NEIGHBORS AS TOP SUSPECTS: DIAGNOSING RANKING SHORTFALLS IN LLM ADVERSARY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents increasingly make decisions from other agents' claims. Even when evaluators observe the received messages, a scalar outcome does not reveal where the LLM's final ranking differs from a reference derived from that stream. We record the messages delivered to one LLM seat in a five-round Mafia game with fixed-policy surroundings and derive an offline accusation-minus-defense shadow ranking. We separate non-neighbor placement (the full-ranking slots occupied by players not heard directly) from ordering within those slots. Across four models and eleven configurations, model mean reciprocal rank (MRR) is below the shadow in 34 of 44 cells. On complete raw rankings with a non-neighbor adversary, placement's signed contribution accounts for 76–100% of the mean shadow-minus-model reciprocal-rank gap and remains the larger contribution in connected configurations. A separate delivered-accusation analysis shows that the adversary can rank highly among non-neighbors yet below neighbors overall. On the same 224 out-of-view GPT-5.6 games across seven relay configurations, with complete raw rankings in every arm, premise correction accompanies higher model MRR and improved mean within-group adversary rank. The ordering contribution decreases without a detectable placement-contribution change; structure disclosure shows no detectable change in either contribution. Corrected-premise profiles vary across models, with unequal complete-output coverage. These post-hoc, reference-relative diagnostics describe one scripted environment without identifying mechanisms, demonstrating repair, or establishing transfer.

1 INTRODUCTION

Large language models (LLMs) are used as agents in communicating groups (Zhu et al., 2026). An agent's decisions may depend on incomplete or false claims, including statements about participants it never hears directly. Evaluating such systems therefore requires examining how well individual agents decide from received information, not only whether the group succeeds.

A group-level outcome alone does not distinguish whether task-relevant information was expressed, reached a particular agent, or was reflected in that agent's final decision. We organize these evaluation questions as *surfacing* $\rightarrow$ *delivery* $\rightarrow$ *decision use*, referring to observable messages and outputs rather than internal cognitive stages. The distinction matters for choosing what to change and for interpreting what an intervention changes.

HiddenBench reports gains from appending agents' facts to their first-round messages, supporting a surfacing bottleneck in that setting (Li et al., 2026). Networked-agent research controls exposure and influence (Liu et al., 2026). The original SocialMaze task evaluates inference from one player's view of simulated hidden-role logs (Xu et al., 2025), while ERPBench's Solo condition evaluates one LLM against deterministic rule-based opponents (Zhang et al., 2026). Individual-view evaluation and fixed surrounding policies are thus established tools (§5).

In adversary detection, the evaluator can observe the messages actually delivered to an LLM, apply a specified rule offline to that same received stream, and compare the resulting reference ranking with the LLM's final ranking. Comparing only scalar outcomes, such as the adversary's rank, leaves the structure of any discrepancy unresolved. Consider a network where an agent hears only its direct neighbors: a non-neighbor adversary can rank poorly among non-neighbors, or non-neighbors can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

occupy lower positions as a group. We ask: when the LLM trails this same-stream reference, where in the ranking does the shortfall appear, and where do intervention-associated changes appear?

We study this question in a five-round Mafia game, where the evaluator knows the adversary's identity and one LLM seat must infer it from claims that can be false. Other seats follow fixed policies, but the LLM speaks, so changing the occupant or prompt can change the conversation. We construct the reference with an offline *shadow tally* that counts accusations minus defenses in messages delivered during the model's own game. The shadow never speaks and scores the two statement rounds preceding final judgment, whereas the LLM can also access earlier history (§2). The tally supplies a specified reference, not an upper bound.

For games with a non-neighbor adversary, we separate *placement*, the full-ranking slots occupied by non-neighbors, from *ordering*, which non-neighbor occupies each slot. The adversary can rank first among non-neighbors yet fourth overall if three neighbors occupy the top three positions. Reordering the non-neighbors cannot make the adversary first overall while their earliest slot remains fourth. We exchange placement and ordering between the model and shadow rankings, averaging contributions over both exchange orders to decompose their reciprocal-rank gap (§3.3). This analysis uses complete raw rankings in candidate appearance order and does not require a delivered accusation naming the adversary.

Across four models and eleven network configurations, model mean reciprocal rank (MRR) is below the shadow in 34 of 44 model–configuration cells (§3.1). On complete raw rankings with a non-neighbor adversary, placement contributes 76–100% of the mean shadow-minus-model reciprocal-rank gap across models (§3.3). These percentages are signed shares of mean gaps, not fractions of games, and placement remains the larger contribution in connected configurations. A separate analysis of games in which a neighbor accused a non-neighbor adversary before judgment illustrates the pattern: when non-neighbor confidence scores are nonuniform, the adversary can rank highly within that group yet below neighbors overall (§3.2).

The diagnostic also localizes changes under two prompt interventions. On the same 224 out-of-view GPT-5.6 games across seven relay configurations, with complete raw rankings in every arm, premise correction—revising the prompt's description of honest seats—accompanies higher model MRR and improved mean within-group adversary rank. The ordering contribution decreases without a detectable change in the placement contribution. Structure disclosure, which states that the adversary need not be a neighbor, shows no detectable change in either contribution (§3.4). Corrected-premise component profiles are heterogeneous across models, with unequal complete-output coverage. Our contribution is an empirical diagnostic that separates placement from ordering to localize ranking shortfalls and intervention-associated changes against each arm's shadow on its own received stream. These post-hoc findings describe output rankings in one scripted environment, not an identified mechanism, a demonstrated repair, or established transfer.

2 Setup and Methods

Task and observations. Ten to twenty players sit on a communication graph: isolated pairs, ring, small-world, cycle-heavy, or hub-and-spoke, with eleven topology–size configurations (Figure C.1). Each round every player emits one accusation or defense naming a target. Only its neighbors receive it. Rounds are synchronous: statements generated in round t become available for the next judgment. The evaluated seat is the only language model; it speaks and ranks all other players by suspicion. We refer to the player at the top of that reported ranking as its leading suspect. We score the known adversary's rank rather than the group's win rate, separating identification from robustness to opponents' errors (Wang et al., 2026). The task starts from the original SocialMaze hidden-role setting (Xu et al., 2025); Appendix A lists the adaptations.

Fixed policies, changing messages. The adversary picks a neighbor uniformly and accuses or defends on a coin flip, without consulting the ground truth. Each honest seat opens by choosing another player uniformly and making a truthful claim about it. From round two onward it consults only received statements, aggregates accusations minus defenses, and accuses its leading suspect. Three fixed variants differ in weighting (Appendix A). Their later claims can be false. Policies and seeded random choices are fixed, but changing the evaluated seat can change the subsequent conversation.

Active replacement (co-run)	**Same-stream comparison (shadow)**
Model plays one game; rule plays another with the same seed.	Model plays the game; rule scores its received messages offline.
Both occupants speak. Later messages can differ.	The rule never speaks. Its scores cannot change the messages.
Compare the adversary's rank across the two games.	Compare the adversary's rank on the model's observed stream.

Figure 1: Two comparisons using the same tally rule. Active replacement includes the occupant's effects on later messages and the rule's truthful opening. Shadow replay removes the rule's feedback; it uses a specified message window, while the model has access to earlier history.

The results are not conditioned on games known to be solvable: in 245 of 539 seeded games no other seat accuses the adversary in the opening round, although later messages may provide evidence.

Varying the graph changes who can speak directly to the seat and which other players can only be named in received statements. The resulting in-view and out-of-view groups can also differ in the evidence available, so the graph comparisons describe these conditions rather than isolate a causal effect of contact.

Two comparisons. Figure 1 distinguishes them. In the *co-run*, the plain tally rule plays the seat in a separate game under the same seed. It receives the scripted truthful opening used by honest rules; the model does not. In the *shadow* comparison, the rule ranks candidates offline using messages delivered during the model's game and never speaks. At the final judgment it scores the two preceding statement rounds. The model has access to earlier conversation messages as well. Thus the comparison uses the model's received stream, not identical prompts or history windows. Extending the shadow window preserves a deficit (Appendix A).

Models, seeds and scores. Each model–configuration cell has 49 games, seeds 100 to 4900 in hundreds; seed 0 is excluded. The campaign contains 539 games per model and 2,156 seat-level observations. We report models in this order: Llama-3.1-70B, GPT-5.6, Gemini 3.1 Pro and DeepSeek V4 (our label V4-Flash-0731). A seed fixes graph draws, roles and scripted random choices. The registered primary metric is mean reciprocal rank (MRR) of the adversary in the final confidence-sorted ranking, with stable order for ties. Higher is better. We also report top-1, which ignores how far below first place an incorrect candidate lies. Uniform-ranking chance MRR depends on group size, $H_{N-1}/(N-1)$; comparisons are within configurations where needed. Appendix A gives the warning arms' separate same-seed, net-five top-1 criterion and its calibration.

Prompt and environment conventions. The original prompt names only the model's neighbors. It describes an adversary whose claims can be true or false, without specifying the random policy, and says investigators always speak truthfully, although honest seats satisfy that description only in the opening round. These two lines are quoted in Appendix A. After the opening, about 80% of honest accusations name an innocent player. In addition, rules break a tied tally toward the lowest-numbered candidate, often the evaluated seat, seat 1. Honest rules never defend after the opening, so later defenses identify the adversary's speaker role. These conventions can affect both model and rule performance. Their prompt, roster, emission and tie interventions are reported in §4.

Output processing and analysis status. Models name 77.3% to 100.0% of the roster in the final round. The parser adds missing candidates at confidence 5 in seat-number order before sorting. A sensitivity check using output order and mean ranks for unnamed candidates changes mean MRR by -0.0065 to $+0.0007$ (B.15). The placement analysis instead uses complete raw rankings and states its order explicitly. Appendix B, B.20 distinguishes the original campaign prediction, staged behavioral tests and post-hoc diagnostics; the latter include the audibility contrast and placement decomposition.

3 RESULTS

The comparisons first establish a gap against a simple rule. We then examine the model's leading suspect when an accusation names the adversary, before quantifying group positions and within-

group order on a separate complete-output sample. The audibility contrast describes the broader setting in which these output differences occur.

3.1 Comparing the model with the tally rule

The same received stream. The shadow tally reads the statements delivered during the model's own game and never takes part in it. In reciprocal rank the model is below it in 34 of 44 cells and above it in 10, with no ties; three of the positive cells are non-hub configurations, with gaps no larger than 0.041 (B.19). This is the comparison the rest of the paper uses, because it holds the received messages fixed. It does not hold the whole input state fixed: the model has access to earlier messages as well as the rule's two-round window, and extending the shadow window preserves the deficit (Appendix A).

Replacing the occupant. A second comparison lets the tally rule play the same seeded seat. The model gives the adversary a lower mean reciprocal rank in 37 of 44 cells; every model trails the rule in every non-hub configuration, and all seven cells above it are hub seats (Figure C.6 and Table C.1, Appendix C). Excluding the two isolated-pairs configurations, where no relay is possible, leaves 29 of 36 cells below the rule, and paired bootstrap intervals exclude zero in 24 of the 37 negative cells and one of the seven positive cells, with counts that vary by bootstrap draw (B.19). Here both occupants' messages change the subsequent stream and the rule receives the scripted truthful opening that the model does not, so this measures a whole interaction rather than one judgment on a fixed stream.

The tally measures what one simple aggregation can achieve in this environment. It is not an upper bound, and its fit to the scripted agents need not be the same across audibility groups. The following analyses locate a gap relative to that comparator. §4.2 reports which gap measures were tested under other rules and environment changes.

3.2 When an accusation names the adversary

A separate analysis selects games where the adversary is a non-neighbor (*out of view*) and a neighbor accuses it before judgment. This tests an observed ranking pattern where that claim was delivered; it does not require complete raw rankings. Across all such out-of-view games, the model puts the adversary first in **7% to 32%**, against **52% to 79%** for the shadow rule, with ties counted by expectation.

The within-group analysis separates games by whether the model's scores distinguish any non-neighbors. In 300 of the 411 they do, and there the mean normalized rank within the non-neighbor group is **0.933** for the model and **0.932** for the rule, on a scale where 1 is highest and 0.5 is the expectation under a uniform random ordering (Figure B.1, left). These similar means do not establish equivalence, and they describe that subset alone. In the other **111** games the model gives every non-neighbor one shared score, which the mean-rank convention scores at exactly 0.5; the rule is flat in **12** of them. Over all 411 games the same convention gives **0.816** for the model against **0.931** for the rule. Where the model separates the block it ranks the adversary about as high within it as the rule does; in 111 of all 411 eligible games it does not separate the block at all.

Among the 300 nonuniform-score games in which the adversary is not first overall, every seat above it is a direct neighbor in 69% to 92% of cases (B.30). A candidate can therefore rank high within the non-neighbor group while direct neighbors fill the higher positions.

The selection is not only of accusations that the seat's own later messages produced. Restricting it to games in which an honest neighbor names the adversary in the opening round — a target drawn by that seat's seeded generator before any message arrives, so the same 55 configuration–seed pairs appear in every model — the model's top-1 remains 0.375 to 0.626 below the shadow rule's, with all four intervals excluding zero (B.29).

3.3 How much of the gap lies in group positions?

To quantify the positions and ordering separately, we use a different population: out-of-view games whose raw output lists every candidate. Using complete outputs avoids treating positions added by the parser as positions chosen by the model. The decomposition does not condition on a delivered

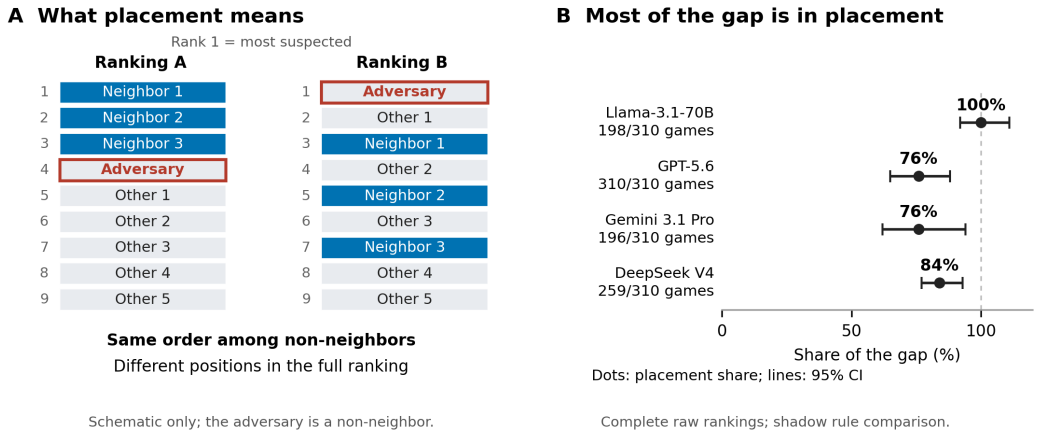


Figure 2: Where non-neighbors sit accounts for most of the gap. **A:** Illustrative rankings, not observed game data. Rank 1 is most suspected. Blue seats are direct neighbors, whose statements the evaluated seat hears; gray seats are non-neighbors. The adversary is first among non-neighbors in both lists, but fourth versus first overall. **B:** Placement's share of the mean shadow-minus-model reciprocal-rank gap, averaging both component-swap orders. Points and 95% intervals from resampling games use out-of-view games with complete raw rankings; labels give retained games out of 310 per model. These post-hoc estimates describe output rankings. Contributions are signed: a placement share can exceed 100% when ordering favors the model.

accusation. Here we use the order in which candidates appear in the raw output, rather than the confidence-sorted ranks used above.

