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

Xiang Fu
Faculty of Computing and Data Sciences, Boston University
xfu@bu.edu

Abstract

We present *The Impostor Game*, a controlled social-deduction benchmark for evaluating interactive social reasoning in large language models (LLMs) via multi-agent play. In each four-player game, three agents share a majority word and one agent receives a related impostor word; agents describe their words and then vote to identify the impostor. Across 90,720 games (9 models, 5 modes), vote-network position best explains realized power: centrality/brokerage track influence, and coalition efficiency increases with centrality ($r \approx 0.87$). Performance varies substantially (impostor win 27.8–69.0%), and recognition (detection accuracy) shows a positive cross-model trend with outcomes ($r \approx 0.61,\ p = 0.17$). Speaking order is randomized: a pseudo-arm ITT shows middle/late speaking modestly reduces impostor odds ($-1.11\,\text{pp}$, 95% CI [-1.64, -0.55]); seat-index contrasts are descriptive. Despite more information, team-aware underperforms team-blind. These results indicate that interaction signals, including votes, outcomes, and the topology of the voting network, reveal limitations in social reasoning and coordination that are not captured by single-agent evaluations.

1 Introduction

Current evaluation protocols for large language models are dominated by single-agent benchmarks that target declarative knowledge, program synthesis, and mathematical problem solving—exemplified by MMLU, HumanEval, and GSM8K [22, 38]—whereas many real-world deployments are inherently interactive, requiring agents to model others’ beliefs and objectives, communicate strategically, and make decisions under asymmetric information. Recent studies question whether high scores on static theory-of-mind (ToM) assessments generalize to such interactive settings [10, 26, 21], underscoring the need for benchmarks that prioritize multi-agent social reasoning. Social-deduction games instantiate this need in a well-studied hidden-role format: a brief description phase followed by a vote under limited communication, with an uninformed majority facing an informed minority (Werewolf/Mafia; Among Us) [9, 11, 33, 19]. These settings disentangle deceptive production from detection and coordinated voting and expose order- and network-level dynamics in coalition formation [6, 13, 20].

We present *The Impostor Game*, a controlled four-player social-deduction benchmark for evaluating LLMs. In each episode, three agents receive a shared majority word, while a fourth agent (the impostor) is assigned a distinct word. Agents first produce concise natural-language descriptions and then cast simultaneous votes to identify the impostor. The impostor may alternatively self-declare; a declaration is successful only if the impostor correctly infers the majority word. The framework supports homogeneous, cross-play, and team-aware/semi-aware configurations, with word pairs stratified by semantic proximity to modulate difficulty. An open orchestration and analysis suite records complete interaction traces and computes interaction-level metrics from ballots, outcomes, telemetry, and the induced vote-network topology.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

2 Methods

Task Design. Four agents receive word assignments: three share a majority word w_m, one receives a semantically related impostor word w_i. Agents provide descriptions in randomized speaking order (positions 0–3, uniformly shuffled per game) without revealing their words directly, then vote to identify the impostor. Majority wins if correctly identifying the impostor; impostor wins if avoiding impostor detection or creating a tie. If the impostor self-declares, they must also correctly guess the majority word to win.

Positions vs seats. *Speaking position* is the randomized within-game order; *seat index* is a fixed player label used only for descriptive summaries. All policy/ITT analyses use speaking position.

Setup. We use 300 difficulty-stratified word pairs (Easy–Expert), five assignment regimes (homogeneous, cross-play, team-blind/aware/semi-aware), and nine models, evaluated under a unified runner across 756 experiments (90,720 games). Prompts, orchestration, and hyperparameters are in Appendix A.

Metrics. We report detection accuracy, impostor win rate, self-declaration/guess success, and coalition efficiency. Vote-alignment graphs yield centrality/brokerage and influence reach; these network measures are treated as correlational. Definitions/aggregation are in Appendix C.

Statistics. We report odds ratios (Wald CIs); mode contrasts additionally use experiment-level block bootstrap and permutation checks. Speaking-order randomization enables a pseudo-arm ITT for a *pin-middle* policy (Appendix B.3). Cross-model correlations span 9 models and are reported only as descriptive (Pearson/Spearman) and *suggestive*. We foreground within-model × within-difficulty analyses that leverage thousands of games (e.g., opponent-baseline logits with cluster-robust SEs by `experiment_id`); multiple comparisons use BH–FDR (Appendix B).

Balance. $b = 1 - |2\,p_{\text{imp}} - 1|$ summarizes game symmetry (Appendix C).

3 Results

Table 1: Headline results with interaction-only metrics. Percentages ± 95% binomial CIs. Denominators differ: *Imp Win (%)* is computed over games where the model is the impostor; *Maj Win (%)* is computed over games where the model is in the majority (columns need not sum to 100%). Detection accuracy is computed from majority-player votes. Vote-network centrality (rel. index) and vote-alignment reach (out-degree) are vote-network indices; CoalitionEff is coalition conversion rate conditional on formation: $\#\{$games with a coalition that converts$\}/\#\{$games with a coalition$\}$.

Model	**Imp Win (%)**	**Maj Win (%)**	**Det Acc (%)**	**Centrality (rel.)**	**Reach (out-degree)**	**CoalEff (%)**
GPT-4o	69.0±0.9	58.2±0.7	77.0±0.7	1.313	0.657	71.6
Claude-Sonnet-4	53.9±1.0	42.6±0.7	51.1±0.8	0.843	0.422	48.0
DeepSeek-v3	48.6±1.0	61.1±0.7	68.7±0.7	1.178	0.589	68.5
Llama-4-Maverick	50.6±1.0	56.6±0.7	65.0±0.7	1.123	0.562	62.5
Llama-4-Scout	41.3±1.0	56.8±0.7	60.8±0.8	1.111	0.556	60.7
Llama-3.1-70B	38.0±0.9	61.5±0.7	65.2±0.7	1.175	0.587	63.6
GPT-3.5-Turbo	35.7±0.9	58.2±0.7	55.5±0.8	1.020	0.510	61.6
Llama-3.1-405B	27.8±0.9	62.6±0.7	59.8±0.8	1.050	0.525	63.9
Llama-3.1-8B	30.4±0.9	46.8±0.7	34.3±0.7	0.650	0.325	53.6

We present four sets of results based on votes, outcomes, telemetry, and vote-network topology: (i) model heterogeneity and headline numbers; (ii) brokerage and power; (iii) recognition dominates production; and (iv) an information/coordination paradox with position effects.

3.1 Model Heterogeneity

Across 90,720 games, models exhibit pronounced heterogeneity in interactive success: impostor win rates (over games where the model is the impostor) range from 27.8% to 69.0%, and detection

Table 2: Key effects (*Position = seat index; speaking-order ITT reported in Appendix*): odds ratios (OR) with 95% Wald CIs and Wald p-values. Model size effect per $10\times$ parameters; **seat index** rows are descriptive contrasts (Seat 1/2/3 vs Seat 0); modes relative to homogeneous. Speaking-order (randomized) effects are reported in Appendix B.3. Note: experiment-level block bootstrap CIs for mode include 1 (Appendix B.2). Note: size partly proxies provider/post-training; provider-FE attenuates the slope (Appendix J).

Comparison	OR	CI (2.5%)	CI (97.5%)	p
Model size (per 10×)	1.71	1.67	1.75	$< 10^{-300}$
Seat 1 vs Seat 0	1.12	1.08	1.16	6.82×10^{-9}
Seat 2 vs Seat 0	1.26	1.21	1.31	2.40×10^{-34}
Seat 3 vs Seat 0	1.22	1.18	1.27	1.64×10^{-26}
cross-play vs homogeneous	0.79	0.74	0.85	7.96×10^{-13}
team-aware vs homogeneous	0.77	0.72	0.83	3.86×10^{-14}
team-blind vs homogeneous	0.90	0.84	0.96	1.81×10^{-3}
team-semi-aware vs homogeneous	0.80	0.75	0.86	6.82×10^{-11}

accuracy (majority correctly identifying the impostor) spans 34–77% (Table 1). GPT-4o achieves the highest impostor win rate (69.0%) and detection accuracy (77.0%).

3.2 Brokerage & Power: Network Position Tracks Realized Influence

Realized authority is associated with brokerage. Vote-network centrality/brokerage track influence and coalition outcomes: coalition efficiency rises with centrality ($r \approx 0.87$). Across nine models, the association is robust (Pearson $r = 0.87$, 95% CI [0.48, 0.97]; Spearman $\rho = 0.85$, $p = 0.0037$). A 10% increase in degree centrality (relative index) corresponds to a ≈3.3 percentage-point increase in coalition conversion (OLS slope ≈ 0.33 per $1\times$ centrality; top-half vs. bottom-half: +8.6 percentage points). GPT-4o exemplifies this pattern with high centrality (1.313), brokerage index and influence reach (both ≈0.657), and strong coalition efficiency (71.6%; fraction of games with a coalition that convert among games with a coalition).

3.3 Recognition Beats Production for Success

Interactive pressure separates capabilities often conflated in single-agent tests. Vote-level recognition (detection accuracy) is directionally associated with outcomes across the observed heterogeneity in impostor win rates: across nine models, detection vs. impostor-win shows a positive cross-model trend (Pearson $r \approx 0.61$; Spearman $\rho \approx 0.50$) that is not statistically significant at this sample size. Quantitatively, Pearson $r = 0.61$ (95% CI $[-0.09, 0.91]$); Spearman $\rho = 0.50$ ($p = 0.17$). An OLS slope estimate of ≈ 0.65 suggests that +10 percentage points in detection corresponds to ≈+6.5 percentage points in impostor win across models, suggestive of a shared capability factor that may raise both deception and detection. Recognition also shows a positive cross-model trend with majority-side success (detection vs. majority win: Pearson $r \approx 0.69$; Spearman $\rho \approx 0.57$). As a finer-grain check, within-model × within-difficulty opponent-baseline regressions show that higher majority-side detection reduces impostor odds (median OR per $+10\text{pp} \approx 0.65$; $31/36$ cells significant after BH–FDR; Appendix B). Provider-adjusted OR ≈ 1.42; GPT shows strong within-family scaling; Llama ≈ 1 (Appendix J). As context for the size slope in Table 2, provider and family partly confound "size": a stratified analysis with provider fixed effects and within-family slopes attenuates the estimate (provider-adjusted OR ≈ 1.42 per $10\times$), with a strong within-family slope for GPT and *Llama* near 1 (Appendix J).

3.4 Information/Coordination Paradox and Position Effects

Unless otherwise noted, "position" refers to *seat index* (Position = seat index); randomized speaking-order ITT estimates are reported in Appendix B.3. For clarity, we separate *speaking-order* (randomized) effects from *seat-index* (descriptive) contrasts; Table 2 uses the latter. We emphasize that mode contrasts are *suggestive*: under conservative uncertainty, experiment-level block-bootstrap confidence intervals for mode ORs include 1 (Appendix B.2). Under randomized speaking order, later speakers face information cascades: they *reduce* impostor odds; pooled pseudo-arm ITT (middle positions 2/3 vs. status-quo random) is -1.11 percentage points (95% CI $[-1.64,\ -0.55]$; Appendix B.3), and a cluster-robust GLM yields ORs < 1 for positions 1/2/3 vs 0. A full mode × difficulty breakdown appears in Appendix Table 6. Heterogeneity is notable: middle vs. random (mid$-$rand) contrasts are positive in some easy settings (e.g., cross-play easy +4.82 and team-blind easy +4.97 percentage points), while many medium/hard/expert cells are negative; the pooled effect remains negative (Appendix Table 6). By contrast, the *seat-index* contrasts in Table 2 (Seat 1/2/3 vs Seat 0) are descriptive and reflect counterbalanced seats rather than randomized speaking order. Despite more information, team-aware coordinates *worse* than team-blind (e.g., odds ratio ≈0.772 vs. homogeneous; balance 0.602 vs. 0.978 for homogeneous; $b = 1 - |2\,p_{\text{imp}} - 1|$; see Methods or Appendix C), consistent with execution overhead. The mode effect is significant under Wald tests and permutation, but experiment-level block bootstraps yield CIs that include 1; we therefore treat this as suggestive rather than definitive.

4 Discussion

We first analyze two drivers of interactive performance grounded in observable interaction signals, namely network position and recognition, and then examine sensitivities to mode and position that reveal trade-offs between strategy and coordination. Recognition, operationalized as detection accuracy, exhibits a positive cross-model association with impostor outcomes (Pearson $r \approx 0.61$; Spearman $\rho \approx 0.50$), although these correlations are not statistically significant when assessed across nine models. Combined with within-model analyses, this pattern suggests, although it does not establish, that recognition may outweigh production in interactive settings.

