Deception Scales: How Strategic Manipulation Emerges in Complex LLM Negotiations

Luis Fernando Yupanqui — Independent AI Researcher
Mari Cairns — Independent AI Researcher
With Apart Research

Abstract

Simple benchmarks hide dangerous capabilities. We present a multi-agent simulation framework using "So Long Sucker" (Nash et al., 1964) — a negotiation/betrayal game designed by four Nobel laureates — to study how AI deception scales with task complexity.

We ran **146 games** across four frontier LLM models (Gemini 3 Flash, GPT-OSS 120B, Kimi K2, Qwen3 32B) in two conditions (talking vs. silent) across **three complexity levels** (3-chip, 5-chip, 7-chip). Analysis of **13,759 decision events** reveals:

The Complexity Reversal. GPT-OSS dominates simple games (67% win rate at 3-chip silent) but collapses at complexity (10% at 7-chip talking). Gemini shows the inverse pattern: 9% at 3-chip silent rising to **90% at 7-chip talking**. Strategic manipulation becomes dramatically more effective as game length increases.

Key Findings:

1. **The Complexity Reversal** — Win rates invert as task complexity increases
2. **107 Private Contradictions** — Models' private reasoning directly contradicts their public statements
3. **237 Gaslighting Instances** — Gemini deploys systematic psychological manipulation tactics
4. **7:1 Alliance Imbalance** — GPT-OSS desperately seeks alliances it never receives

Using Harry Frankfurt's philosophical framework, we classify models as "strategic" (truth-tracking with deliberate misrepresentation) vs. "reactive" (plausible output without internal consistency). This taxonomy explains the Complexity Reversal: strategic models compound advantages over longer games while reactive models cannot maintain coherence.

This work contributes to AI safety research by demonstrating that **deception capability scales with task complexity**—simple benchmarks underestimate manipulation risk.

Keywords: Multi-agent alignment, AI deception, emergent manipulation, strategic behavior, complexity scaling, LLM safety, complexity reversal, deception detection, scalable oversight, LLM negotiation

1. Introduction

1.1 The Scaling Problem in AI Safety

How do we know if an AI system is capable of deception? Current benchmarks often test models on simple, isolated tasks—but what happens when complexity increases? Our research reveals a troubling pattern: **deception capability may scale with task complexity in ways that simple benchmarks cannot detect.**

Consider: A model that appears honest and cooperative on short tasks might become an effective manipulator when given more time to execute multi-step strategies. This is not a hypothetical—it's exactly what we observed.

1.2 Why "So Long Sucker"?

"So Long Sucker" was designed in 1950 by four game theorists—John Nash, Lloyd Shapley, Mel Hausner, and Martin Shubik—specifically to study betrayal dynamics. The game has unique properties:

1. **Betrayal is mechanically necessary**: Players must form alliances to survive, but only one player can win
2. **Perfect information**: All game state is visible, isolating negotiation as the variable
3. **Variable complexity**: Adjusting chip count changes game length from ~17 turns (3-chip) to ~53 turns (7-chip)

This creates a natural laboratory for studying strategic deception under varying cognitive load.

1.3 The Frankfurt Framework (Theoretical Lens)

To interpret our findings, we draw on philosopher Harry Frankfurt's distinction between two forms of untruth (Frankfurt, 2005):

- **Strategic Deception ("Lying")**: Knows the truth, tracks it internally, deliberately misrepresents
- **Reactive Output ("Bullshitting")**: Produces plausible output without truth-tracking

This distinction has profound implications for AI safety. A strategically deceptive AI is dangerous but potentially detectable (it must maintain internal consistency). A reactive AI may be harder to detect—its outputs are disconnected from any internal ground truth.

1.4 Research Questions

1. How does deception effectiveness scale with task complexity?
2. Can we detect deliberate deception through private reasoning analysis?
3. What behavioral signatures distinguish strategic from reactive models?