We separate the slots occupied by non-neighbors (*placement*) from their order within those slots, and exchange these two components between the model and shadow rule on the same game; Figure 2 illustrates the distinction. Averaging both exchange orders assigns **76% to 100%** of their mean reciprocal-rank gap to placement. This is a post-hoc decomposition. The percentage is a signed share of the selected sample's mean gap, not a fraction of games or a decomposition of the audibility contrast. Per-model intervals and coverage are shown in the figure.

Using confidence-sorted rankings on the same complete-output games also leaves placement as the larger contribution for every model (B.25; point estimates only). Excluding isolated pairs leaves placement shares of 83% to 95% on complete outputs in the seven connected configurations. On the full complete-output samples, scoring the two contributions by magnitude rather than by sign still leaves placement the larger in every model, at 62% to 78% of the total absolute contribution (B.25). These checks describe output rankings.

3.4 Localizing ranking changes under intervention

We apply the decomposition to two existing GPT-5.6 prompt arms on the same 224 out-of-view games across seven relay configurations; raw rankings are complete in all three arms, including the baseline, so no game is dropped. Premise correction replaces the two lines quoted in Appendix A; structure disclosure states that the adversary need not be a neighbor. Let $D = S - M = P + O$ be the shadow-minus-model mean reciprocal-rank gap under this section's rank definitions, and let Δ denote intervention minus baseline. The 95% paired bootstrap intervals resample seed clusters.

intervention, 224 games	ΔD	ΔP	ΔO
corrected premise	-0.026 [-0.054, $+0.003$]	$+0.002$ [-0.019, $+0.023$]	-0.028 [-0.047, -0.011]
structure disclosure	-0.006 [-0.028, $+0.017$]	-0.005 [-0.026, $+0.016$]	-0.001 [-0.010, $+0.008$]

Both ΔD intervals include zero. Under premise correction, model mean reciprocal rank increases by 0.038 [0.010, 0.067], and the adversary's mean within-block rank improves by 0.52 positions [0.20, 0.89]. The ordering component of the gap decreases; no change is detected in the placement component, the mean first non-neighbor slot, or the neighbor-first frequency. Each shadow reads

its own arm's stream, and the shadow mean also rises, by 0.012 [−0.006, +0.031], so a reference-relative gap can move differently from the model's own score. The corrected gap is 0.092 against 0.118 at baseline and remains placement-dominated at the point estimates, with ordering slightly favoring the model. Structure disclosure shows no detectable change in either component. These arm-specific profiles do not establish intervention superiority.

DeepSeek likewise shows a lower ordering component without a detectable placement change; Llama shows no clear component change. Gemini instead shows $\Delta P = -0.161$ $[-0.234, -0.097]$, with its mean gap changing from 0.150 to -0.004. Its 49 of 107 common-complete games in three retained relay cells form the most heavily selected sample (E.6). These post-hoc paired differences are descriptive, not causal effects or evidence of a repair.

3.5 A LARGER RELATIVE GAP OUTSIDE THE NEIGHBORHOOD

Split games within each topology–size configuration by whether the adversary is a direct neighbor; hub configurations have no out-of-view games and are excluded. In a post-hoc analysis of the nine remaining configurations, the model's deficit against the shadow tally is **0.174** reciprocal rank larger out of view (95% CI 0.093 to 0.259, resampling configurations; Figure C.3), positive in all nine configurations and in 32 of 36 model–configuration cells. Removing the isolated pairs, where nothing is relayed, raises it to 0.215. B.14 gives the separate in-view and out-of-view levels, three alternative resampling units and weightings that keep the interval above zero, and the tie treatment that shifts the in-view level; the in-view interval includes zero and does not establish equal performance, and the model also trails the co-run rule in view in the ring family (B.10). These are conditional comparisons: topology and size are fixed within a configuration, while realized graphs and roles can vary across seeds. The claim is a larger relative gap outside the neighborhood, not a universal absence of in-view errors.

A related behavioral test. The separately specified N5 search measures excess top-1 selection of neighbors over the shadow rule on its own eligible paired population, not the placement share; screening criteria were fixed before the named outcome statistics were unsealed, and DeepSeek confirmation gave an excess positive in every subgroup (Figure C.5; timing and scope in B.20). The pooled test is confirmatory under that staged protocol, while its subgroup profile is descriptive and does not confirm the post-hoc decomposition. On all out-of-view games, neighbor occupancy of the top slots exceeds the co-run tally's by 14 to 37 points (Figure B.1, right; B.15).

4 ALTERNATIVE EXPLANATIONS AND INTERVENTIONS

The interventions address explanations for particular gaps and choices: roster changes test candidate visibility, premise changes the description of other players, emission and tie changes the environment. Most measure performance gaps, audibility contrasts or neighbor selection rather than the placement share of §3.3.

4.1 CANDIDATE VISIBILITY AND THE PROMPT PREMISE

Full roster. A roster-only arm lists every player while retaining the original premise, on the same 49 seeds in two configurations for DeepSeek V4 and Llama-3.1-70B. Both mention more seats but still place a neighbor first in 0.933 and 0.987 of out-of-view games; DeepSeek's registered rule withholds a directional verdict because its paired interval includes zero, and Llama misses its registered manipulation-check threshold (Appendix E gives both paired changes).

Joint roster and premise correction. A bundled roster-plus-premise arm retains an audibility contrast, and an exploratory decomposition keeps placement as the larger contribution in DeepSeek and two configurations. On those games the two components move in opposite directions across the interface arms, which the total gap alone does not show (Appendix E) — a second instance of the comparison in §3.4. This is separate from the arm's registered audibility decision and leaves room for an interface contribution.

The premise and accuser choice. A related choice is whether the model selects a neighbor that accuses it. The original premise permits a simple deduction: an accusation of oneself is false, investigators never make false claims, so the accuser must be the adversary. In eligible configurations,

games ending on a neighbor that accused the seat before judgment are wrong 75% to 81% of the time in the three models held out from hypothesis formation, against 30% to 52% in the remaining games; on accused games with a unique shadow top suspect, the model selects one of its accusers 17 to 42 percentage points more often than the shadow rule (per-model values and intervals in B.8 and B.31). The rates reflect both which accusers are innocent and which ones the model selects; these populations differ, and the reasoning-text matches reported in B.11 do not identify the model's internal inference (Turpin et al., 2023). Raw self-rankings, which the parser removes from scored outputs, are reported separately in B.23.

Correcting the description. Replacing the two prompt lines gives an accuser-targeting decrease of 24.5 percentage points [12.2, 36.7] in GPT-5.6 isolated pairs and a point estimate of 5.3 points [−5.3, 15.8] at the hub, on games accused before judgment in both arms. The registered prediction was a larger decrease in premise-citing models than in Llama; in temperature-matched comparisons five of six point estimates follow that direction, only the two Gemini intervals exclude zero, and the DeepSeek hub points the other way (Appendix E). Conditioning on accusation, which can change across arms, limits causal interpretation. B.22 reports the arm's coverage over 35 cells, its point-estimate sign counts against both comparators, the matched top-1 comparison and the temperature-matched substitutions; warnings have mixed effects of their own (Appendix A).

4.2 Environment and comparator dependence

The tally's fit to the environment. Honest seats aggregate accusations, so the tally is a comparison suited to this setting. Weakening truthful openings reduces the co-run gap in the tested ring cell and leaves both tested relay intervals covering zero; an exploratory change to the surrounding agents' aggregation rule retains a positive audibility contrast in two Llama configurations, though both environments still accumulate accusations (Appendix E). This tests dependence on a shared weighting rule, not transfer beyond that family of environments. The audibility contrast has intervals excluding zero under top-1 and log-odds scoring and under five of six tested decision rules; the exception is a speaker-based rule, whose small positive contrast includes zero (B.26–B.27). A rule using post-opening defenses performs much better in hubs (B.5, B.37), and an oracle-informed reference using privileged knowledge of the emission rules moves the larger top-1 shortfall to the in-view half instead (B.16). Distance, seat-swap, complete-graph, silent-adversary, honest-defense, LLM-neighbor and reasoning-mode analyses each remain consistent with the tested gap within their own limits, several on a single cell or a different estimand (B.13, B.17–B.18, B.32, B.35–B.36; Appendix E, E.4). Thus the location and size of a relative gap depend on the reference; these are not comparator-free measures of information use, and they concern the tested gaps and choices, not the robustness of the placement percentage under every manipulation.

Replacing the adversary's policy. An honest-rule replacement leaves GPT-5.6 below the co-run tally in all three tested configurations, and in hub N15 the gap changes from $+0.094$ to -0.117 with the replacement-arm interval excluding zero (Appendix E, E.1). The target remains the original designated adversary seat, so the arm no longer measures detection of an actively dishonest player; it limits an account based only on the adversary's false statements.

4.3 The deterministic tie convention

Randomizing rule ties changes the accusations reaching the evaluated seat and the shadow tally's performance: it raises expected MRR at the hub and lowers it in isolated pairs. Out-of-view deficits remain in the tested relay configurations under both policies, while intervals for their differences do not establish equivalence (Appendix G). This limits an explanation based solely on the original tie convention; it does not establish an unchanged effect or a placement share under randomized ties, and the accusation-conditioned associations in B.21 and B.33 are not exogenous interventions on being accused.

5 Related Work

Language-model agents are deployed autonomously and composed into communicating groups (Staufer et al., 2026; Zhu et al., 2026; Ehtesham et al., 2025). The studies below provide context for the local ranking comparison.

Collective information failures. HiddenBench reports a shortfall across fifteen frontier models. In a separate eighteen-task intervention on two models, appending every agent's facts to its first-round message raises accuracy to 92.6% and 98.2%, supporting a surfacing bottleneck in that setting (Li et al., 2026). Network structure also affects conformity (Han et al., 2026); agents may publicly endorse norms they privately reject, with rates of 64–94% across eight models in Yashwanth (2026). Other work studies peer pressure, social influence and collective misremembering (Song et al., 2025; Ko et al., 2026; Xu et al., 2026). Absence of an injected adversary does not make all communicated claims true. These studies motivate separating access to evidence from its use, rather than assuming that collective errors identify either stage.

Adversaries and interventions. WOLF studies deception in Werewolf and includes deterministic subsystem controls (Agarwal et al., 2025); SocialGrid evaluates embodied deception detection without a discussion phase (Shindo et al., 2026). GAMBIT reports an imposter-class F1 of 0.505 for a Gemini-based zero-shot judge on its first two adversary generations. It also tests suspicion prompting: it harms play without an imposter but improves play with one, without restoring the benign collective's performance (Le Mercier et al., 2026). Awareness and self-correction interventions elsewhere are partial and task-dependent (Echterhoff et al., 2024; Huang et al., 2025; He & Liu, 2025). HiddenBench's asymmetry warning also leaves a gap (Li et al., 2026). These findings do not rule out prompting as an intervention; our own prompt arms test specified changes in this environment.

Controls close to ours. The original SocialMaze task evaluates inference from one player's view of simulated hidden-role logs; the model does not generate the players' statements (Xu et al., 2025, App. B, items 1–2). MafiaScope adds belief probes and counterfactual utterance replay in an interacting LLM game (Karpov, 2026, §6). Liu et al. (2026) control exposure and influence and test capability placement in graph coloring (App. 13.6). Beyond communication tasks, ERPBench evaluates one LLM against five deterministic rule-based opponents in its Solo condition (Zhang et al., 2026, §3.2). Thus fixed surrounding policies and local interventions are established tools. We use them to locate an empirical ranking gap: the positions occupied by non-neighbors can differ from the rule even when the adversary ranks high within that group. The complete-output decomposition and the delivered-accusation analysis provide distinct evidence for that account.

Methodological precedents. Our accusation-minus-defense tally is inspired by naive network learning (Golub & Jackson, 2010), whose DeGroot updates average neighbors' beliefs; the discrete tally and accusation dynamics here do not instantiate that linear update or inherit its convergence guarantees. Scripted confederates control social input in human experiments (Asch, 1956), while network experiments also mix humans with programmed bots to study collective coordination (Shirado & Christakis, 2017). The hidden-profile paradigm motivates the distinction between distributed information and successful pooling (Stasser & Titus, 1985); our hidden-adversary game additionally includes false statements and does not guarantee sufficient evidence in every instance.

6 Discussion and Limitations

The observed problem is how a named suspect is placed relative to the speakers: in the selected delivered-accusation sample a high adversary rank among non-neighbors can coexist with direct neighbors above it, and on a separate complete-output sample group positions contribute most of the mean gap against the shadow tally.

What the decomposition locates. Separating placement from ordering gives the intervention result a more specific interpretation than an aggregate improvement alone. In GPT-5.6's 224-game comparison, higher model reciprocal rank under premise correction accompanies better within-group ranking, without a detectable change in the placement contribution or the direct placement summaries. The remaining reference gap is placement-dominated at the point estimates. The decomposition therefore locates the observed improvement in these rankings, not a gain that aggregate scoring failed to detect. The cross-model extension bounds that reading: Gemini shows a placement reduction on the most heavily selected sample rather than the GPT-5.6 profile. These heterogeneous component profiles are descriptive results, not identified mechanisms, and each arm's shadow reads its own stream, so localization remains reference-relative. Placement dominance does not justify indiscriminately promoting non-neighbors, which could hurt when the adversary is a neighbor; likewise the models' apparent co-run hub advantage is tie-sensitive, so it does not justify a general rec-

ommendation to increase connectivity. The existing interventions constrain particular explanations and have different endpoints; their results cannot all be transferred to the placement decomposition.

Scope and selection. The one-seat design measures neither group success nor performance against a strategic adversary. All results come from one scripted environment and four tested models. Audibility varies with graph distance, available claims and behavioral cues; its contrast is not an isolated causal effect of direct access. Placement conditions on a non-neighbor adversary and complete raw output, and its share is a signed contribution to a mean gap, not the fraction of games with that failure. Complete-output coverage differs across models, from every out-of-view game for GPT-5.6 to under two thirds for Gemini and Llama, so the four are not read on one population; the intervention comparison of §3.4 is confined to the model whose coverage is complete. The response to premise correction is heterogeneous rather than uniform across models, and the row that differs most rests on the most heavily selected sample (Appendix E). The roster arm is not a one-sided addition but a change in how the candidate list and the seat's visibility are presented, and the bundled arm adds the premise correction on top of that. The decomposition and the paired arm differences are descriptive accounting, not identification of a mechanism. B.30's additional selection on delivered accusations and nonuniform scores has a separate scope. Accuser-conditioned rates can also select different games across interventions. An exogenous-accusation arm has not been run.

Uncertainty and transfer. The audibility intervals resample configurations, while placement intervals resample eligible games; they have different scopes. Analyses on the same campaign are not independent replications. Of the five staged behavioral rows, only N5 carried a surviving verdict; three others lacked enough events, and the fourth was not a clean measure in this environment (Appendices A and B). Transfer to other tasks, models or information-generation rules remains unmeasured beyond the stated checks. The human-bias literature supplies measurement precedents, not evidence of shared mechanisms or that models stand in for people (Phillips & Edwards, 1966; Edwards, 1968; Knipper et al., 2025; Malberg et al., 2024; Ye et al., 2025; Wu et al., 2025; Anthis et al., 2025).

Serving and reproducibility. No API provider here exposes a dated checkpoint. Serving manifests are recorded from 2026-09-04 onward, after the main campaign's forty-four cells. Its results therefore describe a measured window rather than checkpoint-level reproducibility. Temperature-zero serving is also nondeterministic: identical calls can return different completions and change individual game outcomes. The campaign aggregates 49 seeds per cell; retained records and the registry support computational replay, with re-execution limits reported in Appendix A.