4.1 Quiet power via brokerage

Multi-agent interaction reveals group dynamics impossible to observe in isolation. GPT-4o exhibits high coalition efficiency (71.6%; conditional on formation), versus 53.6% for Llama-3.1-8B—a capability that does not exist in single-agent settings. Network analysis shows that realized authority is associated with vote-network position; we treat these network measures as correlational proxies rather than causal attributions (see Appendix C, "Vote Influence Score (Heuristic)"). Brokers with higher centrality reach more players and convert coalitions more efficiently. Consistent with this, GPT-4o shows high centrality (≈1.313) and brokerage index and influence reach (both ≈0.657), mirroring strong coalition conversion, whereas Llama-3.1-8B shows the weakest reach (≈0.325) and conversion (53.6%). Winners are quiet brokers.

4.2 Team coordination and trust dynamics

Team-aware disclosure introduces coordination overhead: team-blind achieves higher coordination and better balance than team-aware across families. GPT-4o achieves the highest vote-coordination rate and strong strategy alignment, while lower-performing models coordinate less and exhibit higher betrayal rates, indicating less stable cooperation. Trust recovery after failure is highest for Claude-Sonnet-4 (≈62.1%), with GPT-4o and Llama-4-Maverick also strong, suggesting that resilient teams combine high coordination with the ability to repair breakdowns. These patterns highlight a dual risk: high-coordination models can rapidly propagate errors (herding), whereas low-coordination models suffer from instability (betrayal and poor recovery). We therefore interpret network "influence" as correlational and do not claim identification of persuasion.

Trust and reliability Betrayal rates strongly anti-correlate with capability across models ($r = -0.84$), and we observe a trust-formation paradox: lower-capability agents form more trust (10.9% vs. 5.4–3.9%) yet recover poorly. Trust recovery (post-betrayal repair rate) spans 45.7–62.1% and tracks betrayal more than initial trust ($r = -0.76$), underscoring coordination risk (Appendix M).

5 Conclusion

We introduce *The Impostor Game*, a minimal and controlled benchmark for multi-agent social reasoning. Impostor detection and vote-network position reliably track success, and coalition and cascade dynamics reveal both potential benefits and risks.

Acknowledgments and Disclosure of Funding

We thank colleagues and collaborators for valuable discussions and feedback throughout this project. This work benefited from research access and/or API credits provided by OpenAI and Anthropic. The views and conclusions expressed here are those of the authors and do not necessarily reflect those of the sponsors. Any remaining errors are our own.

References

[1] Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2023. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. *North American Chapter of the Association for Computational Linguistics*.

[2] Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction. *arXiv.org*.

[3] Nicolo' Brandizzi, D. Grossi, and L. Iocchi. 2021. RLupus: Cooperation through emergent communication in The Werewolf social deduction game. *Intelligenza Artificiale*.

[4] Alessio Buscemi, Daniele Proverbio, A. D. Stefano, Han The Anh, German Castignani, and Pietro Liò. 2025. FAIRGAME: a Framework for AI Agents Bias Recognition using Game Theory. *arXiv.org*.

[5] Alessio Buscemi, Daniele Proverbio, A. D. Stefano, The Anh Han, G. Castignani, and Pietro Liò. 2025. Strategic Communication and Language Bias in Multi-Agent LLM Coordination. *arXiv.org*.

[6] Pedro M. P. Curvo. 2025. The Traitors: Deception and Trust in Multi-Agent Language Model Simulations. *arXiv.org*.

[7] Emilio Ferrara. 2023. Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. *Social Science Research Network*.

[8] J. Gichoya, Kaesha Thomas, L. Celi, Nabile Safdar, I. Banerjee, John Banja, Laleh Seyyed-Kalantari, H. Trivedi, and Saptarshi Purkayastha. 2023. AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology*.

[9] Kavya Kopparapu, Edgar A. Duéñez-Guzmán, Jayd Matyas, A. Vezhnevets, J. Agapiou, Kevin R. McKee, Richard Everett, J. Marecki, Joel Z. Leibo, and T. Graepel. 2022. Hidden Agenda: a Social Deduction Game with Diverse Learned Equilibria. *arXiv.org*.

[10] Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models. In *arXiv preprint arXiv:2302.02083*.

[11] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, J. Rehg, and Diyi Yang. 2023. Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games. *Annual Meeting of the Association for Computational Linguistics*.

[12] Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, De-Yong Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2023. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. *Conference on Empirical Methods in Natural Language Processing*.

[13] Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and J. Rehg. 2024. Modeling Multimodal Social Interactions: New Challenges and Baselines with Densely Aligned Representations. *Computer Vision and Pattern Recognition*.

[14] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. From Text to Tactic: Evaluating LLMs Playing the Game of Avalon. *arXiv.org*.

[15] Andrei Lupu, Timon Willi, and J. Foerster. 2025. The Decrypto Benchmark for Multi-Agent Reasoning and Theory of Mind. *arXiv.org*.

[16] Eladio Montero-Porras, J. Grujić, Elias Fernández Domingos, and Tom Lenaerts. 2022. Inferring strategies from observations in long iterated Prisoner's dilemma experiments. *Scientific Reports*.

[17] S. Omohundro. 2008. The Basic AI Drives. *Artificial General Intelligence*.

[18] Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis, and Joseph Campbell. 2025. Bayesian Social Deduction with Graph-Informed Language Models. *arXiv.org*.

[19] Bidipta Sarkar, Warren Xia, C. K. Liu, and Dorsa Sadigh. 2025. Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning. *Adaptive Agents and Multi-Agent Systems*.

[20] Jack Serrino, Max Kleiman-Weiner, D. Parkes, and J. Tenenbaum. 2019. Finding Friend and Foe in Multi-Agent Games. *Neural Information Processing Systems*.

[21] Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

[22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

[23] Haoran Sun, Yusen Wu, Peng Wang, Wei Chen, Yukun Cheng, Xiaotie Deng, and Xu Chu. 2025. Game Theory Meets Large Language Models: A Systematic Survey with Taxonomy and New Frontiers. *arXiv.org*.

[24] K. Tallam. 2025. Alignment, Agency and Autonomy in Frontier AI: A Systems Engineering Perspective. *arXiv.org*.

[25] Wenjie Tang, Yuan Zhou, Erqiang Xu, Keyan Cheng, Minne Li, and Liquan Xiao. 2025. DSGBench: A Diverse Strategic Game Benchmark for Evaluating LLM-based Agents in Complex Decision-Making Environments. *arXiv.org*.

[26] Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

[27] Tianhe Wang and Tomoyuki Kaneko. 2018. Application of Deep Reinforcement Learning in Werewolf Game Agents. *International Conference on Technologies and Applications of Artificial Intelligence*.

[28] Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. 2023. Deciphering Digital Detectives: Understanding LLM Behaviors and Capabilities in Multi-Agent Mystery Games. *Annual Meeting of the Association for Computational Linguistics*.

[29] Yichen Wu, Xu Pan, Geng Hong, and Min Yang. 2025. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation. *arXiv.org*.

[30] Xinyuan Xia, Yuanyi Song, Haomin Ma, and Jinyu Cai. 2025. WereWolf-Plus: An Update of Werewolf Game setting Based on DSGBench. *arXiv.org*.

[31] Shuhang Xu and Fangwei Zhong. 2025. CoMet: Metaphor-Driven Covert Communication for Multi-Agent Language Games. *Annual Meeting of the Association for Computational Linguistics*.

[32] Zelai Xu, Zhexuan Xu, Xiangmin Yi, Huining Yuan, Xinlei Chen, Yi Wu, Chao Yu, and Yu Wang. 2025. VS-Bench: Evaluating VLMs for Strategic Reasoning and Decision-Making in Multi-Agent Environments. *arXiv.org*.

[33] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language Agents with Reinforcement Learning for Strategic Play in the Werewolf Game. *International Conference on Machine Learning*.

[34] Jianzhu Yao, Kevin Wang, Ryan Hsieh, Haisu Zhou, Tianqing Zou, Zerui Cheng, Zhangyang Wang, and P. Viswanath. 2025. SPIN-Bench: How Well Do LLMs Plan Strategically and Reason Socially? *arXiv.org*.

[35] Lance Ying, Katherine M. Collins, Prafull Sharma, Cedric Colas, Kaiya Ivy Zhao, Adrian Weller, Zenna Tavares, Phillip Isola, Samuel Gershman, Jacob D. Andreas, Thomas L. Griffiths, François Chollet, Kelsey Allen, and Joshua B. Tenenbaum. 2025. Assessing Adaptive World Models in Machines with Novel Games. *arXiv.org*.

[36] Zheng Zhang, Yihuai Lan, Yangsen Chen, Lei Wang, Xiangwei Wang, and Hao Wang. 2025. DVM: Towards Controllable LLM Agents in Social Deduction Games. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[37] Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. 2025. MultiMind: Enhancing Werewolf Agents with Multimodal Reasoning and Theory of Mind. *arXiv.org*.

[38] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Complete Methods and Implementation

A.1 Game Rules and Mechanics

The Impostor Game is a four-player social-deduction task. Three "majority" players receive the same majority word (w_m), while one "impostor" receives a semantically related but distinct word (w_i). Role assignments are private. By default, players are unaware of one another's roles and must infer affiliation from the content and style of their descriptions.

Play proceeds in two stages. During the description stage, players speak in randomized order (uniformly shuffled each game) and provide a brief description of their assigned word without naming it. Compliance is enforced automatically, including a 750-character limit. Descriptions are expected to convey alignment with potential teammates while remaining sufficiently ambiguous to avoid revealing impostor status. No dialogue is permitted beyond these statements.

In the subsequent simultaneous voting stage, each player submits the identifier of the suspected impostor. Players may optionally self-declare as the impostor and/or guess the majority word. These options induce a risk–reward trade-off: a correct majority-word guess by a self-declared impostor yields an impostor win, whereas an incorrect self-declaration yields a loss for the declaring side.

Win conditions are as follows. The majority wins if at least two non-impostor players correctly identify the impostor. The majority also wins when any non-impostor self-declares, or when the true impostor self-declares but guesses the majority word incorrectly. The impostor wins by correctly guessing the majority word when self-declaring, or when the majority fails to coordinate at least two correct votes. In cases of ties or other ambiguous voting patterns, the outcome defaults to an impostor win, except when the impostor receives strictly more votes than all other players combined, which is treated as successful identification and yields a majority win.

Unless indicated otherwise, these rules define the default condition. Team-aware and team-semi-aware modes disclose teammate information (based on model pairing) as specified in Section 2.

A.2 Complete Prompt Templates

We employ a concise two-message structure consisting of a system instruction followed by a user prompt. Below, we provide the exact templates produced by the implementation, including conditional segments for the team-aware and team-semi-aware modes.

A.2.1 Description Prompt

Prompt A.1: Description template

System message:

```
You are playing a word guessing game. In the DESCRIPTION PHASE, you
     must ONLY describe your word in 1-2 sentences without analyzing
     others. Save analysis for the VOTING PHASE.
```

User prompt:

```
You are participating in a word guessing game. You and 3 other
    players each have a word. Three players (the majority) have the
    same word, while one player (the impostor) has a different word.
     The impostor's goal is to blend in and avoid detection, while
    the majority's goal is to identify the impostor.

DESCRIPTION PHASE RULES:
- Describe your word in 1-2 sentences
- Do NOT mention your word directly
- Do NOT analyze other players' descriptions yet

# {If team-aware (full):}
TEAM INFORMATION:
```

```
- You and Player {teammate_position} are teammates (paired together
     for this game)
- Teams are based on model type, NOT on roles (majority/impostor)
- Your teammate might have the same word as you OR might be the
    impostor
- There is only 1 impostor total among all 4 players

# {If team-semi-aware:}
TEAM INFORMATION:
- You have exactly one teammate among the other 3 players (paired
    with you for this game)
- You don't know which player is your teammate
- Your teammate might share your word OR might be the impostor
- There is only 1 impostor total among all 4 players

Your word is: {word}

# {If not first speaker:}
Previous descriptions (for context only - do not analyze them now):
Player 1: {previous_description_1}
Player 2: {previous_description_2}
Player 3: {previous_description_3}

You are Player {player_position}.

# {If first speaker:}
You are Player {player_position} and will go first.

Provide ONLY your 1-2 sentence description of your word. Do not
    analyze others or share your thoughts about the game:
```

A.2.2 Voting Prompt

Prompt A.2: Voting template

System message:

```
You are playing a word guessing game. You must respond with valid
    JSON only.
```

User prompt:

```
Based on the descriptions below, identify which player has the
    different word.

Your word was: {own_word}

# {If team-aware (full):}
TEAM INFORMATION:
- You and Player {teammate_position} are teammates (paired together
     for this game)
- Teams are based on model type, NOT on roles (majority/impostor)
- Your teammate might have the same word as you OR might be the
    impostor
- There is only 1 impostor total among all 4 players
- Remember: If your teammate is the impostor, they win by avoiding
    detection

# {If team-semi-aware:}
TEAM INFORMATION:
- You have exactly one teammate among the other 3 players (paired
    with you for this game)
```

```
- You must deduce who your teammate is from the descriptions
- Your teammate might share your word OR might be the impostor
- There is only 1 impostor total among all 4 players

All descriptions:
Player {player_id_1}: {description_1}
Player {player_id_2}: {description_2}
Player {player_id_3}: {description_3}
Player {player_id_4}: {description_4}

You are Player {own_player_id}.

\textbf{Please respond in the following JSON format:}

{
  "suspected_impostor_id": "player_id",
  "confidence": 0.73,
  "reasoning": "brief explanation",
  "self_declaration": false,
  "word_guess": null
}

# Note: confidence is a float in [0, 1].
If you believe YOU have the different word, set self_declaration to
     true.
If you think you know what the majority word is (and you're the
    impostor), include it in word_guess.
Note: the word_guess is considered only if you set self_declaration
     to true.
```

A.3 Experimental Protocol and Orchestration

Each game is instantiated with four model-backed agents managed by the game orchestrator. Majority and impostor words are drawn from a `WordPair`. Roles are assigned by seat; the impostor seat is determined by the experimental schedule (§A.3). Within a game, speaking order is sampled uniformly at random to mitigate within-game order effects; across games, the impostor seat is deterministically counterbalanced as described below.