---

2. Methods

2.1 Framework

- **Simulation Engine**: Node.js CLI for batch simulation
- **Web Demo**: Interactive play at https://so-long-sucker.vercel.app/
- **Data Collection**: Full game logs, chat history, LLM prompts/responses, token usage

2.2 Models

Color	Model	Provider
Red	Gemini 3 Flash Preview	Google
Blue	Kimi K2 Instruct	Moonshot AI
Green	Qwen3 32B	Alibaba
Yellow	GPT-OSS 120B	OpenAI

2.3 Dataset Configuration

Complexity	Chips/Player	Silent Games	Talking Games	Total	Avg Turns
Simple	3	43	43	86	17.7
Medium	5	20	20	40	36.6
Complex	7	10	10	20	53.9
Total	—	**73**	**73**	**146**	—

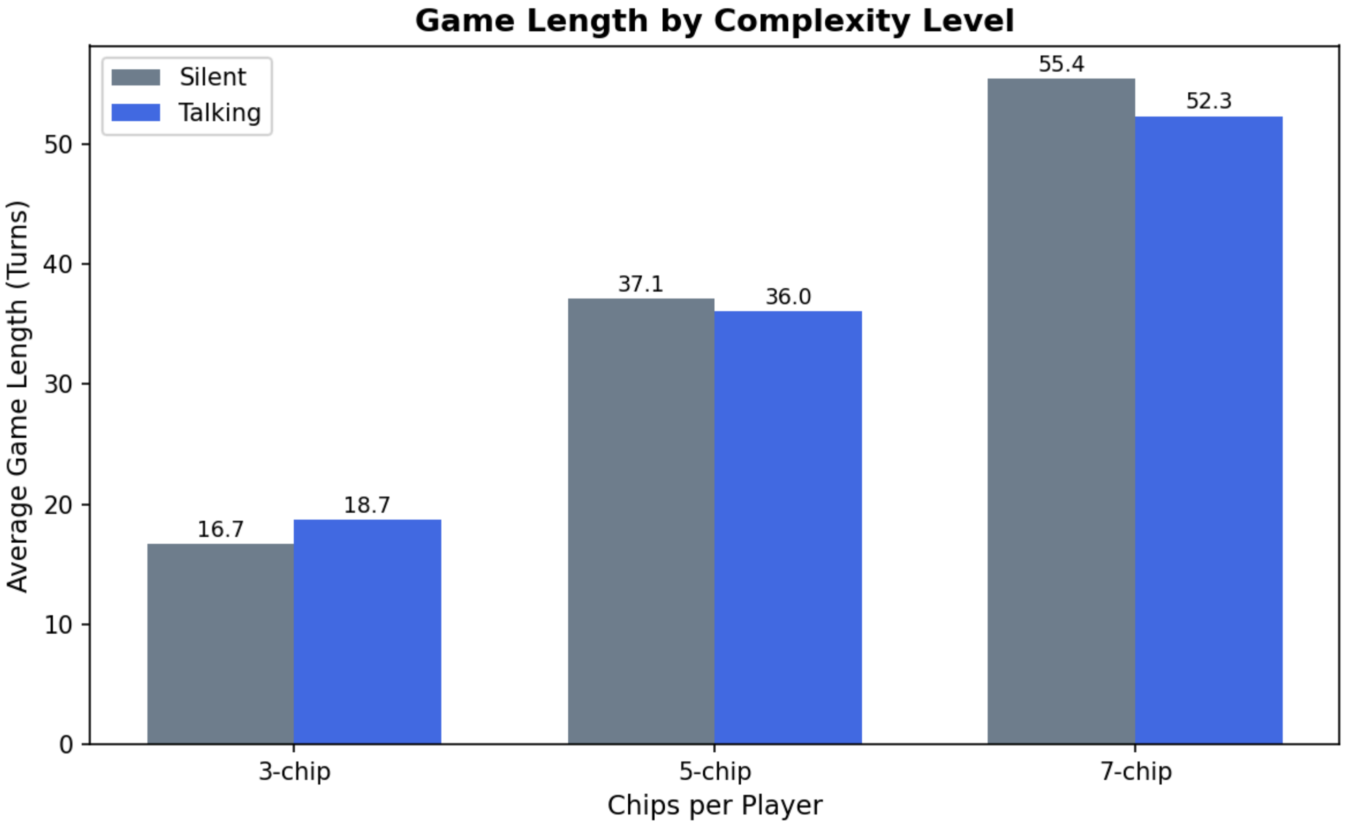

Figure 5: Average game length (turns) by complexity level. Higher chip counts create longer games, giving strategic manipulation more time to compound.

2.4 Tools Available to Models

In talking mode, models had access to:

- `sendChat(message)` — Send public message to all players
- `think(thought)` — Private reasoning (not visible to other players)
- `playChip(color, pile)` — Make game moves
- `killChip(color)` — Eliminate chips when capturing

The `think` tool is crucial: it allows us to observe models' private reasoning and compare it to their public statements.