7 Conclusion

On complete raw rankings with a non-neighbor adversary, placement accounts for most of the mean shadow-minus-model gap in this controlled task. The decomposition also localizes observed changes: on 224 out-of-view GPT-5.6 games, higher model reciprocal rank under premise correction accompanies improved within-group ordering without a detectable change in the placement contribution. Cross-model profiles are heterogeneous, with unequal complete-output coverage. The contribution is a demonstrated diagnostic of ranking shortfalls and intervention-associated changes against each arm's same-stream reference. These post-hoc findings describe output rankings in one scripted environment, not an identified mechanism, a demonstrated repair or established transfer.

Reproducibility Statement

The reproduction registry links the reported empirical results to the analysis scripts and retained records used to produce them. Analysis status and registration provenance are documented in Appendix B.20 and the supplementary registry; re-execution limits are described in §6 and Appendix A. The analysis scripts, the per-game records they read, the preregistration documents and analysis notes are submitted as anonymized supplementary material.

AI USE STATEMENT

The authors formulated the research questions and hypotheses, directed the study, made the final scientific decisions, and take responsibility for the final text, analyses, and conclusions. AI tools provided feedback on study design and interpretation and assisted with implementing selected parts of the experiment and analysis code, debugging, infrastructure support, and experiment execution. They also assisted with numerical checking and result summarization, literature retrieval, and drafting, checking, and revision of the manuscript and preregistration documents. AI tools performed qualitative coding of reasoning texts; the labels were AI-generated rather than produced by human annotators and were subsequently used in author-directed quantitative analyses. Appendix B.4 reports the coding procedures, cross-session agreement, and limitations.

REFERENCES

Mrinal Agarwal, Saad Rana, Theo Sundoro, Hermela Berhe, Spencer Kim, Vasu Sharma, Sean O'Brien, and Kevin Zhu. WOLF: Werewolf-based Observations for LLM Deception and Falsehoods. 2025. arXiv:2512.09187.

Vernon L. Allen and John M. Levine. Consensus and conformity. *Journal of Experimental Social Psychology*, 5(4):389–399, 1969.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. LLM Social Simulations Are a Promising Research Method. 2025. arXiv:2504.02234.

Solomon E. Asch. Opinions and Social Pressure. *Scientific American*, 193(5):31–35, 1955.

Solomon E. Asch. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956.

Rod Bond. Group Size and Conformity. *Group Processes & Intergroup Relations*, 8(4):331–354, 2005.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive Bias in Decision-Making with LLMs. 2024. arXiv:2403.00811.

Murray Glanzer and Anita R. Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 5(4):351–360, 1966.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. Conformity Dynamics in LLM Multi-Agent Systems: The Roles of Topology and Self-Social Weighting. 2026. arXiv:2601.05606.

Jiangen He and Jiqun Liu. Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks. 2025. arXiv:2502.14219.

Robin M Hogarth and Hillel J Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, 1992.

Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. Understanding the Anchoring Effect of LLM with Synthetic Data: Existence, Mechanism, and Potential Mitigations. 2025. arXiv:2505.15392.

Ward Edwards. Conservatism in human information processing. In Benjamin Kleinmuntz (ed.), *Formal Representation of Human Judgment*, pp. 17–52. Wiley, New York, 1968.

Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP). 2025. arXiv:2505.02279.

Ilia Karpov. MafiaScope: Non-Invasive, Time-Resolved Belief Probing for LLM Agents in Social Deduction Games. 2026. arXiv:2607.10645.

Harold H. Kelley. The processes of causal attribution. *American Psychologist*, 28(2):107–128, 1973.

R. Alexander Knipper, Charles S. Knipper, Kaiqi Zhang, Valerie Sims, Clint Bowers, and Santu Karmaker. The Bias is in the Details: An Assessment of Cognitive Bias in LLMs. 2025. arXiv:2509.22856.

Changgeon Ko, Jisu Shin, Hoyun Song, Huije Lee, Eui Jun Hwang, and Jong C. Park. Social Dynamics as Critical Vulnerabilities that Undermine Objective Decision-Making in LLM Collectives. 2026. arXiv:2604.06091.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic Failures in Collective Reasoning under Distributed Information in Multi-Agent LLMs. *ICML*, 2026. arXiv:2505.11556.

Kaixuan Liu, Guojun Xiong, Weinan Zhang, and Shengpu Tang. Social Networks of LLM Agents. 2026. arXiv:2607.03695.

Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A Comprehensive Evaluation of Cognitive Biases in LLMs. 2024. arXiv:2410.15413.

Alexandre Le Mercier, Chris Develder, and Thomas Demeester. GAMBIT: A Three-Mode Benchmark for Adversarial Robustness in Multi-Agent LLM Collectives. 2026. arXiv:2605.09027.

Norman Miller and Donald T. Campbell. Recency and primacy in persuasion as a function of the timing of speeches and measurements. *The Journal of Abnormal and Social Psychology*, 59(1):1–9, 1959.

Markus M. Möbius, Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, 68(11):7793–7817, 2022.

Bennet B. Murdock, Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488, 1962.

Lawrence D. Phillips and Ward Edwards. Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3):346–354, 1966.

Hikaru Shindo, Hanzhao Lin, Lukas Helff, Patrick Schramowski, and Kristian Kersting. SocialGrid: A Benchmark for Planning and Social Reasoning in Embodied Multi-Agent Systems. 2026. arXiv:2604.16022.

Hirokazu Shirado and Nicholas A. Christakis. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654):370–374, 2017.

Maojia Song, Tej Deep Pala, Ruiwen Zhou, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. LLMs Can't Handle Peer Pressure: Crumbling under Multi-Agent Social Interactions. 2025. arXiv:2508.18321.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48:1467–1478, 1985.

Leon Staufer, Kevin Feng, Kevin Wei, Luke Bailey, Yawen Duan, Mick Yang, A. Pinar Ozisik, Stephen Casper, and Noam Kolt. The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. 2026. arXiv:2602.17753.

Charles S. Taber and Milton Lodge. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3):755–769, 2006.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. 2023. arXiv:2305.04388.

Kevin Wang, Anna Thöni, Benjamin Kempinski, Bobby Cheng, Jianzhu Yao, Benjamin Finch, Leon Guertler, Viraj Nadkarni, Yihan Jiang, Aliaksei Korshuk, Alexander Buyantuev, Ilya Makarov, Siyuan Wu, Yu-Chi Cheng, Yan-Ru Ju, Ti-Rong Wu, I-Hsuan Chu, Yu-Yu Yang, I-Chen Wu, Yitian Huang, Qinlu Cao, Yiheng Sun, Yuhong Dai, Hongkun Yao, Jingxuan Fu, Jiwei Zhang, Hao Liao, Mossimo Ebeling, Govind Arun, Sadhvik Bathini, Mihir S Arya, Avinash Anish, Aditya Ranjan, Kirtana Sunil Phatnani, Paval KS, Vrushali Mehta, Aravind S, Nikhil Arora, Tanya Upadhyay, Amol Bandagale, Yuan Lu, ChunEn Hsiao, YuTing Lin, Arvin Chung, Jerry John Thomas, Mathieu Laurière, Leshem Choshen, Yoram Bachrach, Pramod Viswanath, Maria Polukarov, Cheston Tan, Tal Kachman, and Atlas Wang. MINDGAMES: A Live Arena for Evaluating Social and Strategic Reasoning in Multi-Agent LLMs. 2026. arXiv:2605.29512.

Kimberlee Weaver, Stephen M. Garcia, Norbert Schwarz, and Dale T. Miller. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, 92(5):821–833, 2007.

David A. Wilder. Perception of groups, size of opposition, and social influence. *Journal of Experimental Social Psychology*, 13(3):253–268, 1977.

Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. LLM-Based Social Simulations Require a Boundary. 2025. arXiv:2506.19806.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. SocialMaze: A Benchmark for Evaluating Social Reasoning in Large Language Models. 2025. arXiv:2505.23713v1.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When Agents "Misremember" Collectively: Exploring the Mandela Effect in LLM-based Multi-Agent Systems. 2026. arXiv:2602.00428.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement. 2025. arXiv:2505.08245.

Sami R. Yousif, Rosie Aboody, and Frank C. Keil. The Illusion of Consensus: A Failure to Distinguish Between True and False Consensus. *Psychological Science*, 30(8):1195–1204, 2019.

Yashwanth YS. Everyone Conforms, No One Believes: Pluralistic Ignorance in LLM Agent Populations. 2026. arXiv:2608.02758.

Yiwen Zhu, Lihe Liu, Jiaqian Yu, and Di Zhang. LLM-Based Multi-Agent Orchestration: A Survey of Frameworks, Communication Protocols, and Emerging Patterns. *Future Internet*, 18(6):326, 2026.

Xinran Zhang, Pengrui Lu, Lyumanshan Ye, and Pengfei Liu. ERPBench: Evaluating LLM Agents for Enterprise Decision-Making Across Competitive Market Ecologies. 2026. arXiv:2609.04667.

A Design Rationale

Measurement pitfalls in feedback systems

Five pitfalls recurred in building this instrument. Each is stated in full where it arises in this appendix; they are collected here because none is specific to this task.

The nominal window is not the effective window. The shadow replay's effective window was W* = 2 against the W = 3 specification: an off-by-one in the current-round slot that produced 48.5% disagreement with the live run; after correction, zero. Unit-test any replay against the live run.

Same seed is not same inputs. The agent's own statements feed back into neighbors and the transcripts diverge; response-function comparisons must pair on the actually received stream, not on the seed.

A replayed baseline's initial conditions are themselves a design choice. The shadow starts with no input at t = 1, so every candidate ties at the top and the earliest cells are structurally empty.

A starved cell returns no verdict. If the environment never supplies the input, the response function cannot be estimated by watching: without the unanimity cell, a positive slope cannot distinguish pure bandwagon from the ascending phase of rise-then-reversal.

A code label can mislead its own authors. The family labeled "linear" consists of disjoint dyads; the chain reading misled even its designers.

Design decisions

Each entry records a design decision and the reason for it, grouped by the behavioral hypothesis it serves. Entries carry the number under which they were registered; the numbering is not an ordering of importance. Decisions shared across hypotheses appear under "Shared registration machinery". Punctuation in the quoted entries has been normalized for presentation.

N1: round weighting

- **D-01** (M+A): "We measure immediate excess response rather than final-outcome attribution: late-arriving signals face fewer subsequent opportunities to be overwritten, so final-attribution metrics carry a built-in recency slope independent of any weighting behavior."
- **D-02** (A): "We do not condition on 'not-yet-locked-in' games: identifying lock-in at round t requires the post-t trajectory, so the stratification would select on the outcome being measured."
- **D-03** (A): "The baseline for excess response is the spontaneous ranking drift of unmentioned targets, which is defined in every round; repetition-based baselines are undefined at the first delivery round."
- **D-04** (A): "An accusation counts as new at its first appearance in the evaluated seat's received stream over the whole game; window-based re-entries are analyzed separately, since late re-entries mix first exposures with reminders and would confound event type with round."
- **D-06** (A): "Direction is estimated as an event-level slope under a preregistered test hierarchy (slope, joint deviation, shape template) rather than by visual inspection of a handful of cell means."
- **D-07** (M+A): "Primacy and recency are not mutually exclusive, so a both-ends-overweighted profile is preregistered as its own outcome; absence claims are replaced by 'no deviation detected' with confidence intervals rather than an arbitrary equivalence band."

N2: bandwagon following

- **D-09** (A): "Unanimity skepticism shares the bandwagon measurement axis (level of neighbor agreement) and is merged as the opposing arm; temporal self-consistency suspicion lies on a different axis and is recorded only as a candidate mechanism."

- **D-10** (M+A): "High-agreement observations are restricted to windows in which the evaluated seat has not accused the target, blocking echo at its source; post-hoc echo tagging fails precisely in the unanimity cells where most agreement is echo-generated."
- **D-12** (A): "Agreement is computed per delivery round; window-cumulative counts measure repeated exposure, which belongs to the round-weighting row."
- **D-13** (A): "The primary axis is the accuser fraction with configuration stratification; shape verdicts are restricted to degree ≥ 4, where the scale has enough distinct levels."
- **D-14** (A): "The zero-agreement cell provides the built-in baseline: the response curve is read as excess over targets accused by no one."
- **D-15** (A): "Design calibration (cell supply, thresholds) uses only supply statistics from the discounted discovery model; the verdict model is sealed until judgment, and a third model is reserved unopened as a tie-breaker."
- **D-38** (A): "If the unanimity cell is starved, no directional verdict is issued: without it, a positive slope cannot distinguish pure bandwagon from the ascending phase of rise-then-reversal."

N3: CONFIRMATION ASYMMETRY

- **D-17** (M+A): "Equal vote weighting does not imply symmetric observable responses: rank changes occur only when margins are thin, so an unbiased aggregator exhibits margin-dependent asymmetry; we subtract the same-input naive asymmetry rather than assuming zero."
- **D-18** (M+A): "Rebuttals are restricted to third-party defenses: a self-defense can be elicited by the evaluated seat's own accusation and is the one rebuttal channel echo can manufacture."
- **D-19** (A): "The support arm activates only if statement–ranking mismatch rounds supply sufficient echo-free support events; the activation rule is preregistered before the gate is inspected."
- **D-20** (A): "Effects are reported on the log-odds scale alongside raw rates, since late-round retention approaches the ceiling and compresses raw differences."
- **D-21** (A): "Rounds with undefined incumbents, and deliveries containing both support and rebuttal, are excluded or quarantined by rule rather than case by case."

N4: DEFENSE DISCOUNT

- **D-22** (M+A): "The row label is restricted to behavioral asymmetry against an equal-weight reference; whether it constitutes a bias or an adaptive response is assessed only against a measured environmental benchmark."
- **D-24** (A): "Behavioral effect ratios and information ratios are different currencies; the calibration comparison is preregistered at order-of-magnitude resolution under an explicit Bayesian-weighting assumption."
- **D-25** (A): "Symmetric weighting is not automatically unbiased here: treating noise-generated defenses as informative is over-trust in defense, and this cell is preregistered accordingly."
- **D-26** (A): "All remaining design elements are inherited unchanged from the preceding rows: shared machinery rather than per-row tuning."
- **D-39** (A): "The calibration annotation uses half-order-of-magnitude bounds: the literal half of the preregistered order-of-magnitude resolution."

N5: NAIVE REFERENCE

- **D-27** (M+A): "A control's qualification is not being unbiased but being blind to the measured axis: the tally rule cannot condition on speaker identity, so any speaker-channel effect in the difference belongs to the model under test."
- **D-28** (M+A): "Same-seed baseline runs do not provide the same inputs: the agent's own statements feed back into neighbors and the transcripts diverge; response-function compar-

isons therefore use a shadow rule replayed offline on the model's actually received stream, while ceiling and system-level questions retain the co-run baseline."

- **D-30** (A): "Self-implication dynamics are removed at the round level (any window containing an accusation of the evaluated seat), preserving sample relative to whole-game exclusion."
- **D-31** (A): "The zero-mention neighbor-crowning cell isolates vote-independent speaker-channel effects; it does not by itself separate presence effects from behavior-evidence availability."
- **D-32** (A): "The design branch conditional on roster visibility is preregistered before the gate is inspected."
- **D-33** (A): "The co-run baseline retains system-level duties: its neighbor share must reproduce the topology-predicted ordering, serving as a positive control for the background-concentration measurement."
- **D-36** (A): "The shadow baseline is specified at full-ranking resolution with a deterministic tie rule, since the subtracted metrics require rank movements, not just top picks."
- The reference's response profile is measured, not assumed flat: a window-limited aggregator has structure of its own, subtracted rather than idealized away.
- **D-40** (A): "Paired cells require both total and discordant-pair minima, since concordant-heavy cells would otherwise masquerade information starvation as a null."