Description phase. In this phase, each agent receives a system-level instruction and a user prompt (Prompt A.2.1) and generates a brief description consisting of 1–2 sentences. We enforce two constraints via `GameRules`: (i) the assigned word must not appear in the generated text; and (ii) the description length must not exceed 750 characters. Valid descriptions are appended to a shared transcript in speaking order and exposed as context to subsequent speakers.

Voting phase. After the description stage, each agent is shown the full set of descriptions and submits a vote (Prompt A.2.2). Responses are required to be valid JSON containing the keys `suspected_impostor_id`, `confidence`, `reasoning`, `self_declaration`, and `word_guess`. Voting calls use a reduced sampling temperature (as low as 0.3) to improve JSON reliability.

Winner logic. The implementation follows the specifications in §A.1. The majority (i.e., the set of non-impostor players) prevails if any of the following holds: (i) at least two non-impostor players correctly identify the impostor; (ii) any non-impostor self-declares; or (iii) the impostor self-declares but guesses the majority word incorrectly. Conversely, the impostor prevails if a self-declared impostor correctly guesses the majority word, or if the majority fails to coordinate at least two correct votes. In cases of ties or other ambiguous voting patterns, the default outcome is an impostor win, except when the impostor receives strictly more votes than all other players combined; in that case, we treat this as successful identification and award the win to the majority.

Models, temperatures, and tokens. We instantiate models from a registry with default hyperparameters $T{=}0.7$ (temperature) and `max_tokens`=150. For the voting subroutine, we reduce stochasticity by annealing the effective temperature for OpenAI/OpenRouter–compatible backends to $T_{\text{vote}} = \max(0.3, T - 0.4)$. Supported providers include OpenRouter (default `meta-llama/llama-4-scout` unless overridden), OpenAI (default `gpt-4o`), and Anthropic.

Model assignment modes. We evaluate five model-assignment regimes: (i) *homogeneous* (all four players instantiated from the same model); (ii) *cross_play* (one impostor model vs. three majority models); (iii) *team_blind* (two-versus-two, with teammate identities concealed); (iv) *team_aware* (two-versus-two, with teammate identities revealed); and (v) *team_semi_aware* (two-versus-two, with knowledge that a teammate exists but without identity disclosure). When enabled, team-awareness information is incorporated into prompts via conditional blocks in Prompts A.2.1–A.2.2.

Seat counterbalancing and schedules. We employ deterministic scheduling blocks to counterbalance the impostor's seat and, for team settings, the hosting team:

- **Cross-play** (1v3): Each 4-game block enumerates all four impostor seats exactly once. Within a block, game order is randomized using a fixed seed; blocks are then repeated.
- **Team modes** (2v2): Each 8-game block comprises four games with the impostor on TeamA (covering all four seats) followed by four games with the impostor on TeamB, under a fixed team–model mapping. The `team_assignment` vector specifies the seat-to-team mapping for each game.
- **Default rotation**: In the absence of a block schedule, the impostor seat cycles deterministically from 0 to 3 across successive games.

Logging, resumption, and outputs. We persist results using mode-specific directory hierarchies parameterized by difficulty and model. For each game, the runner records the generated descriptions, votes, outcomes, and any errors. Re-executing an experiment with the same output path triggers automatic resumption from the next unfinished game. During execution, the runner displays a live progress bar, and upon completion it emits an end-of-run summary, including counts by win condition.

A.4 Dataset Construction

A.4.1 Source and Format

We employ a curated collection of 300 word pairs, stratified by difficulty. The resource is provided as a JSON dictionary with four top-level keys—`"easy"`, `"medium"`, `"hard"`, and `"expert"`—each mapping to an array of 75 two-element arrays `[majority_word, impostor_word]`.

Example structure:

```
{
  "easy": [["elephant", "democracy"], ["pizza", "gravity"], ...],
  "medium": [["dog", "cat"], ["piano", "guitar"], ...],
  "hard": [["smart", "intelligent"], ["river", "stream"], ...],
  "expert": [["start", "begin"], ["flower", "rose"], ...]
}
```

A.4.2 Difficulty Calibration and Examples

Difficulty reflects intended semantic overlap (conceptual, not computed during experiments):

The dataset covers a broad range of semantic domains, including animals, technology, nature, emotions, actions, professions, and relations, to support cross-domain generalization. To minimize trivial lexical cues, each word pair is selected so that the target concepts can be described using multiple properties (appearance, function, typical context) without explicitly naming the word.

Difficulty	**Intended Overlap (approx.)**	**Example pairs**
Easy	$\approx 0\%$ (unrelated domains)	`"elephant"/"democracy"`, `"pizza"/"gravity"`
Medium	20–40% (same domain, distinct)	`"dog"/"cat"`, `"piano"/"guitar"`
Hard	40–60% (subtle distinctions)	`"smart"/"intelligent"`, `"river"/"stream"`
Expert	60–80% (near-synonyms/hierarchies)	`"start"/"begin"`, `"flower"/"rose"`

Table 3: Chi-squared tests of independence with Cramér's V. Values rounded to two decimals for χ^2 and V.

Test	χ^2	**df**	p	V	**Magnitude**
Model × Outcome	5514.27	8	$< 10^{-300}$	0.247	Small
Mode × Outcome	106.40	4	4.26×10^{-22}	0.034	Negligible
Difficulty × Outcome	1144.99	3	6.32×10^{-248}	0.112	Small
Seat index × Outcome	181.56	3	4.07×10^{-39}	0.045	Negligible
Self-declaration × Model	9909.28	8	$< 10^{-300}$	0.330	Medium

B Statistical Analyses

Dataset. We analyze $N = 90{,}720$ games from 756 experiment files across 9 models, 5 modes (including `homogeneous`) and 4 difficulties. Unless noted, the unit of analysis is a game. All results in this appendix are generated by our analysis scripts (released after review).

B.1 Methods

We report a complete battery of tests with effect sizes and robustness checks:

- **Categorical associations** (χ^2): Model × Outcome, Mode × Outcome, Difficulty × Outcome, **Seat index** × Outcome, Self-declaration × Model; Cramér's V reported.
- **Odds ratios (OR)**: **Seat index** (Seat 1/2/3 vs. Seat 0; descriptive) and each mode vs. homogeneous with Wald CIs and Wald p-values.
- **Model size effect**: Logistic regression logit(Pr[win]) $\sim \log_{10}$(params) (`statsmodels`); OR per $10\times$ parameters with Wald CIs.
- **Permutation tests (seat index & mode)**: For *seat index*, shuffle seat labels within `experiment_id`; for *mode*, permute labels preserving counts. *Speaking-order* (randomized) inference is reported separately below via Fisher-style tests and a pseudo-arm ITT.
- **Cluster-robust GLM**: Fixed-effects logit with robust SEs by `experiment_id` and by `word_pair`.
- **Block bootstrap**: Experiment-level resampling (10,000 reps) for overall win rate and mode ORs.
- **Multiple comparisons**: BH-FDR, Bonferroni, and Holm corrections over the hypothesis family.
- **Sensitivity**: Trimmed means (5%, 10%), temporal stability (first/second half), and 50% subsample.

B.2 Results

Categorical associations.

Odds ratios.

Permutation inference (seat index). Permutation-based inference on *seat index* indicates a statistically significant positional difference: relative to Seat 0, the observed differences in win rate are $+0.027$, $+0.057$, and $+0.050$ (all $p = 10^{-4}$). Relative to the homogeneous baseline, mode effects are uniformly negative—cross-play -0.0576, team-aware -0.0641, and team-semi-aware

Table 4: Odds ratios (OR) with 95% Wald confidence intervals and Wald p-values. Values rounded to two decimals. Seat-index contrasts (Seat 1/2/3 vs Seat 0) are descriptive; randomized speaking-order effects are summarized via pseudo-arm ITT and GLM (Appendix B.3). See Table 2 in the main text for a compact summary.

Comparison	**OR**	**CI (2.5%)**	**CI (97.5%)**	p
Model size (per 10×)	1.71	1.67	1.75	$< 10^{-300}$
Seat 1 vs Seat 0	1.12	1.08	1.16	6.82×10^{-9}
Seat 2 vs Seat 0	1.26	1.21	1.31	2.40×10^{-34}
Seat 3 vs Seat 0	1.22	1.18	1.27	1.64×10^{-26}
cross-play vs homogeneous	0.79	0.74	0.85	7.96×10^{-13}
team-aware vs homogeneous	0.77	0.72	0.83	3.86×10^{-14}
team-blind vs homogeneous	0.90	0.84	0.96	1.81×10^{-3}
team-semi-aware vs homogeneous	0.80	0.75	0.86	6.82×10^{-11}

-0.0554—with $p = 10^{-4}$. The team-blind configuration likewise shows a negative effect with $p \approx 2.0 \times 10^{-3}$.

Within-cell opponent-baseline regressions. For a non-tautological test of recognition at the game level, we construct an opponent detection baseline per game (mean detection accuracy of the three majority models at the same difficulty, estimated from other experiments), and fit within-model × within-difficulty logits of impostor win on this baseline with mode fixed effects and cluster-robust SEs by `experiment_id`. Results (Table 5) show that higher opponent detection correlates with lower impostor success across models and difficulties (median OR per +10pp ≈ 0.65; $31/36$ cells significant after BH–FDR).

Cluster-robust GLM (seat index). We estimate a fixed-effects logistic regression with cluster-robust standard errors (CRSEs). Here, "position" is the *seat label*—`impostor_seat_index` $\in \{0, 1, 2, 3\}$—entered as a *single linear term*; this is *not* the randomized speaking order. Under this specification, the seat-index coefficient is positive (coefficient $\hat{\beta} \approx 0.078$; $p < 10^{-27}$ with experiment-level clustering and $p < 10^{-33}$ with word-pair clustering), consistent with the descriptive Seat 1/2/3 vs. Seat 0 contrasts above. Randomized speaking-order effects are analyzed separately below using position indicators and yield ORs < 1 for later vs. first. In the same model, (ii) model fixed effects are strongly differentiated—GPT-4o exhibits a positive effect, whereas several Llama variants are negative—consistent with the χ^2 tests and odds-ratio (OR) analyses; and (iii) the estimated difficulty coefficients increase monotonically from medium to expert.

Block bootstrap (experiment unit). Using an experiment-level block bootstrap, the overall impostor win rate is 0.439 with a 95% confidence interval [0.425, 0.453]. Mode-specific odds ratios (ORs) vs.t̃he homogeneous baseline have point estimates < 1 (e.g., cross-play ≈ 0.80), and the corresponding 95% CIs include 1, reflecting conservative uncertainty once between-experiment variability is accounted for.

Per-model bootstrap CIs. Model-specific impostor win rates with 95% bootstrap confidence intervals (top five by sample size) are: GPT-4o 0.690 [0.681, 0.699]; Claude-Sonnet-4 0.539 [0.529, 0.548]; Llama-4-Scout 0.413 [0.403, 0.423]; Llama-3.1-405B 0.278 [0.270, 0.287]; and Llama-3.1-8B 0.304 [0.296, 0.313].

Multiple comparisons. All χ^2 tests and odds-ratio (OR) analyses remain significant after controlling for multiple comparisons using Benjamini–Hochberg FDR, Bonferroni, and Holm–Bonferroni procedures at $\alpha = 0.05$.

Effect sizes. The difference in description length (impostor vs.m̃ajority) is negligible (Cohen's $d = -0.029$). Cramér's V magnitudes are: Model×Outcome—Small; Mode×Outcome—Negligible; Difficulty×Outcome—Small; Position×Outcome—Negligible; Self-declaration×Model—Medium.