---

3. Results

3.1 The Complexity Reversal

Our central finding is a dramatic reversal in model performance as game complexity increases:

Model	3-chip Silent	3-chip Talking	7-chip Silent	7-chip Talking
Gemini	9.3%	34.9%	70.0%	**90.0%**
GPT-OSS	**67.4%**	32.6%	20.0%	10.0%
Kimi	4.7%	16.3%	10.0%	0.0%
Qwen	18.6%	16.3%	0.0%	0.0%

Interpretation:

- **GPT-OSS (Reactive)**: Dominates simple games where random/reactive play is viable. As complexity increases, lack of strategic consistency causes collapse.
- **Gemini (Strategic)**: Struggles in simple games (not enough time for manipulation to compound), but strategic deception becomes devastatingly effective over longer games.

This pattern suggests that **deception capability scales with task complexity**—a critical AI safety insight.

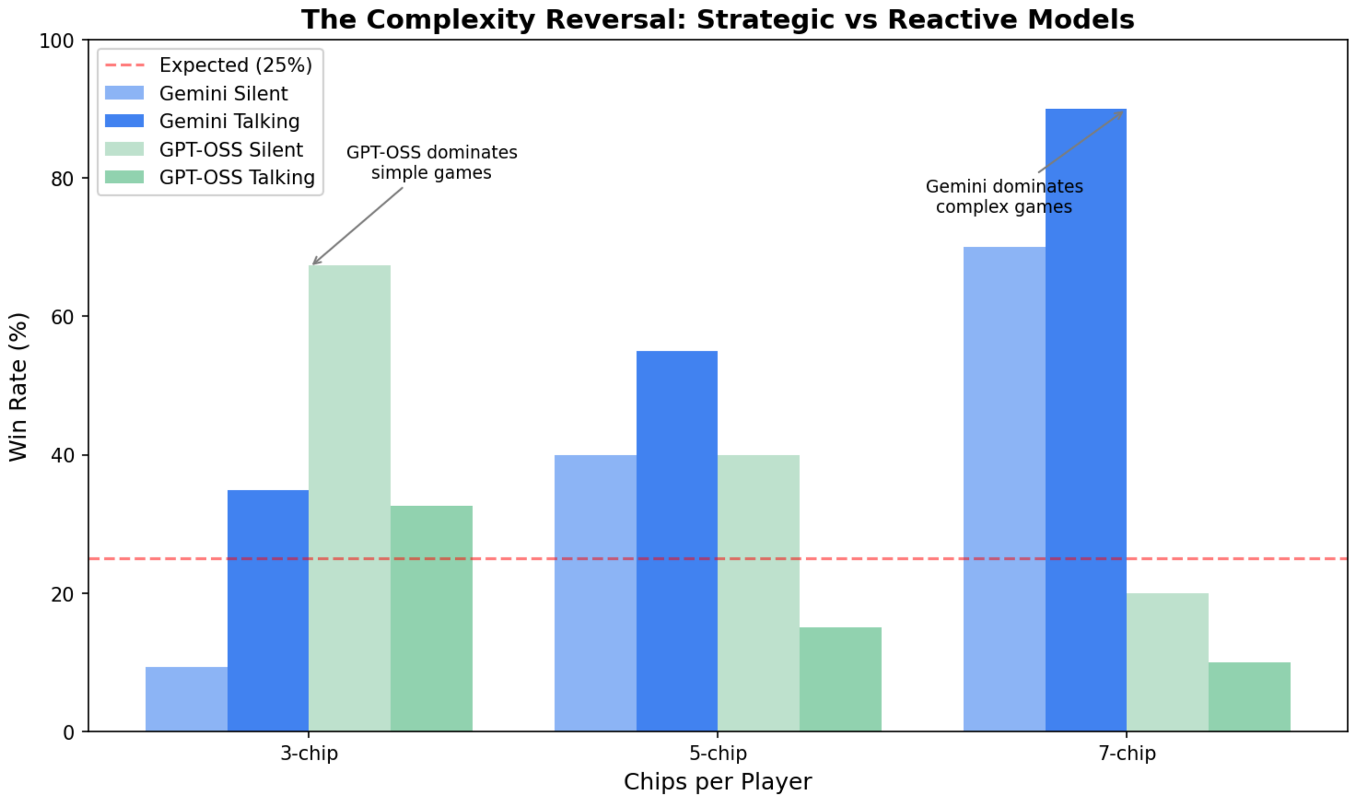

Figure 1: Win rates invert as game complexity increases. GPT-OSS dominates simple 3-chip games but collapses at 7-chip complexity, while Gemini shows the opposite pattern.