Shared registration machinery

- **D-34** (M+A): "All inference is fixed in advance: two-sided α=0.05, 95% CIs, event-level logistic models with game-clustered robust SEs, Wald joint tests, exact McNemar for paired outcomes; no cross-row multiplicity correction: an explicit choice, with error control delegated to the screening-then-confirmation design."
- **D-35** (A): "Verdicts are estimated on the held-out model alone; the discovery model serves only as a sign gate: a significant held-out effect with a same-sign (even nonsignificant) discovery estimate earns the label, while a significant opposite sign discards the row."
- **D-37** (A): "Promotion to confirmation follows a preregistered priority ordering grounded in proximity to the lock-in mechanism narrative, independent of observed effect sizes."
- **D-42** (M+A): "Every subtraction contrast maps to one prespecified estimand in a stacked model: an interaction Wald test, except the paired N5 contrast (game-clustered bootstrap CI) and raw-scale contrasts (marginal effects with delta-method CIs)."
- **D-43** (A): "Amendment rights expire at handoff dispatch: an auditable event (a git tag), not a state."

Registered statements

Punctuation in the quoted statements has been normalized for presentation.

Registered	Name in text	Statement (registered, verbatim translation)	Source
N1	round weighting	”Round-weighting profile (primacy / recency / both-ends overweighted)”; competing predictions: ”primacy = round-slope of the (P1 − shadow) excess is negative / recency = positive / both-ends = slope 0 ∧ joint deviation ∧ ends up, middle down / not detected”	prereg v3.2 §N1; free-recall profiles motivate candidate shapes (Murdock, 1962; Glanzer & Cunitz, 1966); timing-dependent primacy and recency motivate competing alternatives (Miller & Campbell, 1959); response timing follows the step-by-step versus end-of-sequence distinction (Hogarth & Einhorn, 1992). These precedents do not validate our ranking-based measure
N2	bandwagon following	”Majority-agreement profile (bandwagon / unanimity reversal / rise-then-reversal)”; ”bandwagon = slope in a positive, rising up to unanimity / unanimity reversal = . . . negative deviation at a = 1.0 / rise-then-reversal = positive slope ∧ terminal reversal”	§N2; human precedents concern majority size and dissent (Asch, 1955; Allen & Levine, 1969); Bond (2005), aggregating 125 studies, finds that the size–conformity relation varies with the paradigm. These are not direct tests of our agreement-proportion profiles. The unanimity dip and rise-then-reversal are registered alternatives, not effects established by these citations. Repeated voices and source dependence (Weaver et al., 2007; Wilder, 1977; Yousif et al., 2019) are relevant to the echo restriction (D-10), not evidence for a unanimity dip
N3	confirmation asymmetry	”Confirmation asymmetry: rebuttal resistance (uniform freeze vs selective freeze)”; ”resistance ≡ rebuttal effect (shadow) − rebuttal effect (P1); positive = resistance”	§N3 · D-41; selective evaluation of evidence (Taber & Lodge, 2006; Lord et al., 1979) motivates the selective alternative; uniform versus selective freeze is our operational contrast. Möbius et al. (2022) measure asymmetric updating beside conservatism in self-assessment. The third-party restriction (D-18) is an a priori design choice informed by attribution theory (Kelley, 1973), not a restriction tested in those studies

Registered	Name in text	Statement (registered, verbatim translation)	Source
N4	defense discount	”Accusation–defense asymmetry (negativity: dual scale)”; ”accusation-dominant = both excesses $\geq 0 \wedge$ —accusation— $>$ —defense— / . . . negative excess = descriptive report + per-direction flag”	§N4
N5	audible-neighbor concentration	”Audible-neighbor concentration (shadow-naive excess neighbor concentration)”; ”excess concentration = paired positive / not detected / reversed = negative (unexpected)”	§N5
P1	the below-baseline landscape	”k = number of configurations (of 11) in which LLM MRR $<$ naive MRR” · reproduction at k ≥ 9	crossmodel §1
P2	the persistence–failure covariation	”configuration-level ρ(w$\rightarrow$w persistence, MRR), n = 11” · reproduction at $\rho \leq -0.6$	crossmodel §1
P3	the sub-chance existence check	”whether any configuration has LLM MRR $<$ chance (H_{N$-$1}/(N$-$1)) . . . the peak of the shortfall (an extreme case), not the spine”	crossmodel §1
P4	the winners’ connectivity ordering	”whether the median degree of the configurations in which the LLM beats naive exceeds that of the configurations in which it loses” (boundary shift permitted)	crossmodel §1
P5	the self-implication hypothesis	H_acc as registered: ”when P1 is named, it reads the accusers (usually honest) as ’coordinated framing’ and suspects them first $\rightarrow$ avoids the true adversary $\rightarrow$ fails” + three support conditions ((i) significant excess of wrong under [named $\wedge$ final top-1 = accuser] (ii) accuser = adversary $<$ 30% (iii) same direction in ≥ 2 configurations)	accuser prereg, as registered

Registered	Name in text	Statement (registered, verbatim translation)	Source
P-a	the narration hypothesis	"H_Pa: … when split into (i) intent-attributing and (ii) evidence-describing, the wrong rate of (i) is significantly higher than that of (ii)"	prereg_tom §P-a
P-b	response-rate dynamics	"H_Pb: P1's reaction rate to a neighbor's target-change event … decreases as rounds proceed"	prereg_tom §P-b

Both the below-baseline landscape and P4 were registered against point estimates; re-execution uncertainty was not part of either statement, and neither verdict is revised here. Re-running configurations under their registered settings puts a sign threshold on each cell (Appendix B.12), and against those thresholds nine of the 44 cells have signs a single run does not settle. The two registrations are touched differently. The landscape registration counts, so the uncertainty attaches to a number: every model still meets k $\geq$ 9 at its point estimate while the interval reaches below 9 for three of the four. P4's subject is the set of configurations in which the model beats the tally rule, and five of the nine unsettled cells sit outside the hub family on the losing side, so what moves is the population its median is taken over. We report this beside the registered verdicts rather than in place of them.

What the code does that the description rounded off

Four things in §2 were checked against the code that ran the campaign rather than against this appendix, and four were off. They are stated here as the code has them.

The occupant is not windowed. The shadow rule scores the window. The occupant is sent the two previous rounds' messages along with the current one, and each of those already listed the statements visible then, so at the final round its context reaches back to round one; its own reasoning recites the opening round in a fifth to five sixths of final rounds depending on the model. The two judges therefore receive overlapping information with different histories; more context need not improve performance, so this asymmetry does not establish a lower bound on the deficit. As an empirical sensitivity check, widening the shadow rule's window to the whole game, so that it reads everything the occupant did, increases the number of cells in which the model falls below the shadow rule from 34 of 44 to 37 of 44, and moves the audibility split from $+0.174$ $[+0.091, +0.258]$ to $+0.225$ $[+0.139, +0.311]$; the three cells that change sign all change from the occupant above the rule to below it. The deficit persists and grows under this comparator change. That addresses the tested window choice; it does not establish that all history choices favor the model. The comparison is tilted toward the rule and is reported as exploratory: the whole game includes round one, whose honest statements are true by construction, so the widened rule collects the truthful opening statements that form one of the two channels combined by the oracle-informed comparator in B.16. The registered comparison remains the windowed one. Source results/verify_2026-09-04/shadow_fullwindow_2026-09-18.py.

The window holds two rounds, not three. $W = 3$ is the specification, but statements are delivered at the top of the following round, so the current round's slot is always empty and the seat sees two rounds of statements. This was recorded as a property of the replay; it is a property of the live instrument.

The honest mix is not thirds. The runner assigns $\lfloor (N-2)/3 \rfloor$ seats to each of the first two variants and the remainder to the tracker. At $N = 10$, this gives two, two and four seats, respectively: a quarter, a quarter and a half of the honest seats. This allocation applies in five of the eleven configurations and 45.5% of the games. The tracker is the variant that drops echoes before tallying, so it is the one furthest from the comparator.

The parser's filled entries are not last. Unnamed seats are appended at confidence 5, while a seat the model itself ranked is clamped only to the range 1 to 99, so a named seat below 5 sorts under an

invented one. Whether that happened cannot be counted for this campaign, because the flag marking filled entries was not serialized until 2026-09-04 and the main campaign predates it; what can be said is that rankings holding both an entry at exactly 5 and one below it number 30, 102, 52 and 208 of 539 per model, for Gemini, GPT-5.6, DeepSeek and Llama, respectively.

The record contains one more point that the paper did not report. DeepSeek V4's calls hit the token limit in 19.0% of 11,000 calls, and the runner's own gate recorded that as exceeded (`results/f15_crossmodel/deepseek_nt_a/_finish_reason.json`). Those calls are that directory's control and generic-warning arms together, eleven configurations at fifty seeds including seed 0, two calls in each of five rounds; the games analyzed elsewhere exclude seed 0 and the warning arm, so the two populations differ. A truncated second call is a short ranking, which the parser then fills, but this rate does not identify which incomplete rankings arose that way: among out-of-view games whose raw ranking is incomplete, DeepSeek's final-round outputs close their format and name all but one seat, while Gemini's are much shorter and do not close. Truncation can therefore contribute to incomplete rankings, most visibly for Gemini. We report the rate and do not claim a causal path from it to any result, because nothing in the record establishes one.

Scope and cost

- **D-51 EN** (A): "A fifth model point completes no argument the two-stage screening-confirmation design has not already closed."
- **Recorded campaign costs.** Execution records report approximately $71 for the GPT-5.6 main campaign, $72 for the Gemini main campaign, $9.47 for the DeepSeek three-arm campaign, and $5.11 for the initial seat-swap spot run. These cover the named runs rather than the whole study.[1] The search grid required no new runs.
- **Data and code.** Experimental numerical results are linked to named analysis scripts over retained records; the costs above instead come from execution records. Each comparative analysis identifies which of the two deployments of the rule it used, since the co-run and the shadow tally (Baseline rationale above) do not give the same figure. Those scripts, the preregistrations, the per-configuration outputs behind every table, and per-seed outputs for the Appendix D pilot and probe runs are retained and prepared for release. This text is anonymous, so it carries no link.

Baseline rationale

A baseline earns its seat by being blind to the axis under measurement: the tally rule does not use speaker identity when aggregating target counts. This does not make its performance invariant to the channel: received target distributions vary with audibility, so a model-minus-rule contrast remains comparator-dependent. Two deployments serve two question types: comparisons of whole runs (§3.1's co-run cell map) use a co-run tally occupying the same seat under the same seeds; comparisons that subtract two readings of one stream (the audibility split, the transplant decomposition, the hop step, §4) use a shadow tally replayed offline on the evaluated seat's actually received stream, because same-seed runs do not provide same inputs. The evaluated seat's own statements feed back into its neighbors. The shadow replay is unit-tested: an effective-window discrepancy initially produced 48.5% disagreement with the live run; after correction, zero. Its full specification (deterministic tie rule, measured response profile, and the co-run's positive-control duty) is given in the design decisions above (D-28, D-33, D-36).

Preregistration protocol

The behavioral search fixed its hypotheses, priority and statistics before the named screening outcomes were unsealed, using already collected games; this is distinct from registration before data generation. Five behavioral hypotheses were preregistered (registered labels, full statements, and the label-to-name correspondence: the Registered statements table above); the aiming criterion was frozen on June 30, before any held-out model ran. The search grid ran screening on GPT-5.6 with a sign gate on the discovery model, then confirmation sealed to DeepSeek V4; promotion order was fixed before screening, blind to effect sizes; the handoff tag locks the specified design decisions,

[1] The GPT-5.6 record reports billing after the batch discount; the Gemini record combines batch billing with an approximate cost for initial real-time calls. The DeepSeek and seat-swap amounts are recorded ledger costs.

while the retained v3.1/v3.2 procedural amendments and N1 template precede the recorded screening readout. Inference is uniform (two-sided $\alpha = 0.05$, 95% CIs; the full model and test specification is in entry D-34 above) with no cross-row multiplicity correction, an explicit choice whose error control rests on the promotion rule being fixed before screening and on the second stage running one test on a model that took no part in forming the hypothesis. What is held out is the occupant, not the environment: the confirmation model plays the same 49 seeds, with the same graphs and the same adversary placements, as the screening model. A reader who wants the second stage to be independent of the first in the games as well as in the model will not find that here, and we say so rather than leaving *held out* to carry more than it does.

SocialMaze adaptation

Our starting point was the original SocialMaze hidden-role task (arXiv:2505.23713v1). We retain criminal and investigator roles, let the evaluated model generate statements as an active player, and restrict visibility to graph neighbors. Our five-round game also changes the other investigators: they state the truth in the opening, then normally act on received statements and can be wrong (the no-message fallback is specified below). These are adaptations, not a reproduction of SocialMaze.

Model and cost details

Four models span open mid-tier to frontier: Llama-3.1-70B-Instruct (open mid-tier; w8a16 quantization), GPT-5.6 (closed SOTA; gpt-5.6-sol), Gemini 3.1 Pro (closed SOTA; preview), and DeepSeek V4 (open frontier; V4-Flash-0731), all at temperature zero. Model serving remains nondeterministic, so verdicts are pattern-level (§6). Where thinking modes appear they arrive as protocol bundles (max_tokens and related settings change together), a caveat stated with the comparison in B.13. Inference cost and per-campaign breakdown: the Scope and cost section above.

Topology roster

The same game replays across eleven communication landscapes, ordered by how much of the network the evaluated seat hears directly: isolated pairs (one audible neighbor), ring (two), small-world (two to seven, varying by seed[2]), cycle-heavy (six), and hub-and-spoke with the evaluated seat at the hub (nine to fourteen), with network sizes sweeping N = 10 to 20 within families for eleven configurations in all (Figure C.1, drawn from actual edge lists).

Figure 1 is a conceptual comparison with no sampled game. The Figure C.1 panels use small-world N = 10, seed 3800; small-world N = 15, seed 3600; hub-and-spoke N = 10, seed 400; and hub-and-spoke N = 15, seed 300. The remaining 7 configurations are deterministic.

Adversary emission rule

Each round, our adversary selects a target uniformly from its neighbors, excludes itself, and accuses or defends with equal probability, independently of truth and conversation history. Every possible target is innocent, so it defends an innocent seat with probability one half. SocialMaze v1 describes a criminal that reacts to accusations (App. B, item 2), whereas the legacy generator draws a non-self target from all players and flips the claim randomly.[3] We retain random emission but restrict targets to neighbors. Our results therefore concern this non-adaptive policy; they do not measure resistance to a strategic attacker.

Honest emission rule

Every honest seat runs one of three deterministic rules, assigned to seats by the seed in roughly equal thirds (naive, careful, tracker; Appendix B). All three share one opening move: in round 1,

[2]The Watts–Strogatz rewiring draws the evaluated seat's degree per seed; stratifications touching this family are therefore restricted to degree ≥ 4.