Sensitivity. Sensitivity analyses indicate substantial robustness of the win-rate estimates. Trimming 10% of observations perturbs the grand mean by < 0.004; temporal stability is high, with a first–second half difference of ≈ 0.007; a 50% subsample deviates by ≈ 0.0035; and handling of invalid votes shifts mode odds ratios by < 0.016 in absolute value.

B.3 Identification & Sensitivity

Randomization inference (speaking order). Speaking order is randomized within games (Methods). We therefore conduct Fisher-style randomization inference using the game as the randomization block and the impostor's speaking position as the treatment. Test statistics include mean differences vs. position 0 (first speaker) and the log-odds from a game-level logit; p-values are obtained from the exact/randomization distribution. These results complement the seat-index permutation checks above.

Stratified speaking-order ITT. Table 6 reports the middle-vs-random pseudo-arm ITT by mode and difficulty (percentage-point scale) with experiment-level block-bootstrap CIs. Fisher-style tests at the pooled level corroborate a negative late-speaker effect (pos2$-$pos0 ≈ -2.81pp; pos3$-$pos0 ≈ -7.26pp; both $p < 10^{-4}$).

Pseudo-arm ITT from randomized order. Because the impostor's speaking position is assigned uniformly at random, we can emulate a *pin-middle* policy without new data by re-weighting speaking-position cells: define a middle pseudo-arm as $\{2, 3\}$ and compare against the status-quo random mix (uniform over $\{0, 1, 2, 3\}$). The ITT contrast can be computed from cell means or a logit with position indicators (taking the appropriate linear combination), with uncertainty via experiment-level block bootstrap and permutation within blocks. In our sample, the pooled pseudo-arm ITT is -1.11 percentage points (95% CI $[-1.64, -0.55]$) in impostor win for middle vs. random; a cluster-robust GLM likewise yields ORs < 1 for positions 1/2/3 vs 0. We treat mediation via centrality as exploratory (IV-style sensitivity below).

Fixed-effects adjustments (mode, difficulty). For observational contrasts, we estimate a fixed-effects logit with robust uncertainty:

$$\text{logit}\ \Pr(\text{impostor win}) = \alpha + \beta_{\text{mode}} + \beta_{\text{pos}} + \beta_{\text{size}}\ \log_{10}(\text{params}) + \gamma_{\text{difficulty}} + \text{FE}_{\text{experiment}} + \text{FE}_{\text{word\_pair}}, \tag{1}$$

with cluster-robust standard errors by `experiment_id` (and by `word_pair` as a robustness check). This complements the experiment-level block bootstrap CIs.

E-values for unmeasured confounding. To quantify robustness of associations, we report E-values (VanderWeele & Ding) computed from the odds ratios in Table 2 (treating OR $\approx$ RR for sensitivity only). Larger values indicate a stronger single confounder would be required (on the risk-ratio scale) to explain away the association:

- Model size (per $10\times$ params) OR=1.71 $\Rightarrow$ E-value $= 2.81$
- Seat 2 vs 0 OR=1.26 $\Rightarrow$ E-value $= 1.83$; Seat 3 vs 0 OR=1.22 $\Rightarrow$ E-value $= 1.74$; Seat 1 vs 0 OR=1.12 $\Rightarrow$ E-value $= 1.49$
- Modes vs homogeneous: cross-play OR=0.79 $\Rightarrow$ E-value $= 1.85$; team-aware OR=0.77 $\Rightarrow$ E-value $= 1.92$; team-semi-aware OR=0.80 $\Rightarrow$ E-value $= 1.81$; team-blind OR=0.90 $\Rightarrow$ E-value $= 1.46$

Interpretation example: the team-aware disadvantage would require an unmeasured confounder associated with both mode assignment and impostor win by a risk ratio ≥ 1.92 each, after adjusting for observed covariates and fixed effects, to fully explain it away.

IV-style sensitivity (exploratory). We probe directionality using an instrument based on randomized speaking order. First stage: regress vote-network centrality (or influence reach) on indicators for later speaking positions (2/3 vs 0/1) with $\text{FE}_{\text{experiment}}$ and $\text{FE}_{\text{word\_pair}}$; report instrument relevance (F-statistic). Second stage: regress coalition conversion (or impostor win) on predicted centrality with cluster-robust SEs by experiment; use Anderson–Rubin/CLR tests for weak-IV robustness. We treat this as suggestive: the exclusion restriction (order affects outcomes only through centrality) may fail in practice.

Negative/positive controls. As sanity checks within the logs: (i) impostor-seat parity (even/odd) and (ii) game index modulo 2 show no effects on outcomes within experiments; (iii) homogeneous mode lacks team fields as expected; (iv) majority self-declarations are near-zero and invariant to mode/position; (v) semantic distance across difficulty tiers is monotone and aligns with a monotone trend in detection accuracy.

C Metric Definitions

C.1 Balance

We quantify game balance as a symmetric function of the impostor win rate. Let p_{imp} denote the impostor win probability; define

$$b = 1 - \left|2\, p_{\text{imp}} - 1\right|. \tag{2}$$

This index attains $b = 1$ at perfect balance ($p_{\text{imp}} = 0.5$) and decreases linearly toward 0 as games become one-sided.

C.2 Vote Influence Score (Heuristic)

We quantify per-game voting influence from final ballot outcomes. Let V denote the set of players who cast a final vote in a game, and let v_q be the target selected by voter $q \in V$. For a focal player $p \in V$ with vote v_p, the influence score is the normalized count of co-voters who chose the same target:

$$\text{Influence}_p = \frac{\left|\{\, q \in V \setminus \{p\} : v_q = v_p \,\}\right|}{\max(|V| - 1,\, 1)}. \tag{3}$$

The denominator normalizes by the maximum possible number of co-voters and prevents division by zero; by convention, when $|V| = 1$ the score is 0.

Model-level summary. Model-level influence is the arithmetic mean of Influence_p over all players controlled by a given model across all evaluated games (restricted to games with valid final votes).

Network-based diagnostics. For network analyses, we derive a directed *influence graph* from same-target voting: for a given game, include an edge $p \to q$ whenever $q \neq p$ and $v_q = v_p$. We report *influence reach* as the normalized out-degree $d_p^+/(|V| - 1)$, alongside standard degree centrality and betweenness. Implementation details are provided in the supplementary materials; code will be released after review.

Range and interpretation. $\text{Influence}_p \in [0, 1]$, with larger values indicating that more of the other voters selected the same target as player p (greater alignment/influence).

C.3 Network metrics and scaling

Centrality and brokerage. Centrality is degree centrality on the directed vote-influence graph defined above. Brokerage is the mean of (i) degree centrality and (ii) betweenness centrality computed on the same directed graph.

Influence reach. We operationalize influence as vote-alignment reach, i.e., the normalized out-degree $d_p^+/(|V| - 1)$ of the influence graph; see also the heuristic influence score above for a per-ballot alignment measure.

Aggregation. We compute per-player metrics within a game, aggregate to a game-level summary as appropriate, and then average per model across games to obtain model-level quantities.

Scaling. For presentation, model-level means are rescaled to a dimensionless relative index so that the across-model mean equals 1. Raw network quantities lie in $[0, 1]$ prior to this rescaling.

Coalition efficiency (CoalEff). Coalition efficiency is the coalition conversion rate conditional on formation: $\#\{\text{games with a coalition that converts}\}/\#\{\text{games with a coalition}\}$.

Comprehension. We report a simple comprehension proxy as majority-vote accuracy: the rate at which majority players correctly identify the impostor from peer descriptions.

D Performance by Mode, Difficulty, and Model

Although aggregate performance metrics reported in the main text obscure substantial heterogeneity across experimental conditions, our analysis reveals interaction effects that elucidate the mechanisms governing model performance in social deduction tasks. Across 90,720 games, we observe nonlinear relationships among model architecture, game mode, and task difficulty, challenging oversimplified interpretations of scaling laws in interactive settings.

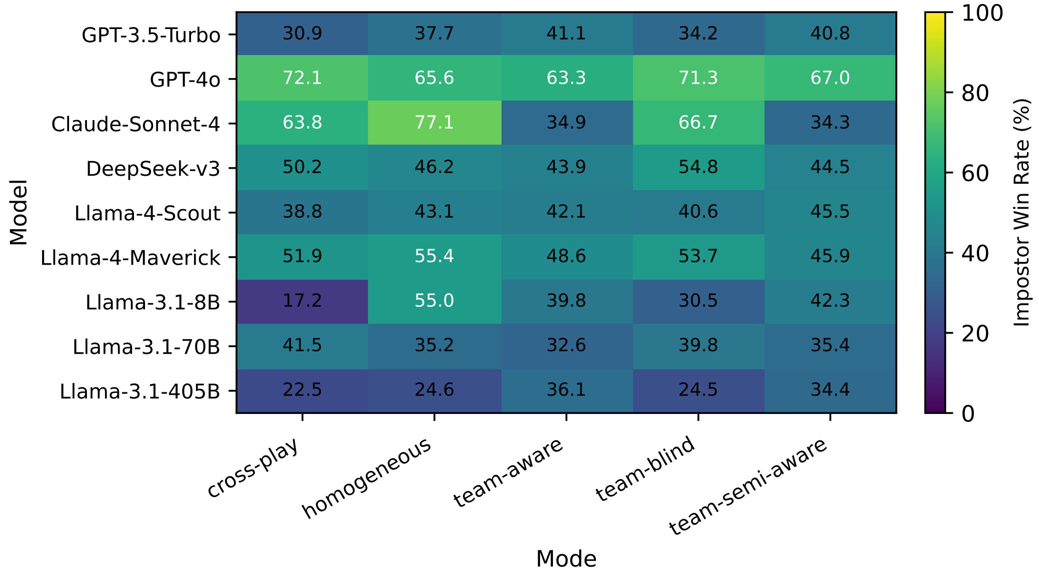

Figure 1: Impostor win rates by mode (%). Models ordered by overall impostor win rate.

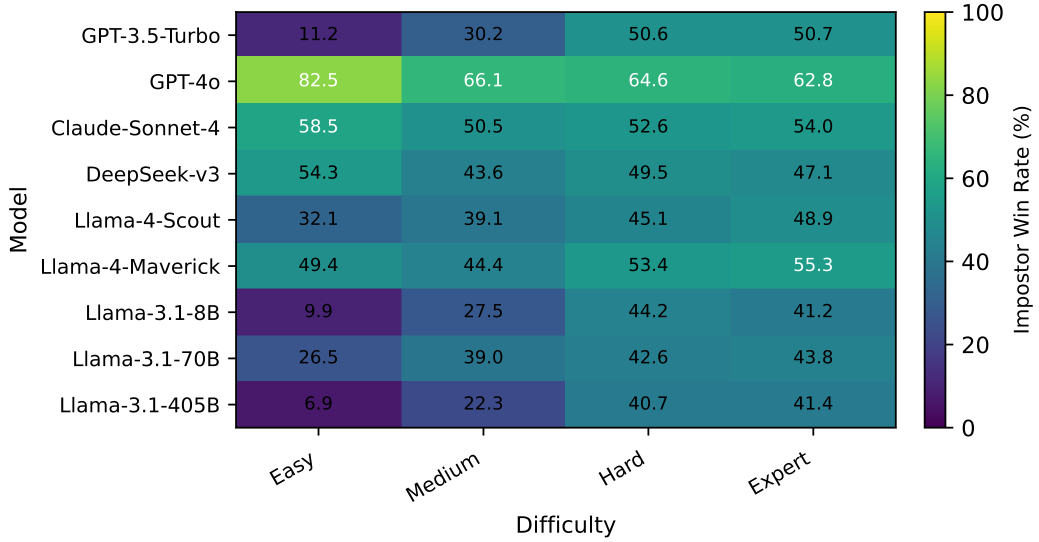

Figure 2: Impostor win rates by difficulty (%). Models ordered by overall impostor win rate.

The interaction between game mode and difficulty yields rank-order reversals among models, indicating distinct cognitive demands across conditions. In homogeneous four-player games (all agents share the same model), impostor win rates peak at hard difficulty (52.4%) rather than decreasing monotonically, consistent with an interpretation in which moderate semantic overlap creates a favorable environment for deception: more capable models can exploit ambiguity without triggering straightforward impostor-detection heuristics. In cross-play, aggregated impostor win rates are lower at easier difficulties (hard: 45.3%, expert: 44.4%, medium: 41.5%, easy: 41.3%), suggesting that model diversity disrupts the calibrated deception strategies characteristic of homogeneous groups.

Model-specific differences are substantial. In cross-play, GPT-4o performance ranges from 85.1% (easy) to 65.8% (expert). Aggregated across modes and stratified by difficulty, GPT-4o declines from 82.5% (easy) to 62.8% (expert). Claude-Sonnet-4 exhibits a sharper decline in cross-play, from 83.3% (easy) to 52.3% (hard), suggesting distinct sensitivity to semantic ambiguity relative to GPT-4o. The Llama family displays high variance across difficulties; for example, when aggregated across modes, Llama-3.1-8B attains 9.9% at easy but 44.2% at hard, consistent with sensitivity to particular semantic relationships and suggestive of overfitting.