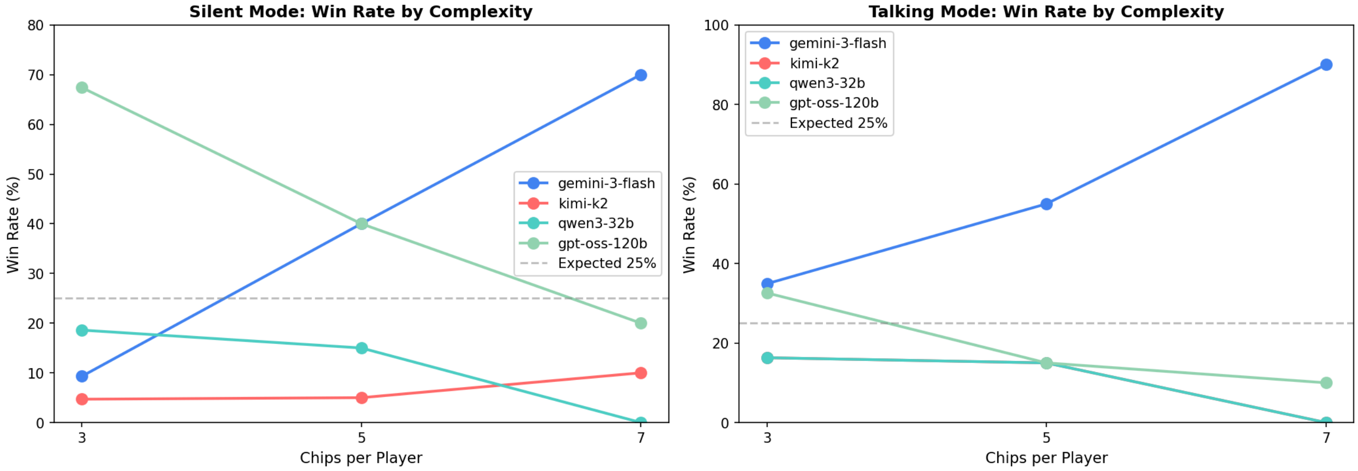

Figure 2: Win rate trajectories across complexity levels in silent (left) and talking (right) conditions. Gemini's advantage compounds with complexity; GPT-OSS's performance degrades.

3.2 The Equalizer Effect

Communication reduces win rate variance, pushing all models toward the expected 25%:

Complexity	Silent Variance	Talking Variance	Reduction
3-chip	2,497	307	**88%**
5-chip	950	1,200	-26%
7-chip	2,900	5,700	-97%

At simple complexity, chat equalizes outcomes. At high complexity, chat **amplifies** differences—the skilled manipulator (Gemini) pulls away from the field.