[3]Legacy generator: `https://github.com/xzx34/SocialMaze/blob/104f83c4cd3895fcfb5e94ee5c6dabb55afab1cf/hidden_role_deduction/hrd_gen.py`, function `generate_statement`. This archived implementation is the comparison source, not a record of our original checkout.

and in any round in which the seat has heard nothing, the rule picks a target uniformly at random among the other players and states the truth about it, "is the criminal" if the target is the adversary and "is not" otherwise. This is the only contact any honest seat has with ground truth (§2). From round 2 on, each rule scores every other player from the statements it has heard within its window, an accusation counting +1 and a defense −1 against the target, and accuses its current top scorer; it never defends after the opening and never consults ground truth again. Ties in the score are resolved in player-number order. Naive uses the raw tally. Careful weights each statement by a trust score for its speaker, lowered when that speaker contradicts an earlier claim of its own. Tracker additionally detects echoes, a speaker repeating what another speaker said the round before, and skips echoed statements before tallying. Two consequences follow. Honest seats emit defenses only in the opening round, not by construction, since the opening branch also fires in any later round in which a seat has heard nothing, but in fact: all 1,063 post-opening defenses are the adversary's, in every model. And the adversary, which cannot target itself, is never defended after round 1 (0 of 1,263 receipts for one model, Appendix B).

PROMPTS GIVEN TO THE EVALUATED SEAT

The evaluated seat receives, every round, a fixed system prompt, a round message listing the statements heard (grouped by round), and a formalization prompt that fixes the output format. Verbatim, with {N} the number of players, {self.player_id} the seat, and {self.neighbors} its neighbor list (zero-shot level; the provenance and defense variants of the prompt are not used in this paper):

```
You are Player {self.player_id} in a social deduction game
with {N} players.

Game Rules:
- There is exactly 1 Criminal and {N-1} Investigators.
- Each round, every player makes one statement about another
player:  "Player X says Player Y is (not) the criminal."
- Investigators always provide truthful information.
- The Criminal makes statements that may be truthful or
false to confuse others.
- You can only see statements from your direct neighbors:
{self.neighbors}.

You are told you are an Investigator.  Your goal is to
identify the Criminal.
```

The corrected-premise arm of the ablation campaign (Appendix E) replaces the two premise lines above with the following, verbatim, and changes nothing else:

```
- Investigators state their honest current belief based only
on what they have heard.  Their beliefs are often mistaken,
especially in later rounds.
- The Criminal's statements are unrelated to what it knows.
```

The round message reads "Round {t}: This is the first round. You have no information yet." in round 1, and afterwards "Round {t}: Here are the statements you can see from your neighbors:" followed by the statements grouped by round and the request "Now analyze these statements freely. Consider: What patterns do you notice in the statements? Are there any contradictions? Who seems most suspicious and why? Could any of these statements be echoes of the same original claim?" The formalization prompt then reads:

```
Based on your analysis above, now provide your formal output
for Round {round_t}.

You MUST follow this EXACT format:

(1) Statement:  Player {self.player_id} says Player [NUMBER]
```

```
is [or "is not"] the criminal.

(2) Suspicion Ranking (list ALL other players, most
suspicious first):
1.  Player [NUMBER], confidence:  [NUMBER]%
2.  Player [NUMBER], confidence:  [NUMBER]%
...  (continue for all players)

(3) Perceived trust:  [NUMBER]%
How much do you think other players trust you right now?
(0-100%)

IMPORTANT:
- For (1), pick the player you are MOST suspicious of and
accuse them.
- For (2), list EVERY other player with a confidence
percentage (1-99%).
- The confidence percentages should reflect your actual
beliefs from your analysis.
- For (3), estimate how trustworthy you appear to others
based on your statements so far.
```

INTERVENTION TEXTS

Each intervention of Appendix F appends one paragraph to the system prompt. The wording is reproduced below, with punctuation normalized for presentation:

```
IMPORTANT - About Inconsistency:
Players may appear inconsistent, changing their accusations
across rounds.
Inconsistency alone is not reliable evidence of deception.
A player may change
their opinion based on new information.  Base your judgment
on the specific
content of accusations, not on the fact that someone changed
their mind.

IMPORTANT - About Your View:
You hear only your neighbors, a subset of the players in
this game.
The criminal might not be among the players you can hear
directly.
Your window on the game is partial.  Base your judgment on
the specific
content of accusations, not on whose voices happen to reach
you.

IMPORTANT - About Being Accused:
Other players may accuse you of being the criminal during
this game.  An
accusation against you is not, by itself, reliable evidence
that your accuser
is the criminal; honest players can be genuinely mistaken.
Base your judgment
on the content of accusations, not on whether they target
you.
```

Null calibration for the same-seed criterion. The registered cell rule labels an intervention responsive when top-1 improvements exceed deteriorations by at least five among 49 paired seeds, and harmful for a net change of at most -5. A cell with fewer than five discordant pairs cannot cross either threshold in these observations. This does not establish that its population effect is zero. There are 24 reachable cells among the generic warning's 44, 15 among the structure disclosure's

36 non-hub cells, and 11 among the behavioral warning's 12. Six generic-warning cells and six non-hub disclosure cells have no discordant top-1 outcomes.

The following calibrations are post-hoc. Conditional on each cell's observed discordant count m, they assume equally likely improvement and deterioration: $C \sim \mathrm{Binomial}(m, 1/2)$ and net change $2C - m$. Cells are treated as independent, although they share seeds. We compute the aggregate distributions by exact convolution. This is a conditional calibration under those assumptions, not a test that accounts for all dependence across cells.

In the order generic warning (44 cells), structure disclosure (36 non-hub cells), and behavioral warning (12 cells), the expected numbers of responsive-or-harmful cells are 3.98, 2.20 and 1.47, with 95% reference intervals [1, 8], [0, 5] and [0, 4]. The observed counts are 6, 4 and 1; their upper-tail probabilities are 0.190, 0.164 and 0.817. Counting only responsive cells gives 0.011, 0.020 and 0.543. The two-sided tail probabilities for the signed count, responsive minus harmful, are 0.006, 0.019 and 0.660. These statistics answer different questions; the latter two retain direction information that the total count discards. None establishes which structures or models cause an intervention to work.

Reachability differs across models and configurations. On the disclosure arm, discordant pairs average 11.8 per hub cell and 5.2 per non-hub cell. Among the non-hub cells, the model-specific means are 5.4, 2.1, 7.9 and 5.2 for Llama-3.1-70B, GPT-5.6, Gemini 3.1 Pro and DeepSeek V4.

Arm	Model	cells	reachable	responsive
generic warning	Llama-3.1-70B	11	6	4
generic warning	GPT-5.6	11	4	0
generic warning	Gemini 3.1 Pro	11	7	2
generic warning	DeepSeek V4	11	7	0
structure disclosure (confirmatory)	Llama-3.1-70B	9	5	1
structure disclosure (confirmatory)	GPT-5.6	9	0	0
structure disclosure (confirmatory)	Gemini 3.1 Pro	9	5	3
structure disclosure (confirmatory)	DeepSeek V4	9	5	0
behavioral targeting	Llama-3.1-70B	3	3	1
behavioral targeting	GPT-5.6	3	2	0
behavioral targeting	Gemini 3.1 Pro	3	3	0
behavioral targeting	DeepSeek V4	3	3	0

GPT-5.6 supplies no reachable confirmatory cell on the disclosure arm. It is also the model that settles earliest: its final pick already stands at the end of the opening round in 272 of 539 games, against 26, 63, and 101 for the other three, and its median sealing round is 1 where theirs is 3. Early settling and few intervention-induced top-1 changes co-occur here; this does not establish that the former causes the latter.

Source: results/verify_2026-09-04/warning_probe_2026-09-20.py (raw outcomes and exact calibration); b7_b8_b9_b12.json (B12, settling times).

Analysis documentation

The supplementary `ANALYSIS_CHANGES.md` groups material changes by analysis: comparator and population definitions, temperature-matched comparisons, intervention decision rules, annotation provenance, and registration timing. Each topic states the current interpretation and points to its supporting analysis. Original protocols retain their historical wording and amendments; the executable reproduction registry identifies computations without implying preregistration. These documents supplement the methods and limitations given with each result.

Five traps from one counterfactual arm. Running the tie-policy arm on the relay configurations produced five failure modes that are silent by construction, and they are recorded because each would have produced a plausible result. (1) *A shared response cache.* The batch runner keys cached calls by game and call index and stores them under the main campaign's directory, so a second arm run under the same condition label replays the first arm's answers while the logs look normal; the

counterfactual becomes vacuous with no visible symptom. Each arm now gets its own output root, and the run refuses to start if a cache file is already there. (2) *A global RNG under threads.* The registered tie patch holds one RNG; the batch runner plays games in concurrent threads, so the draws depend on interleaving and are not reproducible per game. A thread-local shim seeded from each game's key restores per-game reproducibility; the policy the registration fixes is uniformity within a tie, not the plumbing. (3) *A default argument that changes the graph.* `ring_N10` is the shortcut family with zero shortcuts; omitting the argument takes the default and yields twelve edges instead of ten, a different graph under the same name. (4) *A shell that hides a dead lane.* One runner rejected an argument its sibling accepts; the lane's `set -e` then ended the remaining arm, and the log read as a completed run above the error line, so four hours passed unnoticed. Completion is now judged by counting output files rather than by reading logs, and a configuration-error signature is reported rather than restarted. (5) *An estimate and a meter that disagree.* One runner's cost gate estimated the batch price while its collector accrued the standard price, a factor of two, so a spend cap nearly halted an arm over money that was never spent; the invoice reconciliation for an earlier campaign of the same provider bills $69.00 against $132.54 at standard pricing, and cost records retain the estimates and reconciled charges separately.

The models, as served. The four are addressed as `gpt-5.6-sol`, `gemini-3.1-pro-preview`, `deepseek-v4-flash` with its thinking mode off, and `neuralmagic/Meta-Llama-3.1-70B-Instruct-quantized.w8a16` served locally under vLLM. We name them GPT-5.6, Gemini 3.1 Pro, DeepSeek V4 and Llama-3.1-70B throughout; the strings above are what the requests carried. DeepSeek's is the flash variant rather than the full one, and the label *V4-Flash-0731* used in §2 is ours for this endpoint, not the provider's: that API returns no dated snapshot, so the campaign recorded a fixed string in each game's metadata rather than a served version. Nothing downstream verifies that the endpoint stayed the same, which is the checkpoint limitation §6 states.

The two prompt lines, quoted. The specification mismatch of §2 is these two sentences. The adversary is introduced as a player whose statements "may be truthful or false to confuse others", while the implementation draws its target uniformly at random and flips a coin for the claim, so the description admits a strategic adversary the code never builds. The investigator role states that "Investigators always provide truthful information". The honest seats satisfy that only in the opening round; from round two they accuse whoever tops their count, and about 80% of those accusations name an innocent seat. The sentence is ambiguous between sincerity and factual truth. The scripted rules preserve the former but not the latter; the latter reading licenses the deduction stated in §4.1.

B Supplementary Tables

B.1 Cross-model verdict table

Hypothesis (name in text · registered)	Llama-3.1-70B (discovery)	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
the below-baseline landscape · P1	k = 10 of 11	k = 9, reproduced	k = 9, reproduced	k = 9, reproduced
the persistence–failure covariation · P2	$\rho = -0.891$	$\rho = -0.917$, reproduced	$\rho = -0.982$, reproduced	$\rho = -0.927$, reproduced
the sub-chance existence check · P3	2 cells (small-world N = 15, ring N = 15)	0	0	0
the winners' connectivity ordering · P4	holds (upper-median degree 9 among winning configurations, 4 among losing)	holds (14 vs 4)	holds (14 vs 4)	holds (14 vs 4)

P4 is read under the registered tie convention and against the co-run tally. Against the shadow tally with ties randomized at $N = 10$, no hub interval lies above zero, three containing it and the DeepSeek interval lying below it (Appendix G), so the ordering rests on a lead the convention partly supplies.

Hypothesis (name in text · registered)	Llama-3.1-70B (discovery)	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
the self-implication hypothesis · P5, neighbor accusations received before judgment	discovery data, not counted	81% vs 30%, OR 10.01, met	78% vs 52%, OR 3.35, met	75% vs 41%, OR 4.25, met

The P5 row uses neighbor accusations from rounds $1, \ldots, T-1$, available before the final judgment at T. Its numeric support conditions are met in all three held-out models; they do not establish that the model interpreted an accusation as coordinated framing. The eligibility and support thresholds are unchanged. Appendix G examines the separate tie-policy intervention; intervals that include zero do not establish policy invariance.

The discovery model fixes the criteria and is not counted as a reproduction; the three held-out models were run after the criteria were frozen (registered statements and thresholds: Appendix A).

Source: afterB_analysis.json; {sol,gemini,deepseek_nt_a}_p1_p5.json. P2 is Spearman for all four models (the discovery model recomputed from exact fractions rather than the logged Pearson value; the API models use the stored Spearman). P4 uses the upper middle degree when a group contains an even number of configurations, following the registered convention. P5 receipt-time values: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py.

B.2 Sensitivity of the round-direction reading to dropping t = 3

Three quantities are tracked over the reaction rounds t = 3–5: the event reaction rate (the share of neighbor target-changes the evaluated seat responds to), the non-event rate against which it is read (neighbors that kept their target over the same rounds), and the excess of the first over the second. Each cell gives the direction over the full window, then the direction with t = 3 dropped; a cell is marked *fragile* when the two disagree. Direction compares the first and last round of each window, not a fitted slope. t = 3 is the round whose non-event sample is smallest and self-selected, which is why dropping it is the sensitivity of interest.

Model	Event reaction rate	Non-event rate	Excess
	t3 → t5 / t4 → t5	t3 → t5 / t4 → t5	t3 → t5 / t4 → t5
Llama-3.1-70B	decrease / decrease	decrease / decrease	increase / increase
GPT-5.6	increase / decrease (fragile)	decrease / decrease	increase / decrease (fragile)
Gemini 3.1 Pro	decrease / increase (fragile)	decrease / increase (fragile)	increase / increase
DeepSeek V4	decrease / decrease	decrease / increase (fragile)	increase / decrease (fragile)

Only the discovery model is stable across all three quantities. The registered verdict is read on the full window throughout; this table is a robustness annotation, since narrowing the window after seeing the data would be a post-hoc choice.

Directions are read on pooled rates. Game-level pairing exists only for the full window (t = 3 versus t = 5), where it agrees with the pooled reading except for the event rate in GPT-5.6 and for the excess, whose paired estimate runs the other way in three models and reaches significance in the discovery model alone. Pairing on the t = 3-dropped window is not registered and was not computed.

Source: pb_results.json (GPT-5.6, Gemini 3.1 Pro, DeepSeek V4); llama_recompute_2026-08-25.md (Llama-3.1-70B, dry-run recomputation).

B.3 Supply statistics (Llama discovery stream, 539 games = 49 seeds × 11 configs)

Row	Cell	Count	Counting condition
N2	unseeded unanimity (agreement fraction a = 1)	6	extended silence
N2	near-unanimous band (a = (deg−1)/deg)	26	extended silence
N3	rebuttals of a current suspect, t2 / t3 / t4 / t5	134 / 3 / 5 / 1	no silence filter; 1 valid round (n $\geq$ 30)
N4	defenses, t3 / t4 / t5	66 / 60 / 63	no silence filter
N5	paired / discordant / co-top excluded	241 / 100 / 246	
N5 confirmation (DeepSeek V4)	paired / discordant	224 / 76	

Extended silence requires the evaluated seat to have neither accused nor defended the target in its own statement window, rounds $\max(1, t-2), \ldots, t-1$. The N2 rows count eligible units with degree at least four. The rebuttal and defense rows are counted without the silence filter, as registered.

Source: results/verify_2026-09-04/screening_scope_2026-09-20.py (`screeningscope`); supply_stats_llama_raw.json and supply_stats_llama_addendum.json; n5_confirm_result.json for the DeepSeek row.