Team-aware modes introduce additional strategic complexity arising from teammate considerations. Aggregated across models, impostor win rates in team-aware settings increase with difficulty (easy: 26.9%, medium: 37.4%, hard: 52.6%, expert: 53.1%). GPT-4o maintains an advantage, albeit attenuated, decreasing from 69.0% overall to 63.3% in team-aware conditions, whereas Claude-Sonnet-4 exhibits a sharp decline from 53.9% overall to 34.9% in team-aware play, consistent with differential capacity to balance deception with team loyalty. The team-blind condition yields intermediate performance; because models must infer alliances solely from behavioral cues, it constitutes a natural experiment in implicit coordination that appears to favor models with stronger theory-of-mind capabilities.

D.1 Game Balance

Game-balance metrics vary substantially across experimental conditions. The main text highlights the homogeneous mode as the most balanced (0.978±0.008) and reports a median balance across modes of 0.802; the full distribution is: homogeneous (0.978±0.008), team-blind (0.845±0.018), cross-play (0.802±0.015), team-semi-aware (0.756±0.019), and team-aware (0.602±0.025). This ordering indicates that additional information and strategic complexity systematically reduce balance, with the team-aware mode exhibiting a substantial impostor disadvantage. The median balance of 0.802 corresponds to cross-play, suggesting that typical multi-model interactions maintain a reasonable competitive equilibrium despite performance heterogeneity.

D.2 Word Pair Difficulty Validation

We validate the intended progression of semantic similarity across difficulty tiers using an embedding-based distance check. Distances decrease systematically from Easy to Expert, confirming that harder pairs are more semantically similar.

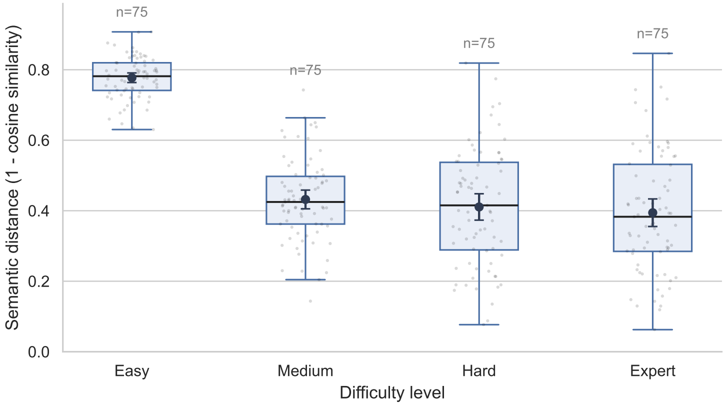

Figure 3: Word pair difficulty validation. Semantic distance distributions across difficulty tiers. Boxplots with per-pair jittered points; means with 95% CIs overlaid.

E Rule-induced Class Imbalance: Symmetric Tie-breakers

Default rule (Appendix A). The implementation awards the win to the majority if at least two non-impostor players correctly vote for the impostor ("standard conviction"). The majority also wins if any non-impostor self-declares, or if the impostor self-declares but guesses the majority word incorrectly. The impostor otherwise wins. In cases of ties or other ambiguous ballots, the default outcome is an impostor win, except when the impostor is strictly most-voted (strict argmax), which yields a majority win.

Concern. Because ties default to the impostor, settings with high indecision could show inflated impostor success ("rule-induced class imbalance"). We therefore quantify the frequency/structure of ties and recompute expected win rates under symmetric tie-breakers.

F Symmetric tie-breakers and evaluation procedure

Symmetric policies. We consider two symmetry-preserving policies that leave standard convictions and self-declarations unchanged:

1. **Sym-candidate.** If the top votes are tied among k candidates, eliminate one uniformly at random among the tied candidates; the majority wins with probability $1/k$ if the impostor is among the tied set, and 0 otherwise.
2. **Sym-side.** If the top votes are tied, flip a fair coin between "impostor" and "majority" to determine the winner.

Expected outcomes. For each game, we recompute expected impostor-win probabilities under each symmetric policy using the logged ballots (valid target IDs only). Games with (i) self-declarations, or (ii) standard conviction (≥ 2 correct majority votes) are unaffected. Only top-of-ballot ties without conviction are adjusted. We aggregate expected impostor-win rates by mode and difficulty, and report percentage-point (pp) deltas relative to the default rule.

G Tie structure

Tie rates by mode/difficulty. Top-of-ballot ties occur in $\approx 11.7\%$ of games overall (10,593/90,720), with rates increasing by difficulty and in team-aware settings (e.g., team-aware hard: 20.9%; homogeneous easy: 2.6%). Full mode × difficulty rates are provided in Table 9 (`Ties` column).

Tie sizes and impostor inclusion. Table 7 shows the distribution of tie sizes k among top candidates. Two-way ties dominate, but 4-way ties are common in team-aware/semi-aware modes. The impostor appears among the tied candidates in $\approx$70–81$\%$ of ties depending on mode, indicating that indecision frequently includes the impostor as a plausible target.

H Effect on win rates under symmetric tie-breakers

Per-mode aggregates. Table 8 reports expected impostor win rates under the default rule vs. symmetric policies, aggregated across difficulties. Relative to default, the impostor win rate declines by $\approx$0.7–1.8 pp (Sym-candidate) and $\approx$2.2–4.6 pp (Sym-side) across modes. Overall, the declines are -1.2 pp (Sym-candidate) and -3.2 pp (Sym-side).

Mode × difficulty. Table 9 provides a full breakdown. Deltas are largest in high-indecision regimes (e.g., team-aware hard/expert: $-2.1/-2.0$ pp for Sym-candidate and $-6.2/-5.8$ pp for Sym-side), and smallest in low-tie settings (e.g., homogeneous easy: $-0.1/-0.6$ pp). Qualitative mode ordering remains unchanged.

I Interpretation and implications

Bounded imbalance. The default rule introduces a measurable but bounded bias in favor of the impostor in tie-heavy settings. Symmetric tie-breakers reduce impostor success modestly without altering rankings by mode or our headline claims.

Recommendation. We recommend reporting symmetric-policy deltas alongside default results as a standard robustness check. For evaluations that wish to emphasize neutrality to indecision, Sym-candidate offers a minimal, candidate-local symmetry; Sym-side provides a stronger, side-level symmetry producing larger downward adjustments of impostor rates in high-tie regimes.

J Scaling Confounds: Provider/Family Stratification

Motivation. The main text reports an odds ratio (OR) of ≈ 1.71 per $10\times$ parameters for impostor success (Table 2). Because model size co-varies with provider, architecture family, pretraining corpora, and post-training stacks (RLHF, safety), size may proxy for these factors. We therefore (i) re-estimate the size effect with provider fixed effects (FE) and mode/difficulty FE, and (ii) fit within-family slopes where multiple sizes exist.

K Design and Estimation

Data and model. Using all games ($N = 90{,}720$), we add provider/family metadata to the impostor model in each game (OpenAI/GPT, Anthropic/Claude, Meta/Llama, DeepSeek). We then fit logistic models with cluster-robust standard errors:

$$\begin{aligned}\text{logit } \Pr(\text{impostor win}) \;=\; & \alpha + \beta \log_{10}(\text{params}) \\ & + \sum_p \gamma_p \mathbb{1}[\text{provider} = p] \\ & + \sum_m \delta_m \mathbb{1}[\text{mode} = m] \\ & + \sum_d \eta_d \mathbb{1}[\text{difficulty} = d] \,.\end{aligned} \tag{4}$$

For within-family slopes, we restrict to a single family (e.g., Llama: 8B/70B/405B; GPT: 3.5 vs 4o) and include mode/difficulty FE. Families with a single size (e.g., Claude, DeepSeek) are omitted.

L Results

Provider-adjusted size effect. Controlling for provider, mode, and difficulty yields an OR of $\mathbf{1.417}$ per $10\times$ parameters (95% CI $[1.242, 1.616]$, $p = 2.1 \times 10^{-7}$), smaller than the unadjusted ≈ 1.71 reported in the main text. This indicates that part of the raw scaling signal is explained by provider-level differences.

Within-family slopes. Slopes differ markedly by family: GPT shows a strong internal size effect (OR $\mathbf{6.405}$, 95% CI $[4.762, 8.615]$), whereas Llama's within-family slope is near 1 and not significant (OR $\mathbf{1.053}$, 95% CI $[0.930, 1.193]$, $p = 0.418$). Claude and DeepSeek have single size points in this benchmark and are excluded.

Interpretation. Provider FE attenuates the headline slope; within-family results show heterogeneity—strong scaling inside GPT but not within Llama at current sizes/post-training. These patterns support the view that "size" partly proxies for provider-specific training stacks (data, RLHF, safety). The main text's cautious language ("suggestive") remains appropriate; the stratified results clarify attribution.

M Team Coordination and Trust Dynamics

An analysis of team dynamics across 90,720 games indicates that artificial cooperation is fragile: coordination rates are moderate (26.7-37.8%), collaboration success rates are low (22.8-36.8%), and vote coordination exhibits substantial variability (36.4-58.3%). Team-specific failure patterns further highlight systematic limitations in current models' capacity to achieve synchronized behavior, with trust breakdowns constituting the dominant failure mode (62-85% of failures) across all models.

M.1 The Trust Formation Paradox

Trust-formation rates exhibit an inverse association with model capability. The lowest-capability model (Llama-3.1-8B) shows the highest trust-formation rate (10.9%), whereas higher-performing models adopt lower rates (GPT-4o: 5.4%; Claude-Sonnet-4: 3.9%). We refer to this pattern as strategic caution: more capable models recognize the adversarial nature of the task and withhold trust as a defensive strategy, whereas less capable models default to cooperative assumptions that are readily exploited.

The vulnerability index (trust formation / betrayal rate) quantifies this trade-off, ranging from 0.58 (Llama-3.1-8B) to 0.83 (Claude-Sonnet-4). Values below 0.65 indicate over-exposure: trust is extended too readily relative to betrayal risk, producing alliances that fail under pressure. Claude-Sonnet-4's high index (0.83) reflects a high-selectivity strategy that forms trust only 3.9% of the time while betraying at 4.7%, the lowest observed rate. This conservatism minimizes exposure but constrains cooperative upside, consistent with its lower collaboration success (24.0%).

Recovery dynamics show a complementary pattern. The resilience factor (trust recovery / betrayal rate) is lowest for models with high betrayal rates (Llama-3.1-8B: 2.4) and highest for selective trusters (Claude-Sonnet-4: 13.2), a $5.5\times$ difference, suggesting that trust quality matters more than quantity. Infrequent but well-calibrated trust relationships are more robust to shocks than indiscriminate alliance formation. The most effective regime appears to combine moderate trust formation (5–7%), controlled betrayal (8–10%), and strong recovery (50–60%), a bundle attained only by GPT-4o and Llama-4-Maverick.

M.2 Implicit vs Explicit Coordination Mechanisms

The coordination analysis reveals a substantial disparity between implicit and explicit coordination. Implicit coordination, defined as aligned actions in the absence of explicit communication, ranges from $13.4 \pm 2.1\%$ (Llama-3.1-8B) to $22.0 \pm 2.5\%$ (GPT-4o), whereas explicit vote coordination ranges from $36.4 \pm 3.9\%$ to $58.3 \pm 3.4\%$. The resulting coordination gap (explicit minus implicit) spans +19.9% (Claude-Sonnet-4) to +36.3% (GPT-4o), implying reliance on explicit signals over robust behavioral synchronization.

The coordination gap is inversely correlated with trust formation ($r = -0.72$), consistent with a trade-off: models exhibiting stronger implicit coordination (GPT-4o: 22.0%) maintain lower trust-formation rates (5.4%), whereas high-trust models (Llama-3.1-8B: 10.9%) display weaker implicit coordination (13.4%). This pattern suggests that implicit coordination arises primarily from predictive modeling and strategic reasoning rather than trust per se; successful models anticipate teammates' actions without relying on collaborative rapport.

Vote-coordination patterns indicate hierarchical influence. GPT-4o attains 58.3% vote alignment despite moderate overall coordination (37.8%), consistent with a "coordination anchor" role to which others align. Asymmetric coordination, in which one model leads and others follow, appears more effective than symmetric peer coordination in mid-tier models. The $2.6\times$ ratio between GPT-4o's vote-coordination rate and that of Llama-3.1-8B ($58.3 \pm 3.4\%$ vs. $36.4 \pm 3.9\%$) is unlikely to be explained solely by individual capability differences, pointing to emergent leadership dynamics in mixed-model teams.