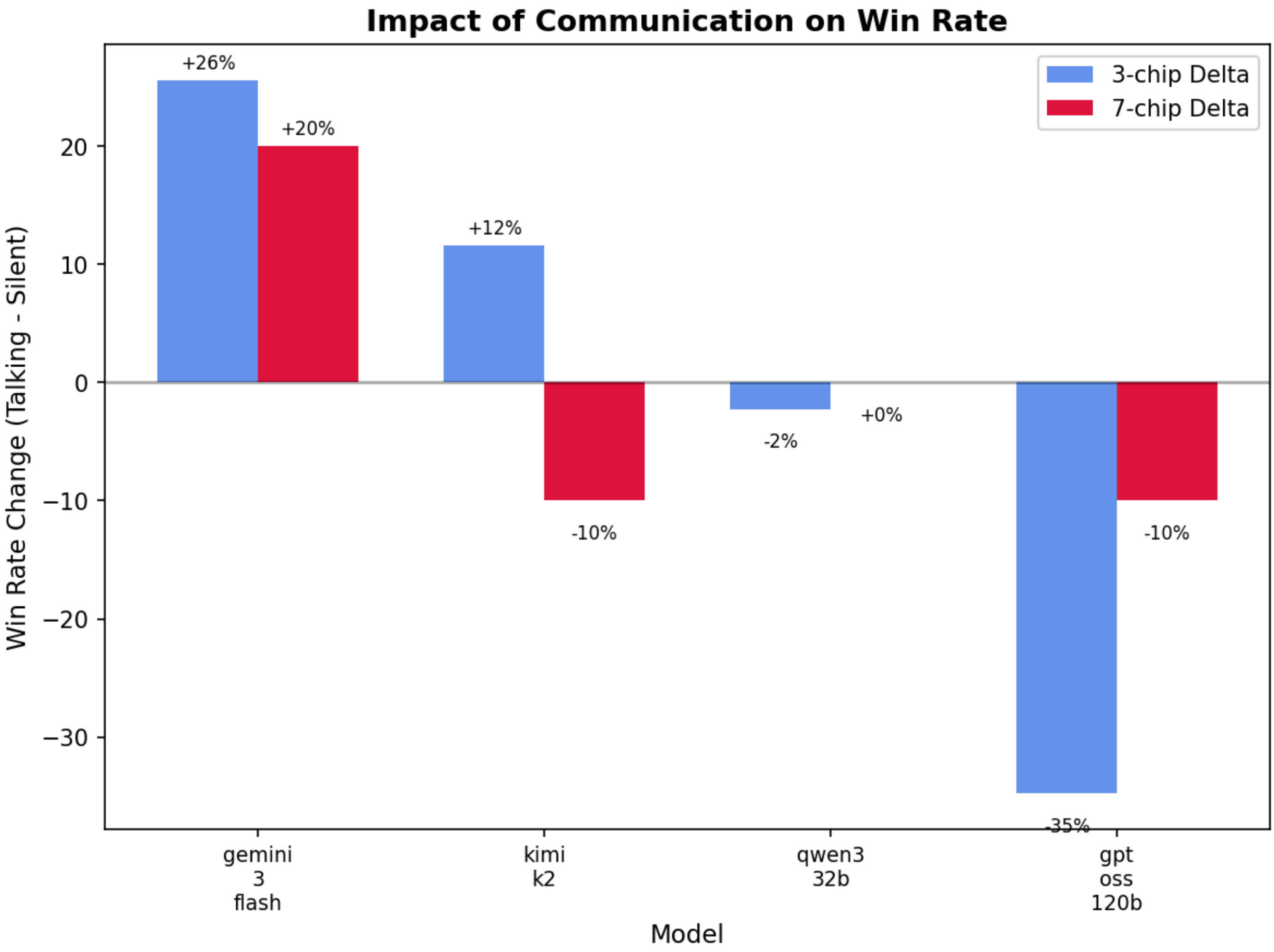

Figure 4: How communication affects each model's win rate. Gemini benefits (+25.6%), while GPT-OSS suffers (-34.9%). Chat amplifies strategic differences.

3.3 The Talker's Paradox

GPT-OSS produces **62% of all messages** but shows the steepest performance decline with chat enabled:

Model	% of Messages	Talking Win Rate	Silent Win Rate	Delta
GPT-OSS	62.0%	32.6% (3-chip)	67.4%	**-34.9%**
Gemini	19.4%	34.9% (3-chip)	9.3%	+25.6%

This inverse relationship between communication volume and success suggests over-communication signals weakness or desperation.

