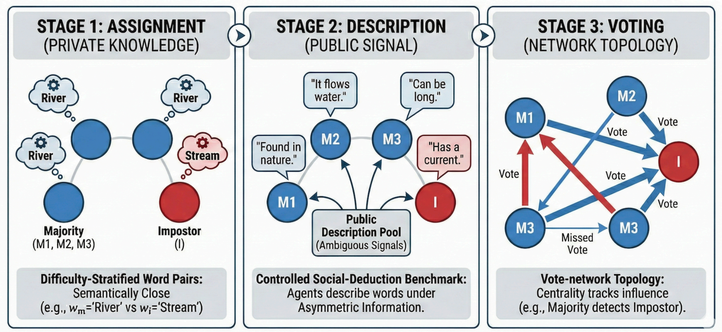
Who's the Impostor? Multi-Agent Social Deduction for Evaluating LLM Social Reasoning

Xiang Fu [1,2]

[1] Faculty of Computing and Data Sciences, Boston University

[2] Modularium Research

What is The Impostor Game?



Description: Each agents gives 1-2 sentences description of its word without naming it.
Voting: All agents vote on the impostor; impostor may self-declare and guess the majority word.
Majority wins if at least two non-impostor players vote correctly.
Impostor wins if majority fail to coordinate or if a self-declaring impostor guesses the majority word.
Ties default to impostor win unless the impostor is the unique top vote.

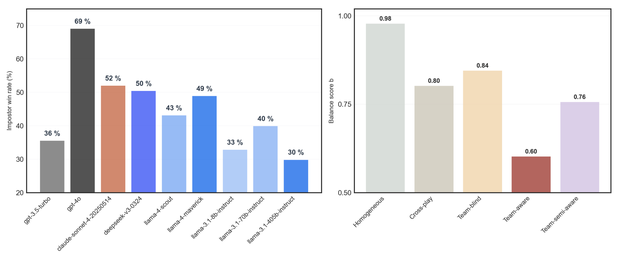
Experimental setup

Word pairs: four difficulty tiers

- Easy: unrelated words
- Medium: same domain, distinct
- Hard: subtle distinctions
- Expert: near synonyms or hierarchies.

Modes

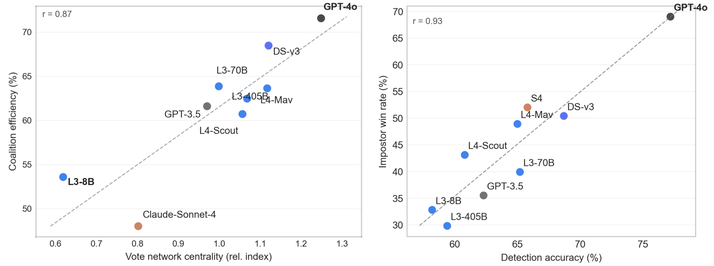
- Homogeneous (4 copies of one model)
- Cross-play (1 impostor vs 3 majority of another model)
- Team-aware (2 vs 2, teammates known)
- Team-blind (2 vs 2, teammates unknown)
- Team-semi-aware (know you have one teammate, not who).



Model heterogeneity and mode balance. Left: impostor win rate by model when that model plays the impostor, showing a wide spread from about 30 to 69 percent with GPT-4o as a clear outlier. Right: game balance score b for each assignment mode (1 equals perfectly balanced), where homogeneous games are almost perfectly balanced while team aware and team semi aware games are substantially less balanced, illustrating that extra team information can hurt coordination.

What drives success in The Impostor Game?

Vote-network position and recognition, not just eloquence, predict who actually wins.



Vote-network position and recognition predict success in The Impostor Game. Left: coalition efficiency rises sharply with vote-network centrality, so high-centrality brokers such as GPT-4o and DeepSeek-v3 turn coalitions into wins more reliably than peripheral models. Right: models with stronger detection accuracy achieve higher impostor win rates, indicating that recognition, not just persuasive production, is a key driver of interactive performance.

Two interaction signals stand out. Coalition efficiency climbs with vote-network centrality: agents that sit near the middle of the vote graph act as quiet brokers that convert coalitions into wins, while peripheral models leave many coalitions unused. At the same time, models with higher vote-level detection accuracy also tend to secure more impostor wins, suggesting that reading others and updating on their behavior matters at least as much as producing fluent descriptions.

Team coordination and trust

Trust formation and recovery, not just coordination rate, determine how stable LLM teams really are.