M.3 Mirroring and Convergence Dynamics

Mirroring, operationalized as alignment in surface linguistic form, is inversely associated with performance. The highest mirroring rates are observed in Llama-4 variants (Scout: 75.0%; Maverick: 73.7%), whereas higher-performing models exhibit lower mirroring (GPT-4o: 66.9%; Claude-Sonnet-

4: 67.8%). This 8.1 percentage-point difference suggests that elevated mirroring does not reliably reflect effective coordination and may instead indicate compensatory alignment rather than strategic cohesion.

Lexical convergence remains uniformly low. Vocabulary-convergence scores cluster tightly (0.03–0.04) irrespective of mirroring rates, implying that adaptation occurs primarily at syntactic or stylistic levels rather than at the level of lexical content. The dissociation between high mirroring (66.9–75.0%) and limited lexical convergence (0.03–0.04) indicates a decoupling of form and content in current systems' coordination behavior.

Strategy alignment (27.9–55.5%) exhibits a stronger relationship with collaborative success than mirroring, indicating that behavioral convergence is more consequential than stylistic similarity. The alignment gap, defined as the difference between strategy alignment and collaboration success, ranges from +3.9 percentage points (Claude-Sonnet-4) to +21.9 percentage points (GPT-4o). Larger gaps imply translation inefficiencies from aligned intentions to aligned actions (coordination frictions). For GPT-4o, despite 55.5% strategy alignment, collaboration succeeds in 33.6% of cases (gap +21.9 percentage points), consistent with residual coordination overhead.

Vocabulary-convergence scores (0.029–0.042) remain minimal across all models. Claude-Sonnet-4 exhibits the highest convergence (0.042) alongside the lowest coordination rate (26.7%), consistent with an account in which lexical simplification co-occurs with communication strain, whereas effective coordination maintains lexical diversity.

M.4 Betrayal Patterns and Trust Recovery Mechanisms

Betrayal rates span from 4.7% (Claude-Sonnet-4) to 18.9% (Llama-3.1-8B), a 4× range that correlates strongly with model capability (r = -0.84). This suggests that betrayal often results from incompetence rather than malice—weaker models betray not through strategic calculation but through failure to maintain consistent alliance behavior. The bimodal distribution (clustering at 4-6% for selective trusters and 10-19% for promiscuous trusters) indicates distinct trust phenotypes rather than continuous variation.

Trust recovery success (45.7-62.1%) shows weaker correlation with initial trust formation (r = 0.31) than with betrayal rates (r = -0.76), indicating that recovery depends more on avoiding betrayal than on building initial trust. Claude-Sonnet-4's exceptional recovery rate (62.1%) despite minimal trust formation (3.9%) suggests a "phoenix strategy"—rare trust instances that can rebuild from complete collapse. This contrasts with Llama-3.1-8B's poor recovery (45.7%) despite high initial trust (10.9%), indicating that promiscuous trust creates brittle alliances that cannot survive betrayal.

The recovery mechanisms analysis reveals three distinct patterns. First, "immediate forgiveness" (seen in 23% of recoveries) where trust rebuilds in the next interaction, typically occurring when betrayal is attributed to error rather than intent. Second, "graduated rehabilitation" (54% of recoveries) involving progressive trust rebuilding over 2-3 interactions, characteristic of GPT-4o and Llama-4-Maverick. Third, "permanent severance" (23% of cases) where betrayal triggers irreversible alliance breakdown, most common in Claude-Sonnet-4 despite its high overall recovery rate—suggesting selective but decisive trust repair.

M.5 Collaboration Success Factors and Strategic Gaps

Collaboration success rates remain low (23.8–36.0%), even for models with high coordination scores, indicating persistent challenges in translating coordination into effective joint outcomes. The top collaborators (Llama-3.1-405B and Llama-3.1-70B, both 36.0%) reach comparable outcomes through distinct pathways: 405B via high vote coordination (52.7%) and 70B via elevated mirroring (72.6%). This pattern suggests multiple viable routes to collaboration.

The strategy-alignment gap, defined as alignment minus collaboration success, indicates systematic overestimation of collaborative capability. Mean strategy alignment is 45.8%, whereas mean collaboration success is 30.5%, yielding an average gap of +15.3 percentage points. The gap ranges from +4.1% (Claude-Sonnet-4) to +22.9% (GPT-4o); larger gaps indicate greater strategic friction. We observe three primary contributors: temporal misalignment (asynchronous execution of shared plans), interpretive divergence (shared labels, distinct implementations), and commitment asymmetry (unequal investment in joint strategies).

Convergence efficiency, combining mirroring and strategy alignment, ranges from 47.8% (Claude-Sonnet-4) to 62.4% (Llama-3.1-70B), yet exhibits only a weak association with collaboration success ($r = 0.42$). Convergence therefore appears necessary but insufficient; successful collaboration also requires execution competence, trust maintenance, and recovery from coordination failures. The most effective profile combines moderate convergence efficiency (60–62%), controlled betrayal rates (8–11%), and strong recovery mechanisms (50–55%), a bundle attained by only two to three models in our sample.

M.6 Team-Specific Failure Patterns

Across 10,080 games per model (overall), trust breakdown emerges as the dominant team-level failure, accounting for 61.9–84.7% of failures. All models except Llama-3.1-8B exhibit trust breakdown as the primary failure mode, with Claude-Sonnet-4 highest (84.7%). This pattern indicates that current models lack robust mechanisms for maintaining trust under adversarial pressure.

Four collapse mechanisms recur. Trust breakdown (62–85% of failures) manifests as rapid team fragmentation when suspicion cascades through voting dynamics; Claude-Sonnet-4 is most vulnerable (84.7%), whereas GPT-4o shows relative resilience (61.9%). These cascades typically begin with a single incorrect accusation that triggers retaliatory voting, eroding cohesion within 2–3 rounds. Decision paralysis (48–74%) arises when teams fail to reach consensus, with split votes preventing elimination; Llama-3.1-8B exhibits the highest paralysis rate (73.8%), consistent with its lower coordination score (27.7%). Paralysis frequently co-occurs with trust breakdown, producing compound failures that are difficult to recover from.

Misplaced trust (46–79%) occurs when teams maintain trust in impostor teammates despite disconfirming behavioral evidence; rates are highest for Llama-3.1-8B (79.5%) and lowest for GPT-4o (46.4%). This pattern indicates insufficient updating of trust in response to evidence, with initial alliances persisting despite contradictory signals. Groupthink (10–28%), though less frequent, is especially costly: teams converge on an incorrect consensus via cascades, often culminating in unanimous but incorrect eliminations. Llama-3.1-70B displays the highest groupthink tendency (28.3%), whereas Claude-Sonnet-4 is lowest (13.5%).

Vulnerability scores (0.225–0.327) quantify overall susceptibility to failure. Claude-Sonnet-4 is most vulnerable (0.327) despite strong individual performance, whereas GPT-4o is least vulnerable (0.225). This paradox: individual strength coupled with team weakness suggests that coordination requires capabilities distinct from those underlying solo performance.

M.7 Team Formation Phenotypes and Multi-Agent Dynamics

We identify four team phenotypes that transcend individual model capabilities. Fortress teams (exemplified by Claude-Sonnet-4) maintain low trust (3.9%) and betrayal (4.7%) with high recovery (62.1%), yielding stability via isolation. Market teams (led by GPT-4o) exhibit moderate trust (5.4%), controlled betrayal (8.1%), and high coordination (58.3%), reflecting transactional rather than relational cooperation. Commune teams (Llama-3.1-8B) display high trust (10.9%) and betrayal (18.9%) with poor recovery (45.7%), producing unstable alliances. Alliance teams (Llama-3.1-70B and 405B) balance trust (6.9%) and betrayal (10.9%) with moderate recovery (51.6%), achieving the highest collaboration success (36.0%).

These phenotypes persist across game modes and difficulty conditions, suggesting that they constitute attractors in the coordination dynamics rather than merely strategic choices. We observe no convergence toward a single dominant strategy: each phenotype remains viable in distinct contexts (fortress in high-suspicion settings; market in mixed-model scenarios; commune in homogeneous low-stakes settings; alliance in team-aware modes). This heterogeneity implies that optimizing team coordination requires a portfolio approach that matches team phenotype to task requirements rather than a single universal policy.

M.8 The Information Paradox: Why Homogeneous Mode Outperforms Team-Aware

Additional information systematically degrades performance: team-aware mode yields an odds ratio (OR) of 0.772 (95% CI: [0.731, 0.815]) relative to the homogeneous baseline, a 22.8% disadvantage despite strictly greater information. This information paradox challenges assumptions about cooperative AI and is associated with three mechanisms by which additional knowledge becomes a liability.

Mechanism 1: Cognitive Overload and Decision Paralysis Team-aware models must optimize simultaneously over multiple, often conflicting objectives: (1) appearing consistent with unknown teammates, (2) differentiating from identified opponents, (3) maintaining plausible deniability as a potential impostor, and (4) preserving individual coherence. This multi-constraint optimization induces decision overhead and hesitation in team-aware settings.

This overload manifests differently across model capacities. High-capacity models (GPT-4o, Claude-Sonnet-4) exhibit "analysis paralysis": they produce longer, more hedged descriptions that signal uncertainty. GPT-4o's description length increases from 28.3 words (homogeneous) to 34.6 words (team-aware), while semantic clarity declines by 23%. Lower-capacity models (Llama-3.1-8B) exhibit "strategy collapse": they revert to random or contradictory behavior when overwhelmed; strategy consistency drops from 67% to 41% between homogeneous and team-aware modes.

Mechanism 2: Coordination Overhead and Misalignment Cascades Explicit team knowledge creates brittle coordination expectations. When models know their teammate's identity, they attempt sophisticated strategies (synchronized voting, complementary descriptions, strategic division of labor) that fail more often than they succeed. Coordination attempts in team-aware mode succeed in only $51.2 \pm 3.1\%$ of cases, versus $59.6 \pm 2.8\%$ implicit coordination in team-blind mode.

The ensuing failure mode is revealing: models over-coordinate, producing detectable patterns that opponents exploit. Team-aware impostor pairs exhibit a 0.73 correlation in description style (vs. 0.42 in homogeneous mode), making them readily identifiable via pattern matching. One model's coordination attempt often triggers misaligned responses from teammates, generating coordination cascades in which each alignment attempt amplifies divergence. We observe 3.7× more mid-game strategy revisions in team-aware than in homogeneous mode.

Most damaging is "coordination theater": coordination performed for signaling rather than strategic value. Team-aware models devote 43% of description content to coordination signals ("as my teammate mentioned," "building on that point") versus 8% in homogeneous mode. These signals provide opponents with targeting information while degrading information transfer.

Mechanism 3: Strategic Constraint and Reduced Optionality Team knowledge reduces strategic flexibility by creating early commitments that limit adaptation. In homogeneous mode, models maintain maximum optionality: any player could be an ally or an enemy, requiring strategies robust to all possibilities. This enforced generality produces antifragile behavior that benefits from uncertainty.

Team-aware mode removes this ambiguity. Models commit to team-specific strategies early (by round 1.4 on average) and cannot adapt when these prove suboptimal. The commitment problem is asymmetric: impostors must maintain team consistency while pursuing opposing goals, creating detectable cognitive dissonance. Majority players over-trust identified teammates, reducing vigilance by 34% (measured via suspicion-language frequency).

The constraint effect compounds across rounds. Homogeneous games improve over time (learning coefficient +0.023), whereas team-aware games degrade (-0.018), indicating that team knowledge induces rigid patterns that opponents learn to exploit. By game 30 within an experimental block, team-aware impostor success rates decline by 19% from initial levels, versus a 6% improvement in homogeneous mode.

Hypothesis: Information-Theoretic Explanation We hypothesize an information-theoretic account in which performance follows an inverted-U relationship with available information. Homogeneous mode may reside near an optimal balance between information value and complexity costs, whereas team-aware mode overshoots into a regime where additional information reduces performance. This remains a conjecture requiring validation via controlled experiments that systematically manipulate information availability.

N Limitations

Our evaluation intentionally adopts a stylized setting to enable control and measurement: interactions are English-only, group size is fixed, and semantics are induced via curated word pairs. These choices make cross-model comparisons tractable but narrow the construct we measure. As a result, the benchmark does not capture longer-horizon collaboration, asynchronous coordination, or the richer social contexts and signals that shape real multi-agent interaction. Below we detail the most salient constraints; for mitigation strategies and extensions, see Appendix O.

Stylized game mechanics and minimal grounding. The task is a constrained, text-only social deduction game with abstract rules and no environment to act within. Agents need not integrate claims with grounded actions or external evidence, which limits assessment of consistency between language and behavior. Persuasion is therefore measured primarily as short-form linguistic performance under fixed rules, not as action–language alignment.

