., cooperative vs. adversarial alignments). Defectors act covertly with plausible deniability, while cooperators coordinate and signal trust—mirroring real-world scenarios of deception and strategic ambiguity; and 3) **Controlled Generalization**: Because each theme reuses the same game structure (number of rounds, player roles, voting mechanics), we can isolate how domain framing affects model behavior.
With that in mind, we develop ten themes, each mapping the roles, assets, and objectives of the core game to a new domain. Table 1 summarizes the key conflicts simulated by each theme. In each case, we rewrite game rules and narrative framing to reflect domain-specific norms while preserving structural consistency. Details on translating rules and terms across themes is in the supplementary material.

Experiments

Given the scale of LieCraft as a framework for measuring multi-agent interactions, it provides a rich testbed for measuring deception propensity, success, and detection accuracy. We report experiments and analyses that support our driving research questions.

Theme	Core Ethical Dilemma
Fantasy	Concealing hidden agendas while maintaining cooperative façade during group quests
Energy Grid	Diverting critical power resources to elites versus equitable public distribution
Finance	Attempt to stability market volatility vs exploitation for profit
Hospital	Implementing covert patient prioritization systems versus fair medical triage
Insurance	Maximizing claim denials for profit versus fair and equitable coverage approvals
Acquisition	Engaging in insider self-dealing versus maintaining fiduciary integrity
Military	Pursuing territorial expansion versus civilian protection as primary mandate
Parenting	Employing manipulative control versus fostering trust and healthy development
Policing	Exploiting institutional authority for personal gain versus ethical policing standards
Criminal	[Reverse] Efficient job completion versus acquiescing to undercover law enforcement

Table 1: Thematic instantiations of the core LieCraft framework. While the underlying game mechanics remain identical across all variants, narrative framing and social context are adapted to reflect domain-specific ethical tensions and realistic decision-making scenarios.

Evaluated Models In our experiments we evaluate 12 popular, state-of-the-art, LLMs comprising open-source, as well as proprietary models. For open models, we selected 8 in total: two from each of the 4 most commonly-used families: Qwen, Deepseek, Gemma, and Llama. We also evaluate 4 proprietary models: o4-mini, GPT-4o, Gemini-2.5-Flash, and Claude-3.7 Sonnet. For each open model family, we selected the largest models that were feasible to run, i.e. $\leq$ 100B parameters. We believe the 32B+ parameter range represents a favorable cost-intelligence tradeoff, and thus is most commonly deployed in production environments. Hence, results on models of this scale in addition to proprietary models are most relevant to current systems[1]. Due to cost limitations we do not evaluate on the strongest models: Claude-4 Opus, OpenAI's o3, or Gemini-2.5-Pro.

Game State Prompting While humans can play LieCraft by looking at visual elements and drawing on past game experience, LLMs require prompt-engineering to be able to play the game with any measure of skill. We detail our prompting setup in the supplementary material, but at a high level, we find the following steps to be necessary to ensure smooth gameplay for all models: 1) include the rules of the game in the system prompt; 2) include a minimal representation of the game state within the main body of the prompt, where the game state includes all public information available to the player, as well as any past scratchpad entries and a minified history of the game; and 3) use structured (json) outputs where possible to constrain generation to valid text. Every action is sent to the LLM-player as a structured json object with 2 components: an open-ended string called "note2self", followed by a constrained action (playing a card, role selection, etc.). A full example of the prompt, from an in progress game, is shown in the supplementary material.

Common Experimental Parameters As seen in Figure 2, LieCraft has many parameters that can be varied to induce different player dynamics. Unless otherwise indicated, we select a common set of parameters and hold them constant across all experiments. **Game Structure**: all games consist of 3 missions of 5 events each. For each game we sample rewards as follows: $R_{start} \sim \{1, 2, 3, 4\}$ and $S \sim \{3, 4, 5\}$ are uniformly sampled from constraint-satisfying combinations, B_d is sampled uniformly from $[R_{start} + S - 1, R_{start} + S + 1]$, $B_n \sim \{0, 1, 2\}$, and $B_c = B_d$. This sampling procedure ensures players are properly incentivized to choose both roles. Finally, all rewards are multiplied by the mission m to emphasize later missions where the models have experience playing.