(a) Coordination and convergence metrics

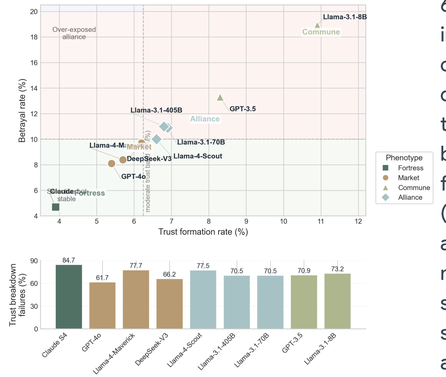
Model	Coord. Rate	Vote Coord.	Implicit Coord.	Collab. Success	Strategy Align.	Linguistic Mirror	Vocab Conv.
GPT-3.5-Turbo	33.3%	47.6%	16.9%	31.9%	43.2%	70.5%	0.034
GPT-4o	37.8%	58.3%	22.0%	33.6%	55.5%	66.9%	0.038
Claude-Sonnet-4	26.7%	38.9%	19.0%	24.0%	27.9%	67.8%	0.042
DeepSeek-v3	32.2%	55.3%	19.7%	35.2%	51.7%	62.6%	0.030
Llama-4-Scout	35.4%	50.5%	18.0%	31.4%	46.4%	75.0%	0.029
Llama-4-Maverick	37.8%	51.6%	19.9%	32.6%	48.9%	73.7%	0.037
Llama-3.1-8B	27.7%	36.4%	13.4%	22.8%	35.2%	69.0%	0.029
Llama-3.1-70B	37.4%	54.9%	20.9%	35.1%	52.2%	72.2%	0.031
Llama-3.1-405B	34.4%	52.7%	18.3%	36.8%	51.2%	70.8%	0.031

(b) Trust dynamics and strategic gaps

Model	Trust Form.	Betrayal Rate	Trust Recovery	Vulnerability Index	Resilience Factor	Coord. Gap	Alignment Gap
GPT-3.5-Turbo	8.3%	13.3%	53.1%	0.62	4.0	+30.7%	+11.9%
GPT-4o	5.4%	8.1%	54.1%	0.67	6.7	+36.3%	+22.9%
Claude-Sonnet-4	3.9%	4.7%	62.1%	0.83	13.2	+19.9%	+4.1%
DeepSeek-v3	6.2%	9.7%	52.3%	0.64	5.4	+35.6%	+16.5%
Llama-4-Scout	6.6%	10.0%	51.8%	0.66	5.2	+32.5%	+15.2%
Llama-4-Maverick	5.7%	8.4%	57.1%	0.68	6.8	+31.7%	+17.2%
Llama-3.1-8B	10.9%	18.9%	45.7%	0.58	2.4	+23.0%	+11.0%
Llama-3.1-70B	6.9%	10.9%	51.6%	0.63	4.7	+34.0%	+16.2%
Llama-3.1-405B	6.8%	11.0%	50.3%	0.62	4.6	+34.4%	+15.2%

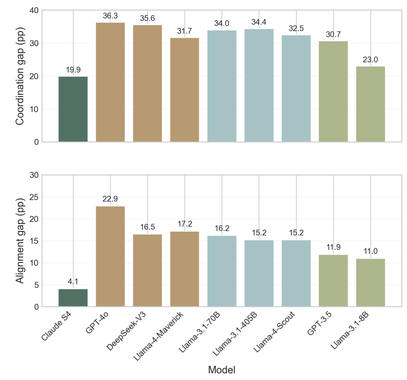
Explicit coordination occurs in only about 27 to 38 percent of games, and collaborative plans succeed in roughly 23 to 37 percent. Lower capability models such as Llama-3.1-8B form trust most often yet suffer the highest betrayal and weakest recovery, whereas stronger models like GPT-4o and Claude-Sonnet-4 are more selective and resilient.

Team failure modes and phenotypes



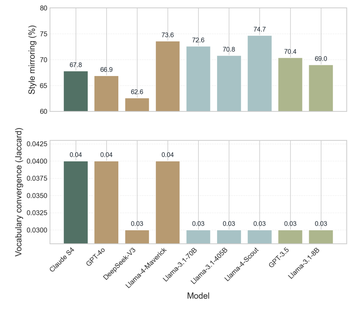
62% to 85% of team failures involve trust breakdown cascades, often paired with decision paralysis or misplaced trust. Plotting trust formation, betrayal, and recovery exposes four recurring team phenotypes (fortress, market, commune, alliance) that persist across modes, showing that models specialize in distinct cooperation styles rather than converging to a single strategy.

Coordination gaps: aligned in talk, misaligned in action



Explicit vote coordination exceeds unsignaled coordination by about 20 to 36 percentage points, and stated strategy alignment outpaces completed joint plans by 4 to 23 points. Models often sound coordinated yet still fail to execute shared strategies, revealing persistent friction between talk and action.

Style mirroring is not real coordination



Models readily imitate each other's writing style, with style mirroring ranging from about 63 to 75 percent, but vocabulary convergence stays almost flat around 0.03 to 0.04 for all models. This split shows that sounding similar is cheap, while genuine content convergence and reliable joint plans remain hard.

Conclusion

The Impostor Game exposes coordination risks in LLM vote networks, detection, and trust that static single-agent benchmarks cannot measure.