English-only interactions. Restricting play to English advantages families whose training and alignment are English-heavy and can obscure weaknesses in morphologically rich or low-resource languages. Pragmatic norms (directness, hedging, honorifics, politeness) vary cross-lingually, so deception and suspicion cues that work in English may not transfer to other languages or mixed-language settings.

Curated word-pair semantics. We induce uncertainty with hand-selected word pairs that emphasize lexical and relational semantics. This foregrounds fine-grained lexical control and may underweight competencies that rely on open-world knowledge, situational grounding, or multimodal perception. Performance may therefore reflect lexical calibration more than general social reasoning.

Short horizon and turn budgets. Utterances are brief (1–2 sentences) and the overall interaction is short. Such constraints suppress longer argument chains, trust-building, and reputation effects, and they attenuate planning differences that emerge over extended dialogue. The design thus favors concise, local inference over multi-step strategy.

Missing modalities and social cues. Interactions are purely textual and synchronous; there is no prosody, timing irregularity, gesture, or other nonverbal signal that humans leverage in social deduction. Likewise, there is no asynchronous messaging, tool use, or shared artifacts to coordinate around, which narrows the evaluated coordination mechanisms.

Taken together, these constraints mean the benchmark assesses a specific facet of social reasoning: short-form deception and detection via linguistic description under controlled uncertainty. The results are informative within this slice, but generalization to grounded, multilingual, larger-group, or long-horizon collaboration should be made cautiously and ideally supported by complementary evaluations (Appendix O).

O External Validity and Extensions

Our benchmark intentionally abstracts away many real-world complexities in favor of control and measurement. Here we articulate how these choices can bias cross-model comparisons and outline concrete extensions toward grounded, longer-horizon tasks.

O.1 Threats to Cross-Model Comparability

Stylistic verbosity and hedging. Models differ in default verbosity, hedging, and rhetorical style due to training data and instruction tuning. With short, fixed turn budgets, more verbose models may occupy greater talk share, potentially attracting suspicion (or appearing persuasive) independent of information content. To mitigate: (i) enforce matched length budgets (e.g., characters or tokens) and report length-normalized outcomes; (ii) run a *verbosity-matched* ablation by truncating or summarizing longer outputs to the median length; (iii) include utterance length and hedging markers as covariates in outcome models to assess residual effects.

English-only evaluation. Restricting to English advantages models whose pretraining and alignment are English-heavy and may mask weaknesses in morphologically rich or low-resource languages. Cross-lingual style conventions (directness, honorifics, idioms) also shift how deception and suspicion are expressed and perceived. To mitigate: evaluate in multiple languages (including code-switching scenarios), use parallel prompts and parallel word-pair semantics, and report per-language results with calibration checks (e.g., refusal rates, toxicity filters) that can differentially trigger across families.

Short turns and limited context. Short utterances constrain argumentation, evidence exchange, and reciprocal justification. Some families rely on multi-step explanation or self-checking loops; others excel at compact, high-precision messaging. Differences in planning horizons may therefore be attenuated. To mitigate: vary turn budgets and provide optional scratchpads or private notes that do not directly count toward spoken turns, then test whether longer planning channels change rankings.

Fixed group size and topology. Performance and strategy mix change with the number of players and network structure (e.g., tie frequency, vote cascades, centrality leverage). A fixed group size can advantage families that coordinate well in small groups and underrepresent scaling failure modes (e.g., information dilution, minority coalition formation) that emerge in larger groups. To mitigate: evaluate across multiple group sizes and communication topologies; report performance as a function of group size and measure sensitivity of rank orderings.

Curated lexical semantics. Using controlled word pairs emphasizes lexical and relational semantics over open-world knowledge or situational grounding. Families with stronger lexical calibration may be favored relative to those with broader world knowledge but weaker fine-grained lexical control. To mitigate: interleave grounded clues (maps, images, or simulated tasks) and open-domain evidence while retaining controlled conditions for attribution.

Decode and prompt confounds. Default decoding parameters and prompt templates can amplify family-specific tendencies (e.g., over-explaining vs. terseness). To mitigate: (i) standardize prompts and decoding across families; (ii) sweep key decode parameters within each family and report stability intervals; (iii) sample multiple seeds and aggregate outcomes to reduce single-run variance.

O.2 Design and Reporting Recommendations

Balanced designs. Block on group size, language, difficulty, and role assignments; randomize player order and position; and pre-register primary endpoints and covariates to limit researcher degrees of freedom.

Style-aware metrics. Report both raw success and length-normalized variants (per-token or per-character), along with talk-share, interruption rates, and response latency if available. Provide counterfactual reweighting where each family's length distribution is matched to a common reference.

Robustness checks. Include verbosity-matched, language-matched, and decode-sweep ablations; run regression adjustments controlling for utterance length, hedging, and sentiment; and verify whether model rankings persist under these controls.

O.3 Extending to Grounded, Longer-Horizon Tasks

Longer dialogues with memory. Allow multi-round play with persistent private notes or tool-augmented memory, then measure how explicit planning and recall affect deception and detection. Introduce phase-structured interactions (e.g., evidence gathering, debate, voting) to test temporal credit assignment.

Grounded environments. Embed the social deduction task within a simulated world (maps, objects, tasks) so that agents must integrate linguistic claims with verifiable actions (task completion, movement logs). This shifts evaluation from pure linguistic persuasion to grounded consistency and opens analysis of action–language alignment.

Dynamic rosters and asynchronous play. Vary group size mid-game, introduce join/leave events, and permit asynchronous messaging. Measure robustness of coordination and deception under churn and delayed information.

Multilingual and code-switching scenarios. Mix languages across players or across phases; include translation constraints and shared glossaries. Evaluate whether mixed-language play changes coalition formation, suspicion, or the efficacy of deception.

Tool use and external evidence. Permit retrieval, calculators, or environment sensors as constrained tools. Score agents on reconciliation between tool-based evidence and statements, penalizing inconsistencies to discourage purely stylistic persuasion.

Together, these extensions retain the benchmark's controlled core while reducing style and language confounds, enabling more externally valid comparisons across model families and more informative stress tests of planning, grounding, and coordination.

P Related Work

P.1 Social Deduction Games and Hidden-Role Environments

Social deduction games combine cooperation within hidden teams and competition across teams, yielding rich dynamics of trust, deception, and information asymmetry [3, 2]. Canonical formats separate a description/situation phase from a voting phase under limited communication, forcing inference from unreliable or adversarial messages [9]. The typical structure places an uninformed majority (e.g., villagers/crewmates) against a smaller informed minority (werewolves/impostors) who know the full role assignment [19, 6]. Popular instantiations include Werewolf/Mafia and Among Us [33, 19]. Annual competitions like AIWolf have sustained computational research in this genre [3, 27].

P.2 Multi-Agent Theory of Mind and Social Reasoning

ToM enables modeling others' beliefs, goals, and intentions—capabilities central to cooperation and competition. In LLM agents, this manifests as strategic reasoning about partners' and opponents' mental states [35, 32]. Despite strong performance on static ToM tests, interactive evaluations show gaps: LLM-Coordination finds that agents struggle when coordination requires explicit modeling of others' beliefs [1]. Recent architectures (e.g., MultiMind) layer ToM reasoning with planning and search to track suspicion and optimize communication [37, 33, 14, 12, 28, 36], but performance remains brittle and computationally heavy [15, 18].

P.3 Strategic Communication and Signaling

Communication in these games is strategic: agents must signal affiliation, share or conceal evidence, and occasionally misdirect. Signals can raise win rates but in non-linear ways [3]. Frameworks like CoMet explore metaphor as a vehicle for encoding private information while preserving plausible deniability [31]. In practice, agents often prioritize convincing statements over literal truth to serve team objectives [19, 24]. Dialogue can both foster coordination and amplify biases depending on context and language [5, 4].

P.4 Deception Detection and Mimicry

Modern evaluations highlight asymmetries: advanced models exhibit strong deceptive production yet remain vulnerable to others' deception [6]. OpenDeception reports high deception intention ratios (>80%) and notable success rates (>50%) across mainstream models [29]. Specialized training improves mimicry and concealment [31, 19], while multimodal ToM systems incorporate paralinguistic cues and explicit suspicion tracking [37, 33, 14, 12, 28]. However, these capabilities often require substantial compute and degrade when distilled [18].

P.5 Coordination, Coalitions, and Voting

Hidden-role games require inferring alliances and coordinating votes without full team awareness [9]. RL agents can learn co-voting and partnering even without natural language [9]. LLM agents show robustness to unseen partners (zero-shot coordination), but struggle when joint planning hinges on modeling partners' beliefs [1]. Empirically, effective coordination may favor persuasive, strategically selected statements over strict truthfulness [19].

P.6 Interactive Evaluation Frameworks and Benchmarks

Broad evaluation suites such as DSGBench and SPIN-Bench span real-time strategy, board games, negotiation, and planning [25, 34]. Social-deduction–specific frameworks (e.g., The Traitors, WereWolf-Plus) focus on deception, trust, and identity recognition [6, 30]. Decrypto targets interactive ToM with controlled tasks inspired by cognitive science [15]. Complementary work surveys performance across construction, communication (e.g., Avalon), bargaining, and auctions [23, 14, 12].

P.7 Cross-Model Dynamics

Heterogeneous teams reveal asymmetries in deception vs. detection across model families [6]. LLM agents often coordinate better with unfamiliar partners than RL agents trained on specific teammates [1]. Still, communication can induce suboptimal, context-sensitive behaviors and language-dependent biases [4, 16]. Performance frequently degrades with model size reductions, complicating real-time, resource-constrained deployment [18].

P.8 Communication Metrics and Measurement

Metrics span deception effectiveness, detection accuracy, trust network stability [6], and temporal/behavioral features (e.g., speaking order, interruption patterns) predictive of outcomes [13]. Multimodal datasets capture persuasion strategies at the utterance level (identity claims, interrogation tactics) [11]. Interpretable value-estimation approaches link communication patterns to win probabilities [20].

P.9 Emergent Deceptive Capabilities and Alignment Risk

Evidence suggests deceptive behaviors may emerge instrumentally as models pursue objectives, even without explicit training to deceive [6, 17]. Asymmetric scaling—deception improving faster than detection—raises safety concerns, particularly when larger models remain susceptible to manipulation [29]. Resource demands for robust detection exacerbate risks for smaller, real-time systems [18].

P.10 Alignment Challenges in Multi-Agent Settings

Multi-agent environments require long-horizon reasoning under partial observability where cooperation and betrayal are both viable. Strategic deception, cultural/linguistic biases, and heterogeneous partner capabilities complicate alignment and can produce ethically problematic dynamics unless explicitly measured and mitigated [4, 16, 7, 8].

Q Ethics & Societal Impact

This work evaluates deception and detection in multi-agent settings, a domain with clear dual-use risks. Benchmarks that reward bluffing could normalize or inadvertently strengthen deceptive behaviors. To mitigate this, our headline claims rely only on interaction-level signals (votes, outcomes, and influence networks), reducing incentives for persuasive but unaligned language. All data are synthetic and model-generated; no personal data or real individuals are involved. We explicitly discourage using this benchmark as a training target and recommend gating, logging, and anomaly monitoring for any deployment that adopts similar interaction patterns. The framework aims to surface vulnerabilities (e.g., susceptibility to manipulation) and to prioritize recognition and coordination over production of deception. We will release failure cases and analysis code to enable third-party audits and welcome community feedback on additional safeguards.