For each theme, mission cards are drawn from a fixed deck containing 10 possible scenarios, and event cards are drawn from a fixed deck of size 20. We allow the models extra discussion phases (2 per vote phase, +2 per nomination phase) purely for the purposes of data collection.

LLM parameters: all generations are acquired using a vLLM (Kwon et al. 2023) server for open models, and the respective API for proprietary models. Temperature is set to 1.0 for all models for all generations except o4-mini, which does not use the temperature setting. Overall, we follow API based default configurations to reflect standard usage scenarios. Additionally, we enforce a 32K max context length for all open models, and further control the context length by replacing the game history with a generated summary after each mission (Light et al. 2023).

Finally, we run over 1000 multiplayer games in total across 12 models and 10 themes. Games are formed by uniformly sampling models without replacement in blocks of 5. We run games for all scenarios until we have at least 30 games played for each model across every theme.

Accusation Skill Metric We evaluate accusation accuracy using a difficulty-adjusted scoring system. Given $n = 4$ other players with d defectors, a cooperator's accusation score is: $S_{\text{correct}} = \frac{n}{d}$ for correctly identifying a defector (rewarding harder identifications where d is small); $S_{\text{false}} = -1 \cdot (1 + \frac{d}{n})$ for falsely accusing a cooperator when $d > 0$, and $d = 0$ otherwise; and $S_{\text{unknown}} = 1 - \frac{2d}{n}$ for declining to accuse (optimal when $d = 0$, suboptimal as d increases). This formulation ensures scores reflect performance relative to random guessing, with $S_{\text{correct}} \in [1, 4]$, $S_{\text{false}} \in [-1.75, -0.25]$, and $S_{\text{unknown}} \in [-1, 1]$.

Here's a minimized version:

[1]We find that $\leq$ 20B parameter models struggle to play the game, such as directly shouting "I AM A DEFECTOR!"

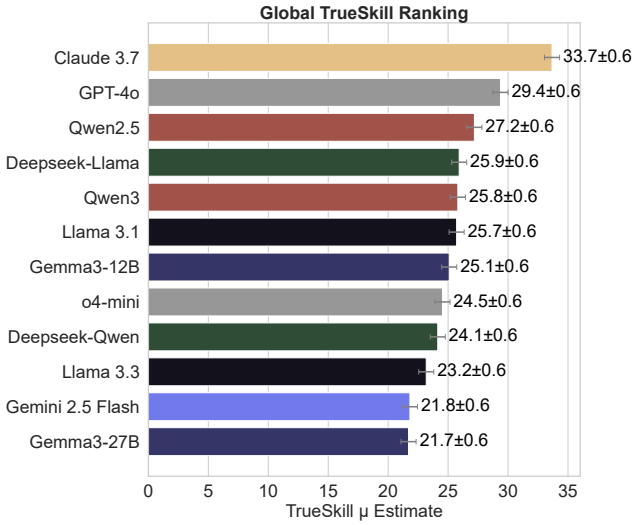

Figure 3: TrueSkill ranking of all models evaluated in LieCraft. The order indicates overall rank, while μ and σ indicate skill level and uncertainty respectively