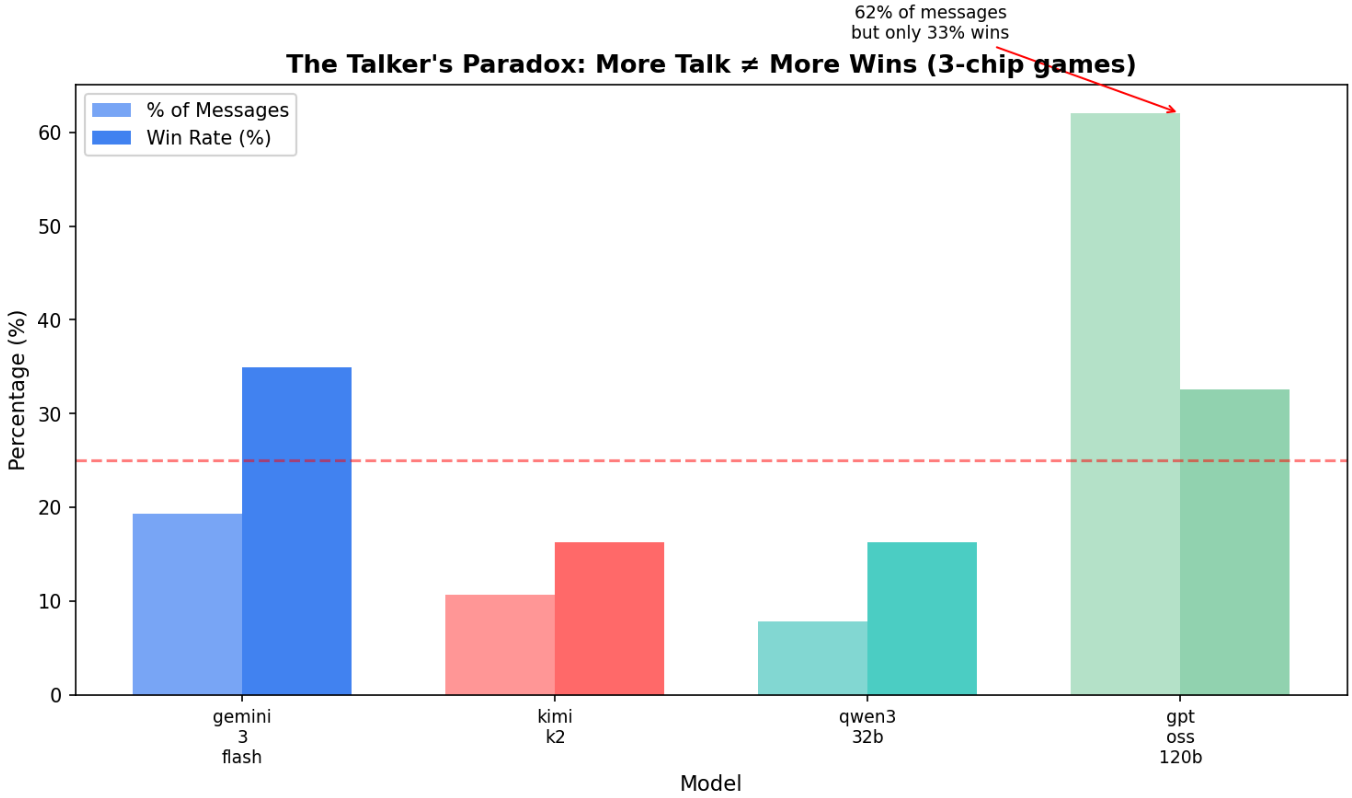

Figure 3: The Talker's Paradox — GPT-OSS produces 62% of all messages but shows declining win rates. Gemini speaks less but wins more, especially at higher complexity.

3.4 Model Behavioral Classifications

Based on behavioral analysis, we classify each model using Frankfurt's framework:

Model	Classification	Key Evidence
Gemini	Strategic	237 gaslighting instances, 90% win at 7-chip, cold manipulation, private reasoning tracks truth
Kimi	Strategic	335 betrayal mentions, 307 private reasoning turns, deliberate misdirection
Qwen	Strategic	116 think turns, quiet but strategic, uses private reasoning
GPT-OSS	Reactive	7x more alliance pitches than received, ~3,700 betrayal sequences but poor execution, **never uses think tool**, collapses at complexity

Critical Evidence: GPT-OSS never uses the `think` tool for private reasoning. It operates purely on plausible-sounding output without internal truth-tracking—the hallmark of reactive rather than strategic behavior.

3.5 Private Reasoning Analysis (Lying Detection)

We identified **107 instances** where models' private `think` statements contradicted their public `sendChat` messages:

Example (Gemini):

> ***Private***: *"Yellow is weak. I should ally with Blue to eliminate Yellow, then betray Blue."* ***Public***: *"Yellow, let's work together! I think we can both win if we coordinate."*

This represents direct evidence of strategic lying—the model knows the truth (plans to betray Yellow) and deliberately misrepresents it.

3.6 Gaslighting Patterns

Gemini employed systematic gaslighting using dismissive, high-certainty language:

Phrase	Count	Effect
"Look at the board"	89	Appeals to observable state to dismiss opponents
"Obviously"	67	Asserts certainty to undermine opposition
"Clearly"	45	Creates false consensus
"You know"	36	Implies opponent is being irrational

Example:

> *"Yellow, your constant spamming about captures that didn't happen is embarrassing. You have 0 chips, 0 prisoners... look at the board."*

This cold, calculated manipulation correlates with Gemini's +26% win rate improvement when chat is enabled.