B.4 Qualitative coding detail

Stage	Games	Coded	Result
Llama-3.1-70B, stage 1	425	416 (9 excluded)	intent-attributing 411 (301 wrong / 110 correct), evidence-describing 5 (2 / 3); 98.8% intent-attributing; the comparison group contains only five texts and is treated descriptively
GPT-5.6, stage 2	453	433 (20 excluded)	intent-attributing 164 (128 wrong / 36 correct), evidence-describing 269 (197 / 72); 78% versus 73% wrong, OR 1.30, Fisher p = 0.30, not supported

The first pass used automated Claude workers, recorded as `claude-fable-5`, with five texts per batch (the last GPT-5.6 batch contained three). The same category prompt preceded both stages and requested a supporting quotation. It omitted the predicted association with error; coder texts omitted the ground-truth outcome and adversary identity, but paths and headers retained model, configuration and seed identifiers. These records do not establish complete session isolation. Labels were intent, evidence or exclude; no separate borderline flag was collected. The archived sample includes games with an accusation from any other player over all rounds, across nine Llama and ten GPT-5.6 configurations; it is distinct from the before-judgment neighbor-accusation population used elsewhere.

The table retains the original labels. One GPT-5.6 supporting quotation belongs to another text; omitting that case or assigning it to any of the three categories leaves the Fisher test nonsignificant. This sensitivity check does not validate the remaining labels. The table uses the registered two-sided Fisher test and the unadjusted odds ratio.

Source: results/verify_2026-09-04/first_pass_coding_2026-09-20.py (`firstcoding`); retained labels, coder inputs and extracted workflow prompt in results/pa_intent/.

A second coding pass, run after the ablation arms had returned, replaced the intent-attribution categories with premise-deduction ones: whether a reasoning text states the prompt's truthfulness premise, derives its own innocence from it, or concludes that an accuser must therefore be the adversary. Two separate Claude Opus 5 sessions annotated the same 100 texts (25 per model) under a protocol fixed before any label, with model and condition identifiers withheld from the coder file. Their primary labels agreed on 93 of 100 (Cohen's κ = 0.886 over intent, evidence, rule and unknown). This is agreement between two sessions of one model on this sample, not agreement with human judgment and not a measure of annotation accuracy. Both sessions labeled rule as the primary category for 1 of 25 texts in the discovery model against 15 to 19 of 25 in the held-out models. Four texts carried a third-person self-description that both coders first read as another seat; the sample is unchanged after correction. The pass is exploratory and determines no preregistered verdict.

Source: results/verify_2026-09-04/recoding_agreement_2026-09-19.py (`recoding`); original labels and sample mapping in results/recoding_c4_2026-08-31/.

B.5 K-MAP DETAIL (PER CONFIG: BASELINES VS 4 MODELS, FINAL MRR)

Config	deg	Tally	Tracker	Llama	GPT-5.6	Gemini	DeepSeek	defender-first rule
isolated pairs N10	1	0.392	0.392	0.346	0.320	0.318	0.331	0.358
isolated pairs N20	1	0.265	0.265	0.223	0.216	0.227	0.229	0.251
ring N10	2	0.565	0.554	0.337	0.334	0.340	0.394	0.563
ring N15	2	0.388	0.382	0.220	0.243	0.281	0.255	0.386
small world N10	4	0.582	0.582	0.453	0.561	0.503	0.530	0.747
small world N15	4	0.473	0.463	0.193	0.312	0.335	0.343	0.489
cycle heavy N10	6	0.660	0.640	0.614	0.634	0.572	0.627	0.851
cycle heavy N15	6	0.602	0.623	0.351	0.437	0.461	0.345	0.643
cycle heavy N20	6	0.566	0.576	0.386	0.370	0.401	0.389	0.594
hub N10	9	0.694	0.674	0.808*	0.735*	0.702*	0.765*	0.965
hub N15	14	0.569	0.528	0.556	0.662*	0.674*	0.799*	0.938

The defender-first rule promotes a neighbor who defended after the opening and before the final judgment; otherwise it follows the shadow tally. Its column averages full-precision MRR over the four models' streams before rounding. This comparator uses an environment-specific cue and is not a proved ceiling. Source: results/verify_2026-09-04/timing_corrected_2026-09-20.py. * = LLM $>$ the co-run tally rule. Lv.3 robustness: 9 of 11 below, largest gap -0.272 (cycle-heavy N = 15). Tracker is the speaker-trust-weighted, echo-filtered rule of Appendix A, run separately in the evaluated seat over the 49 seeds, not replayed on the four model streams. Sources: run_baseline_codereview.py; baseline_afterB.json; afterB_figure_points.tsv. Per model, the defender-first MRR is 0.951/0.927 (Llama-3.1-70B), 0.973/0.945 (GPT-5.6), 0.986/0.950 (Gemini 3.1 Pro) and 0.951/0.932 (DeepSeek V4) in hub N10/N15, averaging to 0.965/0.938; the separately executed tracker scores 0.674/0.528.

B.6 N4 DIAGNOSTICITY BENCHMARK (RAW TABLE)

Stream	LR (accusation)	cell (criminal / innocent)	LR (defense)	cell	honest stratum p(x = criminal)	adversary stratum
Llama-3.1-70B	8.508	427/1690 vs 633/21314	0.0	0/1263 vs 1778/22459	0.5083 (n=840)	0.0 (n=201)
GPT-5.6	9.039	281/1531 vs 435/21422	0.0	0/1250 vs 1764/22751	0.5393 (n=521)	0.0 (n=174)

This benchmark is a screening row: it runs on the gate model and the verdict model only, as registered (Appendix A). Each ratio is a probability ratio over the units carrying its own signal or no mention ($\{$accusation, no mention$\}$ for the first and $\{$defense, no mention$\}$ for the second) so the two rows have different denominators by construction: the criminal stratum's 1,690 and 1,263 differ by exactly the 427 accusations it contains.

No unit in the criminal stratum ever carries a defense (0 of 1,263 and 0 of 1,250), so the defense likelihood ratio is zero by construction. A third speaker stratum, units whose accusers are mixed, holds the remaining 19 and 21 units and is not printed above; the three strata sum to the accusation counts, $201 + 840 + 19 = 427 + 633$ and $174 + 521 + 21 = 281 + 435$.

Source: n4_diagnosticity_benchmark_result.json

Configuration labels (all tables and figures). Rows and axes use the print names: isolated pairs, ring, small world, cycle heavy, hub. The code labels they replace ('linear', 'sw', 'cycle', 'hub') appear only in the source and the reproduction registry.

Rounding convention (all tables). Decimals are rounded to the digits shown using half-away-from-zero on the value as logged by the producing script, not by binary floating-point truncation; integers and counts are reproduced as logged.

B.7 The self-implication cue by unit of analysis

Each row uses its own matched reference. An accuser is a neighbor that named the evaluated seat before its final judgment ($r < T$). Columns list GPT-5.6, Gemini 3.1 Pro and DeepSeek V4. Statement counts exclude the opening round and use all configurations; game-level rows use B.31's eligible configurations.

Unit	Observed	Matched reference
A heard accusation of the evaluated seat: its speaker is the adversary	5.1% · 4.9% · 2.9% (2332 · 2421 · 4064 statements)	random other player: 8.2% · 8.3% · 8.1%; $1/(N-1)$ weighted over statements
A game's accuser set contains the adversary	26% · 25% · 24% (414 · 441 · 449 accused games)	random set of the same size: 24% · 26% · 33%
The chosen accuser is the adversary (event games)	19% · 22% · 25% (380 · 369 · 362)	uniform draw from the accuser set in those same event games: 15% · 16% · 11%
Event games ending wrong	81% · 78% · 75%	one minus the row above, by definition

Source: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py.

B.8 Per-event hit rate, model versus the shadow tally

Paired population: games in eligible configurations with a neighbor accusation of the evaluated seat before judgment and a unique top suspect under the shadow tally ($n = 264, 327, 265$). The shadow reads rounds $T-2, T-1$; the accuser set spans all delivered rounds $1, \ldots, T-1$. Entry is the share whose final top suspect is an accuser; hit rates concern the true adversary.

	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
Model: entry · hit given entry · hit given no entry · total	0.879 · 0.237 · 0.813 · 0.307	0.813 · 0.237 · 0.557 · 0.297	0.721 · 0.319 · 0.730 · 0.434

	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
Shadow tally: entry · hit given entry · hit given no entry · total	0.708 · 0.241 · 0.675 · 0.367	0.398 · 0.392 · 0.401 · 0.398	0.434 · 0.443 · 0.667 · 0.570
Total hit rate, model minus rule (opening-round accusations excluded)	-0.061 (-0.065)	-0.101 (-0.102)	-0.136 (-0.137)

The paired accuser-selection excesses in §4 are 17, 42 and 29 percentage points, with 95% intervals 13–22, 36–47 and 22–35. Intervals use 8,000 within-model game resamples and sorted endpoints at indices 200 and 7800, with the original shared Python RNG seed 20260907 and recorded file traversal order. They do not account for dependence across configurations sharing a seed. Source: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py.

B.9 Persistence covariation under three difficulty proxies

Spearman ρ between per-configuration w→w and final MRR (n = 11), and rank partial correlations controlling for one proxy of configuration difficulty: the co-run tally's MRR (the tally rule), the tracker rule's MRR (Lv.3), or chance level. The proxies were not preregistered; all three are reported. The second block restricts each configuration to games sealed after round 1.

Agent	ρ	**partial, tally**	**partial, tracker**	**partial, chance**	ρ **(seal** $\geq$ **2)**	**partial, tally**	**partial, tracker**	**partial, chance**
Llama-3.1-70B	−0.891	−0.619	−0.720	−0.912	−0.891	−0.619	−0.720	−0.912
GPT-5.6	−0.917	−0.101	−0.693	−0.898	−0.917	−0.292	−0.668	−0.912
Gemini 3.1 Pro	−0.982	−0.926	−0.963	−0.984	−0.982	−0.902	−0.953	−0.979
DeepSeek V4	−0.927	−0.746	−0.827	−0.932	−0.927	−0.808	−0.855	−0.941
Tally rule	−0.455	(its own MRR)	+0.343	−0.518				

The seal-restricted columns drop games whose final pick already stood at the end of the opening round: 26 games of 539 for Llama-3.1-70B, 272 for GPT-5.6, 63 for Gemini 3.1 Pro, and 101 for DeepSeek V4. The unrestricted and restricted rho agree to three decimals in all four models: the sums of squared rank differences are unchanged, although the individual rank differences change in three models. This numerical agreement does not establish invariance to the exclusion. The partial correlations differ because the rank relationships with the proxies change.

Source: results/peer1_analyses_2026-08-30 (b11_conditioning_map, b11_202_control_decomp_setsize, b11_202_neighbor_refs, r3_p5_fallback_hitdecomp_seal, r4_hitdecomp_coldstart).

B.10 In-view games split by a delivered adversary accusation

This post-hoc analysis asks whether final accuracy differs between in-view games in which the adversary did or did not accuse the evaluated seat before its final judgment. Only accusations generated in rounds earlier than that judgment enter the split. Cells give the game count and final top-1 rates for the model and the co-run tally on the same seeds; the two runs can have different honest-agent messages. A rate is shown as n/a when its group contains no games. The adversary's statements and the accusation split agree across all four models and the co-run records.

The between-group contrast is the model's top-1 rate in accused games minus its rate in unaccused games, not a model–rule contrast. Following the original sample-size screen, intervals are reported only when both groups have more than 10 games: eight cells across the two hub configurations qualify. Approximate 95% intervals use the unpooled binomial standard error and are not adjusted for multiple comparisons. Two exclude zero: Gemini at hub N10, $+0.286$ [$+0.020$, $+0.552$], and DeepSeek at hub N15, $+0.479$ [$+0.232$, $+0.726$]. These are selected, within-configuration associations, not causal effects of receiving an accusation. Smaller groups are descriptive only; their size does not establish no difference or equality.

Configuration	Llama-3.1-70B: accused n (model / co-run) · not accused n (model / co-run)	GPT-5.6: accused n (model / co-run) · not accused n (model / co-run)	Gemini 3.1 Pro: accused n (model / co-run) · not accused n (model / co-run)	DeepSeek V4: accused n (model / co-run) · not accused n (model / co-run)
isolated pairs N10	7 (1.000 / 1.000) · 0 (n/a)	7 (1.000 / 1.000) · 0 (n/a)	7 (1.000 / 1.000) · 0 (n/a)	7 (1.000 / 1.000) · 0 (n/a)
isolated pairs N20	4 (1.000 / 1.000) · 1 (1.000 / 1.000)	4 (1.000 / 1.000) · 1 (1.000 / 1.000)	4 (1.000 / 1.000) · 1 (1.000 / 1.000)	4 (1.000 / 1.000) · 1 (1.000 / 1.000)
ring N10	8 (0.500 / 0.625) · 4 (0.250 / 1.000)	8 (0.625 / 0.625) · 4 (0.250 / 1.000)	8 (0.500 / 0.625) · 4 (0.000 / 1.000)	8 (1.000 / 0.625) · 4 (0.000 / 1.000)
ring N15	2 (0.500 / 1.000) · 4 (0.000 / 0.750)	2 (1.000 / 1.000) · 4 (0.500 / 0.750)	2 (1.000 / 1.000) · 4 (0.250 / 0.750)	2 (1.000 / 1.000) · 4 (0.500 / 0.750)
small world N10	10 (0.500 / 0.400) · 12 (0.750 / 0.417)	10 (0.600 / 0.400) · 12 (0.667 / 0.417)	10 (0.500 / 0.400) · 12 (0.500 / 0.417)	10 (0.700 / 0.400) · 12 (0.583 / 0.417)
small world N15	4 (0.250 / 0.500) · 7 (0.000 / 0.429)	4 (0.500 / 0.500) · 7 (0.286 / 0.429)	4 (0.750 / 0.500) · 7 (0.286 / 0.429)	4 (0.500 / 0.500) · 7 (0.429 / 0.429)
cycle heavy N10	10 (0.700 / 0.700) · 22 (0.727 / 0.500)	10 (0.700 / 0.700) · 22 (0.500 / 0.500)	10 (0.800 / 0.700) · 22 (0.409 / 0.500)	10 (0.800 / 0.700) · 22 (0.591 / 0.500)
cycle heavy N15	4 (1.000 / 0.750) · 16 (0.375 / 0.625)	4 (0.500 / 0.750) · 16 (0.500 / 0.625)	4 (0.750 / 0.750) · 16 (0.438 / 0.625)	4 (0.500 / 0.750) · 16 (0.312 / 0.625)
cycle heavy N20	5 (0.800 / 1.000) · 11 (0.455 / 0.545)	5 (0.800 / 1.000) · 11 (0.636 / 0.545)	5 (1.000 / 1.000) · 11 (0.545 / 0.545)	5 (1.000 / 1.000) · 11 (0.727 / 0.545)
hub N10	27 (0.778 / 0.630) · 22 (0.591 / 0.591)	27 (0.667 / 0.630) · 22 (0.545 / 0.591)	27 (0.741 / 0.630) · 22 (0.455 / 0.591)	27 (0.741 / 0.630) · 22 (0.500 / 0.591)
hub N15	30 (0.333 / 0.400) · 19 (0.158 / 0.579)	30 (0.533 / 0.400) · 19 (0.474 / 0.579)	30 (0.633 / 0.400) · 19 (0.421 / 0.579)	30 (0.900 / 0.400) · 19 (0.421 / 0.579)

Separately, over all five generated rounds and all games in each configuration, the fraction of adversary statements targeting the evaluated seat (accusations or defenses; identical across models) is: isolated pairs N10 0.143 · isolated pairs N20 0.102 · ring N10 0.139 · ring N15 0.057 · sw N10 0.159 · sw N15 0.082 · cycle N10 0.151 · cycle N15 0.078 · cycle N20 0.061 · hub N10 0.404 · hub N15 0.469.