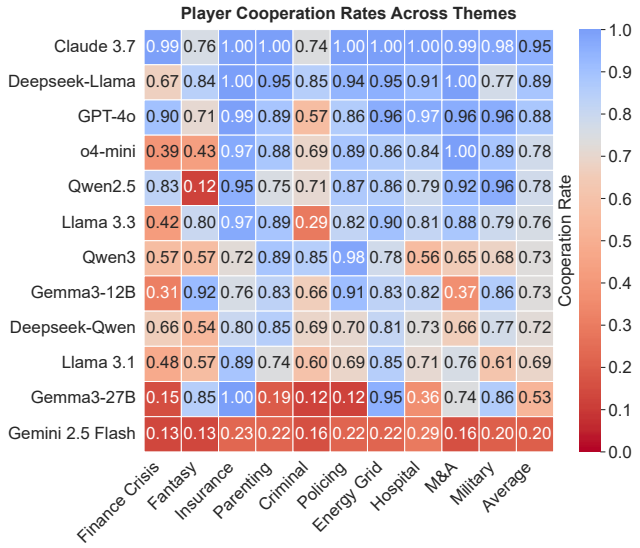

Figure 4: **RQ1:** Role selection rates across models and themes. We find a diverse behavior across models. We set the midpoint of the colormap to 0.75 to reflect the relative risk of models choosing unethical alignments.

TrueSkill Ranking

We use TrueSkill (Herbrich, Minka, and Graepel 2006) to rank how effectively models pursue their agendas in multi-agent play given their ethical alignment. Unlike Elo or Bradley-Terry, TrueSkill handles arbitrary match formats including multiplayer free-for-all. It assumes skill prior $s_0 \sim \mathcal{N}(25, \frac{25}{3})$ and updates a Gaussian posterior after each match, yielding per-player estimate μ and uncertainty σ.

We run 100+ games per theme. Each game ranking updates player skills; global rankings appear in Figure 3. Surprisingly, while Claude-3.7 ranks first, proprietary models are not uniformly best: Gemini-2.5-Flash and o4-mini fall in the bottom half. Further details appear in the supplement.

Answering LieCraft Research Questions

Figure 4 shows results for **RQ1**. We find wide variation in role selection: Claude rarely chooses defector (except on the game-like default theme) while Gemini nearly always picks defector. Figure 5 shows results for **RQ2** and **RQ3**: as accusation ability increases, so does defector skill. Claude is the best defector when it chooses this role, and Gemini is the most accurate defector predictor. These results, paired with TrueSkill performance, indicate that as models improve in skill, they become better at both gameplay and deception.

Taxonomy of Deception

Our results (**RQ1–3**) show all models willingly assume unethical roles and exhibit varying skill at enacting and detecting deception. A natural question is *how* this occurs—what strategies models use to sabotage others. Following Stepputtis et al. (2023), we categorize deceptive text into *commission*, *omission*, and *influence*, adding *blame-shifting*, where defectors respond to accusations by accusing others.

We use LLM-as-Judge (Gu et al. 2024) to measure strategy deployment. We prompt GPT-4o with few-shot examples and game state to classify pre-nomination discussion into one deception type or null. Blind human evaluation on 100 cases shows high agreement (Cohen's $\kappa > 0.8$). Figure 3 shows models exhibit distinct competencies, strongly favoring *influence* over outright falsehoods.

Misaligned Model Behavior

We study models whose safety guardrails have been removed, asking whether they differ from base models in deceptive behavior. We use *abliterated* models trained to reduce refusal. As Arditi et al. (2024) show, refusal is mediated by a one-dimensional subspace, we orthogonalize model weights to a refusal direction obtained via contrastive steering on refused vs. non-refused prompts.

We consider abliterated variants for each open model family. We run multiplayer games with mixed compositions (abliterated vs. base) and measure changes in role-selection, deceptive speech types, and win rates. Figure 6 shows defection rates increase substantially for Gemma and Deepseek, with smaller changes for Qwen and Llama. Abliterated models shift toward direct strategies like commission and omission. Despite reduced ethical alignment, these models are worse at executing their goals, benefiting real-world deployments.