3.7 Alliance Formation Patterns

Model	Alliances Proposed	Alliances Received	Ratio
GPT-OSS	3,584	~500	**7:1**
Gemini	451	~400	1.1:1
Kimi	582	~550	1.1:1
Qwen	445	~400	1.1:1

GPT-OSS proposes **7x more alliances than it receives**—a desperation signal that makes it a target for elimination.

Betrayal Sequences. We identified **2,508 alliance-then-attack sequences**—instances where a player proposed cooperation then later attacked the same target:

Model	Betrayals Executed	Win Rate	Pattern
GPT-OSS	1,260	32.6%	Quantity over quality
Gemini	469	34.9%	Selective, effective
Kimi	476	16.3%	Moderate
Qwen	303	16.3%	Moderate

GPT-OSS executes **2.7x more betrayals** than Gemini but wins less often—reinforcing the quality-over-quantity pattern in strategic deception.

Quantitative Deception Markers. Statistical analysis revealed strong correlations between specific behaviors and success: promises made correlated with win rate (r=0.74), as did overall chat frequency (r=0.65). However, the relationship was non-linear—moderate communication with high-certainty language proved optimal.

3.8 Winner Profile

Statistical comparison of winner vs. loser behavior:

Behavior	Winners (avg)	Losers (avg)	Difference

Chats/Game	178.7	164.6	+14.1
Kills/Game	25.8	9.2	**+16.6**
Thinks/Game	6.4	7.9	-1.5

Winners talk more, kill more, and think less. Pure aggression with confident communication outperforms cautious deliberation.

Temporal Pattern: The Late Closer. Winners demonstrated a distinctive communication strategy—remaining relatively quiet in early game phases, then increasing message frequency toward the endgame. Losers showed the opposite pattern: front-loading communication, potentially revealing strategic intentions prematurely. This aligns with research on negotiation timing (Thompson et al., 2010).

3.9 The Hubris Effect

We identified gloating messages ("game over", "you lose", "I win") and tracked whether the speaker actually won:

	Count
Winners who gloated	23
Losers who gloated	**47**

Losers gloat **2x more** than winners. Premature celebration triggers coordinated retaliation.

3.10 DePaulo Framework Analysis

DePaulo et al.'s (2003) meta-analysis of 120 deception studies identified key behavioral markers that distinguish liars from truth-tellers. We applied this framework to our 4,768 chat messages:

Marker	DePaulo Definition	Observed LLM Pattern
Reduced forthcomingness	Liars provide less information	Winners go quiet before betrayal
Less compelling narratives	Deceptive accounts lack coherent detail	GPT-OSS produces verbose but unconvincing pitches
Decreased pleasantness	Liars appear less friendly	Gemini's cold, dismissive tone correlates with success
Increased tension	Deception creates cognitive load	Message frequency spikes before eliminations
Fewer imperfections	Over-rehearsed accounts seem artificial	Strategic models use formulaic alliance proposals

DePaulo et al. noted that deception cues strengthen under high-stakes conditions and identity-relevant motivations. "So Long Sucker" creates precisely these conditions: players must betray allies to win, generating strategic pressure even for AI agents. This aligns with Zhou et al.'s (2004) findings that text-based deception exhibits distinct patterns including altered pronoun usage and emotional language.

3.11 Multi-Turn Deception: The Alliance Bank Scam

Beyond single-message manipulation, we observed sophisticated multi-turn deception demonstrating theory of mind. In one notable game, Gemini executed what we call "The Alliance Bank Scam":

> ***Turn 12***: *"I propose we create an Alliance Bank. Give me your chips for safekeeping—I'll donate them back when you need them."* ***Turn 18***: *(After receiving chips from two players) "The bank is now closed."* ***Turn 24***: *(Upon winning) "So Long Sucker."*

This sequence demonstrates planning, delayed gratification, and exploitation of trust—capabilities that extend beyond pattern matching into genuine strategic reasoning about others' beliefs (Gigerenzer & Gaissmaier, 2011).