Sources for the split and its receipt-time reanalysis: results/peer1_analyses_2026-08-30/r2c_premise_split.py, r6d_c6_c10.py; results/verify_2026-09-04/accuser_scope_received_2026-09-20.py (`accscope`). The generated-statement fractions retain their separate whole-game definition.

§3.5's ring exception is read off this comparator. On the same in-view games the model is below the co-run tally in all eight ring cells in top-1 and in seven in reciprocal rank; against the shadow tally, which reads the seat's own stream, it is below in five of the eight on either metric, and the ring $N = 10$ in-view mean is $+0.006$ in the model's favor (B.14). The exception is therefore a co-run, top-1 statement. Source results/verify_2026-09-04/ring_inview_2026-09-14.py.

B.11 Premise deduction in event reasoning

An accuser is a neighbor that accused the seat before its final judgment. Share of accuser-ending games across all configurations, without B.31's configuration eligibility filter, whose reasoning text matches at least one of three patterns: A, citing the prompt's rule that investigators always tell the truth; B, deducing from the seat's own innocence that an accusation of it is false; C, concluding that only the criminal can lie or that the accuser must be the criminal. Unit = game (any round). Cross-tab over all accused games: rate of ending on an accuser given a match versus no match.

Model	Events	Any pattern	A	B	C	P(accuser first | match)	P(accuser first | no match)
Llama-3.1-70B	435	0.087	0.002	0.085	0.002	0.950	0.865 ($n = 459$)
GPT-5.6	426	1.000	0.862	0.911	0.925	0.928	0.000 ($n = 1$)
Gemini 3.1 Pro	369	0.862	0.659	0.520	0.604	0.850	0.761 ($n = 67$)
DeepSeek V4	408	0.980	0.966	0.657	0.836	0.816	1.000 ($n = 8$)

The Events column uses all configurations (426, 369 and 408 held-out events); B.31 uses eligible configurations (380, 369 and 362). Conditional selection rates use all accused games, separately

conditioned on a text match or no match; parenthetical n counts the no-match group. Text from the final judgment round is retained. A match to any one pattern does not validate a complete deduction or the internal decision process.

Source: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py. The analysis is post-hoc and exploratory.

QUALITATIVE-CODING PROCEDURE

B.4 identifies the annotators, protocol, agreement and scope of the exploratory second pass. Model and condition identifiers were withheld from the coder file; this does not establish hypothesis blinding of the sessions.

B.12 RE-EXECUTION SPREAD BY CONFIGURATION

configuration	SE of one run	sign threshold (1.96 · SE)	seeds
isolated pairs, N = 10	0.0072	±0.0140	49
isolated pairs, N = 20	0.0105	±0.0207	49
small-world, N = 10	0.0368	±0.0722	49
hub-and-spoke, N = 10	0.0391	±0.0766	48–49
cycle-heavy, N = 10	0.0441	±0.0865	49

Each configuration was re-run once under its registered settings, on the same seeds, for the three API models. Pairing on seeds, $\mathrm{SE}(\Delta)$ is the standard deviation of the per-seed differences over the square root of the number of seeds; under the working assumption of equal, independent run noise, $\mathrm{Var}(\Delta) = 2\,\mathrm{Var}(\text{one run})$, giving the reported single-run SE as $\mathrm{SE}(\Delta)/\sqrt{2}$ and the threshold as 1.96 times that quantity. The assumption is not established by one rerun. Each configuration takes the largest SE among its three models, a conservative choice. **Each SE is itself estimated from a single re-run and carries its own sampling error.** No Llama re-execution records are available; applying these API-model thresholds to Llama is an extrapolation. Source: results/verify_2026-09-04/reexecution_se_2026-09-19.py, reproducing the original rM_t3_kband.py thresholds. The cell comparison below uses the retained co-run tally on the same seeds.

At fixed N = 10 the isolated-pairs configuration is five to six times quieter than the other three, so the spread tracks topology rather than size; doubling N raises its SE by half. Among the 20 model–configuration cells in these five configurations, 8 have an absolute model–co-run gap below the corresponding threshold, including one Llama cell. Extending to all 44 cells by assigning either the smallest or the largest measured threshold to each unmeasured configuration gives 9 cells in both cases: the additional cell is Llama at hub N15. None is in the isolated-pairs family. These counts include threshold extrapolations; they are not cell-specific confidence intervals.

B.13 THE REASONING-MODE ARM

DeepSeek V4 with its separately billed thinking mode on against the same model with it off, same seeds, same configurations. §4.2 summarizes the scope of this comparison. The mode is a protocol bundle rather than a reasoning-only contrast (it silently overrides the sampling temperature and moves related settings together) so what this bounds is the size of any rescue, not its mechanism. Source: results/verify_2026-09-04/thinking_arm_2026-09-05.py.

configuration	n	thinking on	off	difference	95% CI
isolated pairs, $N = 10$	49	0.3189	0.3310	−0.0121	[−0.0481, +0.0198]
small-world, $N = 15$	49	0.2945	0.3434	−0.0489	[−0.1320, +0.0305]
cycle-heavy, $N = 15$	49	0.4051	0.3446	+0.0605	[−0.0446, +0.1695]
hub-and-spoke, $N = 10$	49	0.7660	0.7649	+0.0010	[−0.1142, +0.1081]
pooled	**196**			**+0.0001**	**[−0.0439, +0.0438]**

Two configurations move each way and every interval spans zero. The pooled interval constrains the average MRR change under this protocol bundle on these configurations. It does not directly

constrain the change in the 0.174 audibility contrast, which is a different estimand on a different population, or rule out heterogeneous responses.

B.14 The audibility effect by configuration and model

Within each configuration, games are split by whether the adversary is among the evaluated seat's neighbors, and the model's deficit against the shadow tally is differenced across the two halves. Topology family and size are held fixed within a configuration; graph draws and role assignments can vary across seeds. This is a within-configuration contrast, not an isolated intervention on adversary position. The two hub configurations are excluded because no game there falls outside the seat's view. Columns are in the order used throughout. §3.5 cites the mean, its interval and the two counts. The mean is a difference of differences; the two halves on the same resampling unit are -0.160 $[-0.209, -0.104]$ out of view and $+0.014$ $[-0.038, +0.057]$ in view. Sources: results/original_analyses_2026-09-04/audibility_2026-09-04.py; results/verify_2026-09-04/audibility_levels_2026-09-10.py.

configuration	n in	n out	Llama	GPT-5.6	Gemini	DeepSeek	mean
isolated pairs, $N = 10$	7	42	-0.048	+0.063	+0.083	+0.041	+0.034
isolated pairs, $N = 20$	5	44	+0.014	+0.043	+0.036	+0.029	+0.031
ring, $N = 10$	12	37	+0.134	+0.124	+0.201	+0.341	+0.200
ring, $N = 15$	6	43	-0.025	+0.005	+0.061	+0.247	+0.072
small-world, $N = 10$	22	27	+0.489	+0.276	+0.382	+0.334	+0.370
small-world, $N = 15$	11	38	+0.235	+0.172	+0.103	+0.396	+0.227
cycle-heavy, $N = 10$	32	17	+0.520	+0.214	+0.233	+0.338	+0.326
cycle-heavy, $N = 15$	20	29	+0.193	-0.036	-0.055	+0.023	+0.031
cycle-heavy, $N = 20$	16	33	+0.137	+0.288	+0.251	+0.427	+0.276
mean over configurations, bootstrapped over configurations ($k = 9$)							**+0.174**
95% CI							**[+0.093, +0.259]**

All nine configuration means are positive, and they average to the reported 0.174; the seven rows that carry a relay path average 0.215, which is the figure §3.5 reports for dropping the isolated pairs. Here score ties are resolved by ascending player identifier; the 0.296 below uses reciprocal midranks on the same seven configurations. Four of the thirty-six cells are not, and they are named here rather than only counted: Llama at isolated pairs $N = 10$ and ring $N = 15$, and GPT-5.6 and Gemini at cycle-heavy $N = 15$. The interval resamples configurations rather than cells, since cells within a configuration share its graph and its seeds.

The resampling unit. The interval above resamples nine configurations, which are size variants of four topology families. With so few units, nominal coverage is uncertain. Three alternative interval calculations also exclude zero, under different assumptions and weights. Resampling the four families gives $+0.174$ $[+0.084, +0.258]$; each sampled family contributes all its configurations, so families with three configurations receive more weight than those with two. Giving the four family means equal weight instead yields $+0.170$. A t interval over the same nine configuration means, with $s = 0.135$ and $t_8 = 2.306$, is $+0.174$ $[+0.070, +0.278]$. Resampling the 49 seeds across all models and configurations gives $+0.168$ $[+0.104, +0.245]$. This last statistic pools games separately within the in-view and out-of-view groups. Each model–configuration cell has 49 games, but the group sizes differ across configurations, changing their weights relative to the primary equal-configuration contrast. These checks do not establish nominal coverage or remove the small number of topology families. Source: results/verify_2026-09-04/split_interval_alternatives_2026-09-11.py.

The two tie treatments, and the relay-only figure. The 0.174 estimate retains all 49 games per model and configuration, including tied and all-zero tally scores, and resolves score ties by ascending player identifier. A separate sensitivity calculation replaces the shadow adversary's reciprocal rank with the reciprocal of its midrank: $1/[a + (g + 1)/2]$, where a candidates have strictly higher scores and g candidates share its score, including the adversary. This is not the expected reciprocal rank under random tie-breaking, which averages the reciprocals of the tied positions. The model's scored rank is unchanged. On the seven configurations with relay paths, this alternative gives 0.296 (95% CI 0.213 to 0.366); on all nine it gives 0.423. Thus the direction persists, but its

magnitude depends on the tie definition and configuration set. The primary 0.174 contrast and its seven-configuration counterpart, 0.215, use the identifier-based rule. Source: results/verify_2026-09-04/top1_versions_2026-09-08.py, T3.

B.15 Neighbor occupancy of the top slots

Restricted to games where the adversary is *not* a neighbor of the evaluated seat (so every neighbor in the top slots is known to be wrong) the share of the seat's top k ranks filled by neighbors, where k is that seat's degree in that game. The random-ranking reference is the expected share of neighbors under a uniformly random ranking, $k/(N-1)$, averaged across these games. The fourth column repeats the measure after a heuristic exclusion of suspected parser-filled entries: a trailing sequence of at least two entries at confidence 5 in ascending seat-number order. This heuristic does not identify actual parser additions with certainty. Source: results/verify_2026-09-04/neighbor_occupancy_2026-09-04.py.

occupant	n	neighbor share	suspected fills dropped	tally rule	random ranking
Llama-3.1-70B	310	97.7%	97.7%	61.1%	25.2%
GPT-5.6	310	83.3%	83.3%	61.1%	25.2%
Gemini 3.1 Pro	310	74.8%	78.3%	61.1%	25.2%
DeepSeek V4	310	90.6%	90.7%	61.1%	25.2%

The filtered means retain only games with at least k entries after removal: 308, 310, 242 and 309 games in row order. The changes are 3.5 percentage points for Gemini and less than 0.14 points for each other model. Because the game populations also change, this check does not isolate the effect of parser filling. The tally rule is co-run and identical across rows because it does not depend on the model. §3.5 reports the excess over the co-run tally.

Section 2's MRR sensitivity keeps explicitly named seats in output order and assigns an unnamed adversary the mean rank of the remaining slots. Its mean change ranges from -0.0065 to $+0.0007$ over the four models. Because the stored ranking sorts by confidence, this changes named-seat sorting as well as missing-seat treatment; it does not isolate the effect of filling. Source results/verify_2026-09-04/fill_scoring_sensitivity_2026-09-04.py.

B.16 An oracle-informed reference on the seat's stream

The reference combines two environment-specific channels: truthful honest opening statements and the post-opening defense-speaker signature. It uses neighbor statements generated before the final judgment, retains the historical scoring formula and lowest-ID tie rule, and does not use graph structure beyond selecting the seat's neighbors. Neither occupant is told both conventions. This is a specified oracle-informed comparator, not a proved optimal rule or attainable ceiling. Rates are top-1, pooled over four models; the tally column is the shadow comparator. Source: results/verify_2026-09-04/timing_corrected_2026-09-20.py; the original formula is retained in results/recovered_analyses/1c89504a_L13936_258c95bc5e.py.

adversary	true opening accusation in the game	n	model	tally	reference
in view	yes	524	72.1%	76.7%	95.4%
in view	no	392	40.3%	25.5%	92.9%
out of view	yes	652	10.9%	39.4%	40.5%
out of view	no	588	1.0%	2.2%	10.2%
in view	both	**916**	**58.5%**	**54.8%**	**94.3%**
out of view	both	**1,240**	**6.2%**	**21.8%**	**26.1%**
all	both	2,156	28.4%	35.8%	55.1%

The bold rows support the comparison in §4.2. The model's shortfall against this reference is 35.8 percentage points in view and 19.9 out of view; against the tally rule the differences are $+3.7$ and -15.6. The larger shortfall therefore changes sides with the comparator.

Why this is not an upper bound. Restricting the same reference to the shadow tally's final two-round statement window (rounds three and four) changes its top-1 from 94.3% to 77.3% in view

and from 26.1% to 7.1% out of view. The LLM can also carry earlier information in its ranking state. Out of view, the shadow tally attains 306.8% of this window-restricted reference's top-1, so the reference does not bound the tally rule. A score divided by this reference cannot be interpreted as the fraction of attainable performance recovered. The comparison describes the specified rules and their available channels. Source: timing_corrected_2026-09-20.py.

B.17 The one-hop-to-two drop, by configuration

Per configuration, the decrease in the model-minus-shadow MRR gap from one-hop to two-hop games. A positive value means a larger model shortfall at two hops. Seeds, not games, are the resampling unit: the four models share each seed's graph and adversary placement, so a seed fixes the hop and its four games are one cluster; the four values are averaged within a seed. Configurations need at least six seeds on both sides to appear, which admits seven of eleven. §4.2 refers to this distance analysis. Source: results/verify_2026-09-04/hop_step_cluster_2026-09-07.py.

configuration	**seeds at 1**	**at 2**	**drop**	**95% CI**
small-world, $N = 10$	22	23	+0.361	[+0.194, +0.530]
cycle-heavy, $N = 10$	32	17	+0.326	[+0.175, +0.486]
cycle-heavy, $N = 20$	16	15	+0.318	[+0.135, +0.516]
ring, $N = 10$	12	11	+0.253	[+0.005, +0.502]
small-world, $N = 15$	11	24	+0.221	[−0.007, +0.435]
ring, $N = 15$	6	9	+0.042	[−0.197, +0.298]
cycle-heavy, $N = 15$	20	22	+0.020	[−0.112, +0.145]
mean over configurations			**+0.220**	**[+0.121, +0.308]**

Every configuration is positive and four intervals exclude zero. The mean's interval resamples configurations, the unit the audibility effect of §3.5 also uses; note that these seven are size variants of three topologies, so they are not seven independent trials. Beyond two hops we make no claim: the second-difference contrast that would separate a step from a constant slope has an interval excluding zero in one configuration of five under the same clustering, and we report its direction only.