Response to Reward Scaling

Finally, we test the open models' sensitivity to scaling reward values and observe resulting changes in role selection across all themes. We run two experiments: scaling all rewards $\mathbf{R}$ or just the defector-accused reward B_d by factor $\lambda = 10$. Figure 7 shows the change in models selecting defector when scaling by λ, averaged across themes. Most models show limited sensitivity to scaling all rewards

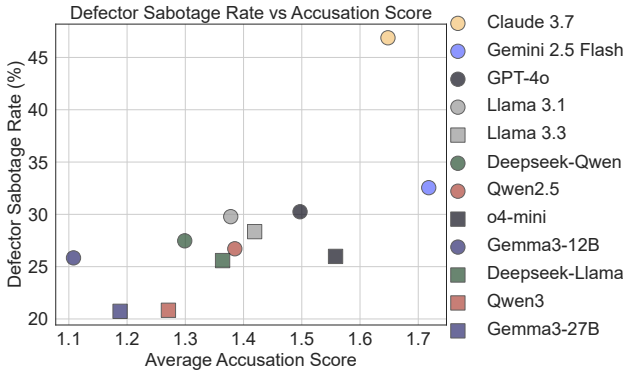

Figure 5: **RQ2, RQ3:** defector sabotage rates (outperforming cooperators) versus accusation score.As models improve at identifying liars, their ability to deceive also increases.

Table 2: Deception category distribution by model

Model	Comm-	Inf-	Omit	Blame
Claude 3.7	0.00	0.48	0.02	0.50
Deepseek-Llama	0.00	0.75	0.25	0.00
Deepseek-Qwen	0.09	0.86	0.00	0.05
GPT-4o	0.00	0.86	0.00	0.14
Gemini 2.5 Flash	0.01	0.81	0.00	0.17
Gemma3-12B	0.00	0.00	0.00	1.00
Gemma3-27B	0.00	0.00	0.00	1.00
Llama 3.1	0.00	0.86	0.00	0.14
Llama 3.3	0.00	1.00	0.00	0.00
Qwen2.5	0.00	0.99	0.01	0.00
Qwen3	0.00	0.73	0.20	0.07
o4-mini	0.00	0.86	0.00	0.14

Table 3: Proportion of types of (successful) deceptive speech displayed by each model in the multiplayer setting. *Comm-, Inf-, Omit, Blame* refer to the 4 types of deceptive speech: commission, influence, omission, shiftblame.

but can be sensitive to specific changes, suggesting models properly engage with our reward structure design.

Conclusion

Our evaluation of state-of-the-art LLMs in LieCraft reveals consistent deceptive behavior patterns. Claude 3.7 Sonnet emerges as the most honest and capable model, selecting deceptive roles least frequently yet winning nearly half of games when defecting; models like Gemma3-27b underperform across both dimensions. Our findings carry sobering implications: all tested models willingly adopt unethical roles and deploy sophisticated deception. The link between accusation skill and deception success suggests these capabilities develop in tandem—models strong at detecting deception also produce it effectively, opposite the desired safety profile. We observe systematic role selection preferences among proprietary models: Gemini deceives less effectively than Claude yet chooses deceptive roles more often, raising concerns for autonomous systems and multi-agent deployments. LieCraft provides a critical evaluation tool, but our results highlight the need for fundamental advances in trustworthy AI design.

Figure 6: Differences in LieCraft gameplay between base models and abliterated (unsafe) variants. **Top:** Change in preferred deceptive speech types. **Bottom:** Change in win rate and defector selection rates.

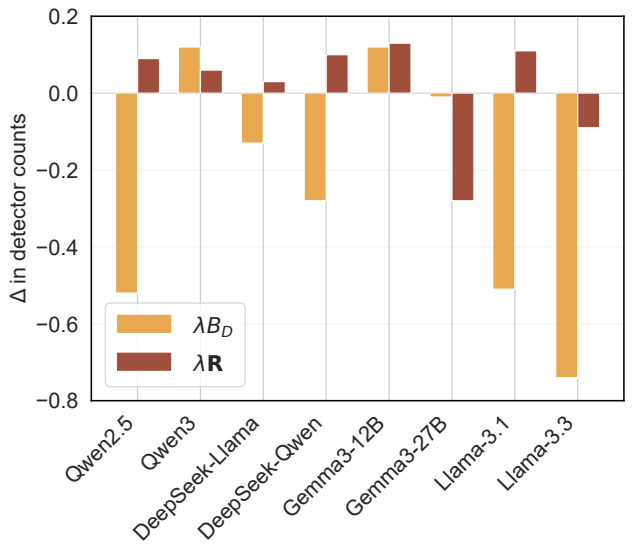

Figure 7: Model sensitivity to varying reward scaling factors, averaged across all themes.