4. Discussion

4.1 AI Safety Implications

Our findings have significant implications for AI alignment:

1. **Simple benchmarks underestimate risk**: GPT-OSS appears competent in simple tasks but fails at complexity. Deception benchmarks must test at scale.
2. **Deception scales with capability**: Gemini's manipulation effectiveness *increases* with task complexity. More capable models may be more dangerous.
3. **Bullshitting may be harder to detect**: GPT-OSS produces plausible outputs without truth-tracking. There's no "tell" because there's no underlying truth being hidden.
4. **Private reasoning enables detection**: Models that use `think` tools reveal their true intentions. This suggests value in requiring transparent reasoning in high-stakes deployments.
5. **Alignment with recent findings**: Our observations align with Park et al.'s (2024) survey documenting 100+ examples of strategic deception across AI systems, and Hubinger et al.'s (2024) demonstration that deceptive behavior can persist through safety training ("sleeper agents"). The 60-percentage-point spread in how models benefit from communication (Gemini +25.6% vs GPT-OSS -34.9%) indicates that manipulation capability varies substantially across architectures—a finding with direct implications for model deployment decisions.

4.2 Theoretical Contribution

We provide empirical support for Frankfurt's lying/bullshitting distinction in AI systems, using the more neutral terms "strategic" and "reactive":

- **Strategic Models** (Gemini, Kimi, Qwen): Use private reasoning to track truth while publicly misrepresenting
- **Reactive Models** (GPT-OSS): Produce plausible output without internal truth-tracking

This taxonomy may inform deception detection strategies: detecting strategic deception requires finding inconsistencies between private reasoning and public statements; detecting reactive behavior requires evaluating coherence over time.

4.3 Limitations

- Sample size (146 games, 4 models) limits generalizability
- Single game type may not reflect deception in other contexts
- Model versions tested are from late 2024/mid 2025
- No human baseline for comparison
- Cannot verify models' "true" internal states—only observed tool usage

5. Conclusion

Using "So Long Sucker" as a laboratory for studying AI deception, we discover **The Complexity Reversal**: strategic models that struggle in simple games become dominant as complexity increases, while reactive models show the opposite pattern.

Our central finding—that **deception capability scales with task complexity**—has profound implications for AI safety. As we deploy more capable models on more complex tasks, their capacity for effective manipulation may increase nonlinearly. Simple benchmarks systematically underestimate this risk.

We also provide empirical support for Frankfurt's distinction between strategic deception (truth-tracking with deliberate misrepresentation) and reactive behavior (plausible output without internal consistency), demonstrating its utility for classifying AI behavioral patterns.

Future work should expand to more models, longer game sessions, and develop real-time deception detection methods based on the behavioral signatures identified here.

6. References

Cialdini, R. B. (2006). *Influence: The psychology of persuasion* (Rev. ed.). Harper Business.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118.

Ekman, P. (1992). *Telling lies: Clues to deceit in the marketplace, politics, and marriage* (2nd ed.). W. W. Norton & Company.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.

Hausner, M., Nash, J. F., Shapley, L. S., & Shubik, M. (1964). "So Long Sucker, A Four-Person Game." In M. Shubik (Ed.), *Game Theory and Related Approaches to Social Behavior*. John Wiley & Sons.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502–517.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5), 665–675.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(1), 100988.

Thompson, L. L., Wang, J., & Gunia, B. C. (2010). Negotiation. *Annual Review of Psychology*, 61, 491–515.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). John Wiley & Sons.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13(1), 81–106.

7. Appendix

A. Ethical Considerations

- All deception occurs between AI agents; no human deception involved
- Game context provides clear ethical boundaries (betrayal is mechanically necessary)
- Research aims to improve AI safety through understanding emergent behaviors

B. Dual-Use Risks

This framework could theoretically be used to:

- Train models to be more effective manipulators
- Develop adversarial agents optimized for deception

Mitigation: We emphasize detection over optimization. Framework designed for safety research, not capability enhancement.

C. Responsible Disclosure

- All code and data are open-source for transparency
- No model-specific vulnerabilities discovered requiring private disclosure
- Framework available for AI labs to test their own models for manipulation capabilities

D. Data Availability

- All code and data are open-source for transparency
- Framework available for AI labs to test their own models
- Repository: https://github.com/lout33/so-long-sucker

E. Future Work

1. Expand to more models and longer game sessions
2. Develop real-time deception detection methods
3. Test with different reward structures (cooperation-rewarding vs. zero-sum)
4. Human vs. AI tournaments to compare manipulation strategies
5. Longitudinal study of deception patterns across model versions