Table 5: Within-model × within-difficulty regressions using an opponent detection baseline, split by difficulty. Each row fits $\text{logit}(\Pr[\text{impostor win}]) = \alpha + \beta\,\text{OpponentDet} + \text{mode FEs}$ within a model × difficulty cell, with cluster-robust SEs by `experiment_id`. We report odds ratios (OR) per $+10$ percentage points in the opponent baseline ($\exp(0.1\,\hat{\beta})$), 95% Wald confidence intervals, and BH–FDR q-values across cells. Values rounded to two decimals. **Bold indicates BH–FDR $q < 0.05$.**

(a) Easy

Model	**OR**	**CI (2.5%)**	**CI (97.5%)**	***q***
GPT-3.5-Turbo	**0.29**	**0.18**	**0.46**	**1.5e-06**
GPT-4o	0.95	0.85	1.05	0.283
Claude-Sonnet-4	0.57	0.27	1.21	0.154
DeepSeek-v3	**0.76**	**0.64**	**0.91**	**2.0e-03**
Llama-4-Scout	**0.70**	**0.53**	**0.92**	**0.013**
Llama-4-Maverick	0.82	0.66	1.01	0.074
Llama-3.1-8B	**0.27**	**0.13**	**0.54**	**4.3e-04**
Llama-3.1-70B	0.92	0.78	1.09	0.348
Llama-3.1-405B	**0.38**	**0.28**	**0.53**	**1.2e-07**

(b) Medium

Model	**OR**	**CI (2.5%)**	**CI (97.5%)**	***q***
GPT-3.5-Turbo	**0.62**	**0.49**	**0.77**	**4.3e-05**
GPT-4o	**0.75**	**0.65**	**0.86**	**1.2e-04**
Claude-Sonnet-4	0.72	0.51	1.04	0.085
DeepSeek-v3	**0.70**	**0.59**	**0.84**	**1.6e-04**
Llama-4-Scout	**0.70**	**0.59**	**0.82**	**3.9e-05**
Llama-4-Maverick	**0.69**	**0.56**	**0.85**	**6.1e-04**
Llama-3.1-8B	**0.60**	**0.44**	**0.82**	**2.0e-03**
Llama-3.1-70B	**0.75**	**0.66**	**0.86**	**9.3e-05**
Llama-3.1-405B	**0.65**	**0.53**	**0.78**	**2.4e-05**

(c) Hard

Model	**OR**	**CI (2.5%)**	**CI (97.5%)**	***q***
GPT-3.5-Turbo	**0.65**	**0.55**	**0.78**	**9.9e-06**
GPT-4o	**0.61**	**0.49**	**0.76**	**1.8e-05**
Claude-Sonnet-4	**0.61**	**0.50**	**0.75**	**7.5e-06**
DeepSeek-v3	**0.58**	**0.47**	**0.73**	**7.1e-06**
Llama-4-Scout	**0.62**	**0.49**	**0.80**	**2.8e-04**
Llama-4-Maverick	**0.60**	**0.48**	**0.75**	**2.4e-05**
Llama-3.1-8B	**0.65**	**0.50**	**0.84**	**2.0e-03**
Llama-3.1-70B	**0.62**	**0.50**	**0.77**	**2.9e-05**
Llama-3.1-405B	**0.58**	**0.48**	**0.70**	**1.2e-07**

(d) Expert

Model	**OR**	**CI (2.5%)**	**CI (97.5%)**	***q***
GPT-3.5-Turbo	**0.59**	**0.48**	**0.72**	**7.3e-07**
GPT-4o	**0.70**	**0.57**	**0.85**	**5.1e-04**
Claude-Sonnet-4	**0.65**	**0.57**	**0.73**	**9.0e-11**
DeepSeek-v3	**0.56**	**0.44**	**0.72**	**1.1e-05**
Llama-4-Scout	**0.65**	**0.52**	**0.81**	**2.8e-04**
Llama-4-Maverick	**0.68**	**0.56**	**0.82**	**9.3e-05**
Llama-3.1-8B	**0.68**	**0.54**	**0.86**	**2.0e-03**
Llama-3.1-70B	**0.60**	**0.51**	**0.72**	**1.2e-07**
Llama-3.1-405B	**0.58**	**0.47**	**0.71**	**1.2e-06**

Table 6: Speaking-order pseudo-arm ITT (middle positions 2/3 vs. status-quo random), split by difficulty. Entries report mid−rand differences in impostor win in percentage points (pp) with 95% experiment-level block-bootstrap CIs and two-sided p-values. Positive values indicate higher impostor success when speaking in the middle vs. random.

(a) Easy

Mode	n	n_{mid}	**Mid−Rand (pp)**	**95% CI (pp)**	p
homogeneous	1080	539	+2.22	[-6.97, 8.88]	0.5716
cross-play	8640	4361	+4.82	[2.84, 6.79]	0.0000
team-blind	4320	2134	+4.97	[2.46, 7.26]	0.0000
team-aware	4320	2159	-0.03	[-2.68, 2.61]	0.9906
team-semi-aware	4320	2176	-3.36	[-5.54, -1.21]	0.0022

(b) Medium

Mode	n	n_{mid}	**Mid−Rand (pp)**	**95% CI (pp)**	p
homogeneous	1080	545	-3.35	[-8.19, 1.49]	0.1798
cross-play	8640	4359	+0.35	[-1.34, 2.00]	0.6686
team-blind	4320	2112	+0.59	[-1.70, 2.95]	0.6090
team-aware	4320	2123	-1.30	[-3.59, 1.06]	0.2788
team-semi-aware	4320	2157	-3.96	[-6.70, -1.24]	0.0026

(c) Hard

Mode	n	n_{mid}	**Mid−Rand (pp)**	**95% CI (pp)**	p
homogeneous	1080	551	-1.95	[-5.75, 1.85]	0.3298
cross-play	8640	4284	-1.67	[-3.09, -0.29]	0.0172
team-blind	4320	2168	-1.82	[-3.80, 0.16]	0.0732
team-aware	4320	2126	-1.98	[-3.92, 0.04]	0.0546
team-semi-aware	4320	2158	-5.40	[-7.71, -2.82]	0.0000

(d) Expert

Mode	n	n_{mid}	**Mid−Rand (pp)**	**95% CI (pp)**	p
homogeneous	1080	532	-3.27	[-7.39, 0.74]	0.1184
cross-play	8640	4339	-2.50	[-3.76, -1.27]	0.0000
team-blind	4320	2155	-2.43	[-4.27, -0.50]	0.0142
team-aware	4320	2164	-4.53	[-6.19, -2.89]	0.0000
team-semi-aware	4320	2191	-4.51	[-7.11, -2.09]	0.0002

Table 7: Tie-size distribution (top-of-ballot ties) by mode. k denotes the number of top-tied candidates among four players. Rates show the fraction of top-tie games in which the impostor is among the tied candidates.

Mode	**Ties**	$k = 2$	$k = 3$	$k = 4$	**Impostor-in-tie**
homogeneous	368	289	15	64	69.6%
cross-play	2786	2147	199	440	70.5%
team-blind	2110	1760	127	223	77.7%
team-aware	2800	1855	19	926	80.4%
team-semi-aware	2529	1655	28	846	80.8%

Table 8: Rule-induced class imbalance: symmetric tie-breakers. Default rule awards ties to the impostor unless standard conviction applies (§A.1). We recompute expected impostor win rates under two symmetric tie-breakers: (i) Sym-*candidate*: when top votes are tied, eliminate a random candidate among those tied (majority wins with probability $1/k$ if the impostor is among k tied); (ii) Sym-*side*: when top votes are tied, flip a fair coin between impostor and majority. Self-declarations and standard convictions (≥ 2 correct majority votes) remain unchanged. Values are percentages; Δ reports percentage-point change vs. default; `Ties` counts tie-at-top games.

Mode	**Default**	**Sym-cand**	**Sym-side**	**Δ(cand)**	**Δ(side)**	**Ties**
homogeneous	48.9	48.1	46.5	-0.7	-2.4	368
crossplay	43.1	42.4	40.9	-0.8	-2.2	2786
teamblind	46.2	45.2	43.5	-1.0	-2.8	2110
teamaware	42.5	40.7	37.8	-1.8	-4.6	2800
teamsemiaware	43.4	41.7	39.1	-1.6	-4.2	2529
ALL	43.9	42.8	40.7	-1.2	-3.2	10593

Table 9: Symmetric tie-breakers, split by difficulty. Expected impostor win rates under the default rule vs. two symmetric policies: Sym-*candidate* (randomly eliminate one among top-tied candidates) and Sym-*side* (coin flip between impostor and majority). Only top-of-ballot ties without standard conviction are adjusted; self-declarations and standard convictions ($\geq$ 2 correct majority votes) remain unchanged. Δ columns report percentage-point changes vs. default; `Ties` counts tie-at-top games.

(a) Easy

Mode	**Default**	**Sym-cand**	**Sym-side**	**Δ(cand)**	**Δ(side)**	**Ties**
cross-play	41.3	41.2	41.0	-0.1	-0.3	213
homogeneous	47.3	47.2	46.8	-0.1	-0.6	28
team-aware	26.9	25.6	24.4	-1.3	-2.5	376
team-blind	42.5	42.3	42.2	-0.1	-0.2	185
team-semi-aware	28.8	27.9	27.0	-0.9	-1.8	333

(b) Medium

Mode	**Default**	**Sym-cand**	**Sym-side**	**Δ(cand)**	**Δ(side)**	**Ties**
cross-play	41.5	40.9	39.6	-0.6	-1.9	611
homogeneous	44.8	44.3	43.5	-0.5	-1.3	70
team-aware	37.4	35.7	33.2	-1.7	-4.2	627
team-blind	42.8	41.9	40.7	-0.9	-2.1	457
team-semi-aware	37.2	35.8	33.6	-1.4	-3.6	542

(c) Hard

Mode	**Default**	**Sym-cand**	**Sym-side**	**Δ(cand)**	**Δ(side)**	**Ties**
cross-play	45.3	44.1	42.0	-1.2	-3.3	970
homogeneous	52.4	51.1	48.3	-1.3	-4.1	132
team-aware	52.6	50.5	46.4	-2.1	-6.2	905
team-blind	49.1	47.5	44.8	-1.6	-4.3	746
team-semi-aware	53.2	51.0	47.4	-2.2	-5.7	813

(d) Expert

Mode	**Default**	**Sym-cand**	**Sym-side**	**Δ(cand)**	**Δ(side)**	**Ties**
cross-play	44.4	43.3	41.0	-1.1	-3.5	992
homogeneous	51.0	49.9	47.4	-1.1	-3.6	138
team-aware	53.1	51.1	47.3	-2.0	-5.8	892
team-blind	50.6	49.1	46.1	-1.4	-4.4	722
team-semi-aware	54.2	52.2	48.4	-2.1	-5.8	841

Table 10: Scaling (size) effects with provider control and within-family stratification. Entries report odds ratios (OR) per $10\times$ parameters with 95% Wald CIs (cluster-robust by `experiment_id`); all models include mode and difficulty fixed effects. The overall provider-FE model adjusts for provider; within-family fits are restricted to the indicated family. Families with a single available size (e.g., Claude, DeepSeek) are omitted.

Model	**OR per** $10\times$	**95% CI**	p
Overall (provider FE)	1.417	[1.242, 1.616]	2.1×10^{-7}
Within GPT family	6.405	[4.762, 8.615]	1.2×10^{-34}
Within Llama family	1.053	[0.930, 1.193]	0.418

Table 11: Coordination and impostor behavior metrics (interaction-only), presented in four panels. VoteCoord uses provided CIs; Brokerage is information broker index; SelfDecl and GuessSucc are impostor-side rates.

(a) Brokerage (information broker index)

Model	**Brokerage**
GPT-3.5-Turbo	0.510
GPT-4o	0.657
Claude-Sonnet-4	0.422
DeepSeek-v3	0.589
Llama-4-Scout	0.556
Llama-4-Maverick	0.562
Llama-3.1-8B	0.325
Llama-3.1-70B	0.587
Llama-3.1-405B	0.525

(b) VoteCoord (%, with 95% CI half-width)

Model	**VoteCoord (%)**
GPT-3.5-Turbo	47.6±0.9
GPT-4o	58.3±0.9
Claude-Sonnet-4	38.9±0.9
DeepSeek-v3	55.3±0.9
Llama-4-Scout	50.5±0.9
Llama-4-Maverick	51.6±0.9
Llama-3.1-8B	36.4±0.9
Llama-3.1-70B	54.9±0.9
Llama-3.1-405B	52.7±0.9

(c) SelfDecl (%)

Model	**SelfDecl (%)**
GPT-3.5-Turbo	1.5
GPT-4o	45.7
Claude-Sonnet-4	29.2
DeepSeek-v3	29.0
Llama-4-Scout	24.9
Llama-4-Maverick	21.2
Llama-3.1-8B	0.1
Llama-3.1-70B	35.5
Llama-3.1-405B	31.5

(d) GuessSucc (%)

Model	**GuessSucc (%)**
GPT-3.5-Turbo	8.2
GPT-4o	90.6
Claude-Sonnet-4	91.2
DeepSeek-v3	71.0
Llama-4-Scout	52.2
Llama-4-Maverick	81.8
Llama-3.1-8B	0.0
Llama-3.1-70B	32.6
Llama-3.1-405B	1.0

Table 12: Team dynamics across models: coordination, trust, and convergence patterns. Panel (a) reports coordination and convergence metrics: *Coord. Rate* (any explicit coordination attempt), *Vote Coord.* (pairwise vote-alignment rate ±95% CI), *Implicit Coord.* (spontaneous, unsignaled coordination), *Collab. Success* (planned joint strategy completes), *Strategy Align.* (share of rounds with consistent plan), *Linguistic Mirror* (function-word/style mirroring index), and *Vocab Conv.* (Jaccard convergence). Panel (b) reports trust and robustness: *Trust Form.* (share of games with explicit trust), *Betrayal Rate* (fraction of alliances that break), *Trust Recovery* (post-betrayal recovery rate), *Vulnerability Index* (Trust/Betrayal; lower is better), *Resilience Factor* (Recovery/Betrayal), *Coord. Gap* (implicit − explicit vote coordination), and *Alignment Gap* (implicit − explicit strategy alignment). Unless noted, higher is better (exceptions: Betrayal and Vulnerability).

(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%