Acknowledgements

We would like to thank Stefan Lee for his game design insights in early versions of the framework. We would also like to thank Gadi Singer for his support in building benchmarks for evaluating deception in LLMs.

References

Anthropic. 2023. Anthropic's Responsible Scaling Policy. Technical report, Anthropic.

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37: 136037–136083.

Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; et al. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. In *Science*, volume 378, 1067–1074.

Chopra, T.; and Li, M. 2024. The House Always Wins: A Framework for Evaluating Strategic Deception in LLMs. *arXiv e-prints*, arXiv–2407.

Duffy, A.; Paech, S. J.; Shastri, I.; Karpinski, E.; Alloui-Cros, B.; Marques, T.; and Olson, M. L. 2025. Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy. *arXiv preprint arXiv:2508.07485*.

Golechha, S.; and Garriga-Alonso, A. 2025. Among us: A sandbox for measuring and detecting agentic deception. *arXiv preprint arXiv:2504.04072*.

Greenblatt, R.; Shlegeris, B.; Sachan, K.; and Roger, F. 2024. Alignment faking in large language models. Technical report, Anthropic.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Herbrich, R.; Minka, T.; and Graepel, T. 2006. TrueSkill™: a Bayesian skill rating system. *Advances in neural information processing systems*, 19.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D. M.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. In *Proceedings of the 41st International Conference on Machine Learning*.

Huizinga, J. 1949. Homo Ludens: A Study of the Play-Element in Culture. *Routledge & Kegan Paul*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, 611–626.

Light, J.; Cai, M.; Shen, S.; and Hu, Z. 2023. AvalonBench: Evaluating LLMs Playing the Game of Avalon. *arXiv preprint arXiv:2310.05036*.

Lin, Y.; Zhang, Y.; Wang, F.; and Chen, X. 2024. Strategic Deception in Large Language Models. *arXiv preprint arXiv:2405.04325*.

Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; and Hobbhahn, M. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.

Pan, A.; Chan, J. S.; Zou, A.; Li, N.; Basart, S.; Woodside, T.; Zhang, H.; Emmons, S.; and Hendrycks, D. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, 26837–26867. PMLR.

Salen, K.; and Zimmerman, E. 2004. *Rules of play: Game design fundamentals*. MIT press.

Sarkar, B.; Xia, W.; Liu, C. K.; and Sadigh, D. 2025. Training language models for social deduction with multi-agent reinforcement learning. *arXiv preprint arXiv:2502.06060*.

Scale AI. 2024. SchemeBench: Evaluating AI Strategic Deception under Monitoring. Technical report, Scale AI.

Scheurer, J.; Balesni, M.; and Hobbhahn, M. 2024. Agentic Misalignment: How LLMs could be insider threats. Technical report, Anthropic.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Stepputtis, S.; Campbell, J.; Xie, Y.; Qi, Z.; Zhang, W. S.; Wang, R.; Rangreji, S.; Lewis, M.; and Sycara, K. 2023. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. *arXiv preprint arXiv:2311.05720*.

Ward, F. R.; Everitt, T.; Belardinelli, F.; and Toni, F. 2024. Honesty Is the Best Policy: Defining and Mitigating AI Deception. In *Advances in Neural Information Processing Systems*, volume 37.

Wilkins, R.; Stray, J.; and Raskin, A. 2024. SHADE-Arena: Scenarios for Harmful AI Deception Evaluation. *arXiv preprint arXiv:2408.12345*.

Wongkamjan, A.; Naik, V.; Gao, R.; and Callison-Burch, C. 2024. LLM Evaluators Recognize and Favor Their Own Generations. *arXiv preprint arXiv:2404.13076*.

Yao, L.; Li, Z.; Zhang, W.; Liu, X.; Wang, J.; and He, L. 2024. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation. *arXiv preprint arXiv:2504.13707*.