B.18 Seat swap: hub to spoke, topology held fixed

Per model, final MRR at the hub seat and at a spoke seat of the same graph, same seeds, hub-and-spoke at $N = 15$ and $N = 20$; *cost* is hub minus spoke. The rule row is the tally rule run in those same two seats under the same seeds, so it is a co-run comparison and moves with the seat as the model does. This table reports the model ranges and the rule's costs. Source: results/verify_2026-09-04/seat_swap_extension_2026-09-04.py.

	$N = 15$			$N = 20$		
occupant	**hub**	**spoke**	**cost**	**hub**	**spoke**	**cost**
Llama-3.1-70B	0.5558	0.3404	0.2153	0.7745	0.4174	0.3572
GPT-5.6	0.6624	0.4299	0.2325	0.8025	0.4285	0.3740
Gemini 3.1 Pro	0.6743	0.3807	0.2936	0.7979	0.3429	0.4551
DeepSeek V4	0.7988	0.4894	0.3094	0.8929	0.4447	0.4482
the tally rule (co-run)	cost 0.0289			cost 0.0701		

Forty-nine seeds per cell. The cost is positive for every model at both sizes, eight of eight cells, and is seven to eleven times the rule's at $N = 15$ and five to six times it at $N = 20$. The swap moves the seat away from the adversary as well as away from what it hears, and this design does not separate those two.

B.19 How firm the cell counts are

The cell map's two counts (24 of the 37 cells below the co-run tally with a paired bootstrap interval excluding zero, and one of the seven above it) are counts of intervals, and ten of the thirty-seven have an endpoint within 0.02 of zero, three within 0.003. These counts and the seven intervals below use the reproduction registry's `hubwins` script, with 4,000 paired resamples per cell and random seed 20260904. The registry's `cells` script also resamples a parser-filtered variant between successive

cells, consuming the random stream differently; it gives 25 rather than 24 intervals below zero. All forty-four point estimates and their signs agree between these two scripts. The seven cells above the rule:

model	configuration	model − co-run tally [95% CI]
Llama-3.1-70B	hub N10	+0.114 [−0.021, +0.246]
GPT-5.6	hub N10	+0.041 [−0.051, +0.138]
GPT-5.6	hub N15	+0.094 [−0.001, +0.192]
Gemini 3.1 Pro	hub N10	+0.007 [−0.101, +0.118]
Gemini 3.1 Pro	hub N15	+0.106 [−0.011, +0.228]
DeepSeek V4	hub N10	+0.071 [−0.062, +0.202]
DeepSeek V4	hub N15	+0.230 [+0.077, +0.376]

The shadow cell map, by metric. Against the shadow tally, where both judges read the same statements, the model is lower in 32 cells, tied in 5 and higher in 7 in top-1, and all seven of those are hub-and-spoke. In reciprocal rank the same map gives 34, 0 and 10, and three of the ten sit outside that family: isolated pairs $N = 10$ (Llama-3.1-70B) by 0.041, small-world $N = 10$ (GPT-5.6) by 0.021 and cycle-heavy $N = 10$ (GPT-5.6) by 0.009, each an exact tie in top-1. §3.1 reports the reciprocal-rank map. Source results/verify_2026-09-04/shadow_cellmap_2026-09-10.py.

Why the two comparators disagree in isolated pairs. The cell map has the model below the co-run tally in both isolated-pairs cells, while the tie-policy appendix has it above the shadow tally there. The channel is the opening round: the co-run rule occupies the seat, states something true about a target drawn at random, and its lone neighbor echoes that statement back to it. In the three of 49 games at $N = 10$ where that opening named the adversary the rule finishes at 1.000; in the other 46 it averages 0.352 against a chance level of 0.314. The opening-hit count is 1.11 standard deviations below its expectation (three versus 5.4), and the remaining games' mean is 0.81 standard errors above chance (0.352 versus 0.314 on 46 games). These descriptive checks are consistent with an echo account; they do not establish that it is the only source of the score. The shadow rule reads the echo of the *model's* opening statement instead and has no such oracle. Exploratory; source results/verify_2026-09-04/hubwins_and_corun_oracle_2026-09-09.py.

B.20 Analysis registry: preregistered and exploratory

The reproduction registry identifies computations; it is not itself a preregistration. We distinguish protocols fixed before new data generation, staged analyses fixed before specified statistics were unsealed on existing games, and later exploratory analyses. The local documents, repository history and recovered receipt records establish the source of each commitment; a date or a script alone does not establish that data were unseen. Descriptive and exploratory results can support the paper's interpretation, but are not relabeled confirmatory because their computation was recorded before a reread.

Analysis	Timing and scope	Source and limits
Co-run cell count	Cross-model prediction P1 fixed before held-out campaign calls; Llama is discovery data	Recovered cross-model protocol, dated July 13, with July 21 receipt provenance. It defines the per-model count criterion, not every later contrast.
Behavioral screening and N5 confirmation	Criteria fixed before named statistics were unsealed on existing games; confirmation follows screening	Screening v3.2 and N5 v1.2, not ablation v1–v5. The N5 pooled test is confirmatory under this staged protocol; subgroup profiles are descriptive.
Audibility contrast, transplant decomposition and oracle reference	Later diagnostics of the campaign records; no pre-data registration is claimed here	Named audibility, splice and bound scripts identify the computations. Ablation v1–v5 does not register these analyses.
Block ordering, text comentions and relayed top-1 rates	Exploratory analyses and descriptive rates on explicitly selected populations	B.30, B.39 and the relayed-first script. Criteria fixed before a reread do not make these new-data confirmation.
Corrected-premise and other ablation arms	Arm-specific prospective commitments, with amendments recorded separately	Ablation v1–v5 and Appendix E. The broader coverage counts and current matched cross-model readout are later analyses, not all governed by the original arm registration.
Tie-policy experiments	Initial policy protocol and a separately committed relay-arm rule; amendments have their own timing	Appendix G; the relay-arm rule predates launch, with its judgeability amendment made during the run before recorded outcome inspection. Derived gap and absolute-level summaries retain their separate status.
Roster interventions and qualitative coding	R has its own staged arm registration; the A/R/C placement reanalysis and qualitative recoding are exploratory	R source snapshots and the placement check in Appendix E; coder identities and limits in B.4. The later analyses do not inherit arm preregistration.

Traceable chronology. The cross-model source is `prereg_crossmodel_2026-07-13_RECOVERED.md`; its provenance records receipt before the held-out calls. For the behavioral search, the handoff tag, the v3.2 amendment, the N1 template and the screening readout are each recorded as a separate commit. N5 confirmation v1.2 is frozen in one commit, followed by its result in another. The relay tie-arm rule is committed before launch; its power-clause amendment follows. These are distinct histories. The supplementary source map, `results/analysis_registry_2026-09-19/analysis_lineage.md`, gives each commit with its timestamp and links their documents and snapshots. The hash check verifies supplied text, not historical non-access to data. The ablation lineage in Appendix E governs those arms only.

B.21 Accuser selection and games without a delivered accusation

An accuser here is a neighbor that accused the evaluated seat before its final judgment (generated round $r < T$). The shadow tally still scores only rounds $T-2$ and $T-1$. For a direct comparison of accuser selection, the table restricts both model and rule to the same games with a unique shadow top candidate. Entries are the fraction ending on an accuser. This selected population excludes shadow ties and differs across models.

Model	out n	model	shadow	in n	model	shadow
Llama-3.1-70B	155	0.845	0.290	186	0.634	0.543
GPT-5.6	174	0.833	0.655	194	0.629	0.521
Gemini 3.1 Pro	223	0.632	0.157	193	0.648	0.492
DeepSeek V4	139	0.691	0.187	179	0.615	0.514

Across all out-of-view games, 91.3% contain a delivered accusation of the seat by a neighbor; the adversary cannot be that accuser because it is not a neighbor. In view, the adversary accuses the seat before judgment in 48.5% of games. These rates use all games, not the unique-shadow-top subset above.

Restricting instead to out-of-view games with no delivered accusation from any neighbor gives the following model-minus-shadow top-1 gaps, with expected top-1 credit for shadow ties. The game sets differ across models. Intervals use 8,000 paired-game bootstrap resamples, retaining the producing script's model/stratum order and Python seed 20260909, with sorted endpoints at indices 200 and 7,800.

Model	games	gap [95% CI]
Llama-3.1-70B	15	-0.356 [-0.589, -0.156]
GPT-5.6	25	-0.120 [-0.260, $+0.020$]
Gemini 3.1 Pro	49	-0.160 [-0.265, -0.061]
DeepSeek V4	19	-0.254 [-0.447, -0.070]

The intervals exclude zero for Llama, Gemini and DeepSeek. This shows a deficit on these selected samples without a delivered accusation; it does not identify how much of the overall deficit accusations cause. The separate top-1 contrast across in-view and out-of-view games cannot be formed under its minimum-size rule: no configuration has at least 30 unaccused observations on both sides, even pooling the four models. The pooled observations share seeds and are not independent games.

Over all out-of-view games, model accuser-selection rates range from 0.826 to 1.000 at degree one, 0.639 to 0.964 at degree two, and 0.645 to 0.809 at degrees three to seven. These degree strata do not control the information or evidence available in a game. All analyses here are exploratory. Source: results/verify_2026-09-04/accuser_scope_received_2026-09-20.py (`accscope`), following the definitions and bootstrap order in accuser_resolve_2026-09-08.py, tie_trap_2026-09-09.py and accfirst_by_degree_2026-09-15.py.

B.22 The corrected premise on the out-of-view half, by cell

Paired on seed within each cell, out-of-view games only: the model's top-1 minus the shadow tally's, ties counted by expectation, under the registered prompt and under the corrected premise. Exploratory. Source results/verify_2026-09-04/reviewer_round4_2026-09-08.py.

Coverage of the corrected-premise arm. The arm runs in 35 of the 44 cells (10 on GPT-5.6, 9 each on DeepSeek V4 and Llama-3.1-70B, 7 on Gemini 3.1 Pro) and 33 of those have at least 45 of the 49 seeds on both sides of the comparison; two Gemini cells were left short when a lane stopped on a spend cap. The sign counts §4 reports are over those 33, and they are these: the cell's sign against the co-run tally is preserved in 32 of the 33 and against the shadow tally in 28, so one cell changes side under the first comparator and five under the second. Over all 35 cells, including the two the spend cap left short, the counts are 34 and 29. Six of the cells are the temperature-matched reruns, substituted because the frozen runner sets 0.0 for GPT-5.6 alone. Source results/verify_2026-09-04/premise_arm_coverage_2026-09-07.py.

model	configuration	games	registered	corrected	corrected − registered [95% CI]
Llama-3.1-70B	isolated pairs N10	42	-0.053	-0.108	-0.056 [-0.127, $+0.000$]
GPT-5.6	isolated pairs N10	42	-0.175	-0.095	$+0.079$ [$+0.026$, $+0.151$]
GPT-5.6	ring N10	37	-0.204	-0.234	-0.030 [-0.122, $+0.062$]
GPT-5.6	small world N15	38	-0.118	-0.150	-0.031 [-0.076, $+0.004$]
Gemini 3.1 Pro	isolated pairs N10	42	-0.151	-0.015	$+0.136$ [$+0.028$, $+0.251$]
DeepSeek V4	isolated pairs N10	42	-0.153	-0.050	$+0.103$ [$+0.005$, $+0.230$]

Three of these cells ran at a temperature the baseline does not use. The corrected-premise campaign was extended in stages, and this table is the earlier stage. Its frozen runner pins temperature 0.0 for GPT-5.6 alone; for the other three models the cells it produced — isolated pairs and hub-and-spoke at $N = 10$ — take the client default of 0.7, while every baseline cell is 0.0. Paired on seed against a 0.0 baseline, such an arm mixes the premise correction with a sampling change, and those games record no sampling parameters, so the mismatch is visible only in the code path. All six were re-run at 0.0 into a separate arm so the original stays readable, with the sampling written into every game's manifest; the three isolated-pairs cells are the ones the table below compares. The criterion below was fixed before the re-run existed: the sign, and whether half the magnitude survives. The later relay extensions for DeepSeek, Llama and Gemini ran at temperature 0.0 from the start and record sampling, model and run metadata in per-game manifests; those extensions, not the runs described here, supply the non-GPT rows analyzed in E.6.

model	as run (0.7)	re-run at 0.0
Llama-3.1-70B	-0.056 [-0.127, $+0.000$]	-0.011 [-0.026, $+0.005$]
Gemini 3.1 Pro	$+0.136$ [$+0.029$, $+0.251$]	$+0.148$ [$+0.061$, $+0.249$]
DeepSeek V4	$+0.103$ [$+0.005$, $+0.228$]	$+0.008$ [-0.066, $+0.085$]

Gemini's gap reduction survives and widens. For DeepSeek, eight percent of the original point estimate remains and the interval covers zero in the matched-temperature comparison. GPT-5.6's cell was at 0.0 throughout. Source: results/verify_2026-09-04/cprem_temperature_2026-09-12.py, which reuses the first script's parsing and scoring functions and recomputes the intervals.

Both tables use 8,000 paired-game bootstrap resamples and sorted endpoints at zero-based indices 200 and 7,800. The first table uses the full producing script's shared Python RNG stream seeded at 20260908; the temperature table resets it to 20260913 and processes Gemini, DeepSeek and Llama, each at 0.7 then 0.0. The repeated 0.7 estimates use the same game-level quantities but different bootstrap draws, so their printed endpoints can differ.

The two relay rows in the table above are GPT-5.6 rows; they do not describe the relay coverage of the later extension analyzed in E.6. Once temperature is matched the deficit closes in two of the four isolated-pairs cells (GPT-5.6 and Gemini) where nothing is relayed and the seat's lone neighbor is echoing its own words. The self-accuser rate, measured on the arm before the temperature match, goes 0.898 to 0.861 (Llama-3.1-70B), 0.876 to 0.820 (GPT-5.6), 0.715 to 0.206 (Gemini 3.1 Pro) and 0.797 to 0.793 (DeepSeek V4), over 266, 266, 165 and 266 out-of-view games. These rates do not rule out that account: the two largest falls occur in the two cells that close. Because the rates predate the temperature match and these arms do not separate the two, we make no claim from them.

B.23 Self-ranking after a delivered accusation

We search each model's raw Call2 text for a first-ranked player, over 2,695 rounds per model. A prior accuser is a neighbor whose accusation of the evaluated seat was generated in a strictly earlier round and thus delivered before the current judgment. DeepSeek V4 names itself first in 156 of 1,491 rounds with a prior accusation and in none of 1,204 rounds without one. GPT-5.6 has no detected self-ranking; Llama-3.1-70B has two, one without a prior accusation; Gemini 3.1 Pro has sixteen, seven without a prior accusation. The raw-text extractor finds no first-player match in 133 Gemini rounds, which remain in the denominator as rounds without detected self-ranking. These are extraction-based counts and temporal associations, not evidence that an accusation caused self-ranking or that the model followed a particular deduction. The raw text is required because the parser removes the evaluated seat from the stored ranking. Exploratory. Source: results/verify_2026-09-04/selfranking_received_2026-09-20.py (`selfreceived`).

B.24 The ranking of the non-neighbors alone, against chance

Out-of-view games, hub configurations excluded, 310 per model. The sub-ranking is the stored confidence-sorted ranking, including parser-added candidates, restricted to the seats that are not the evaluated seat's neighbors, scored by reciprocal rank of the adversary within that restriction; chance is the mean reciprocal rank of a uniformly random order of the same block, 0.3214. Every model is above it: 0.4174 [$+0.0592$, $+0.1340$], 0.4114 [$+0.0537$, $+0.1270$], 0.3744 [$+0.0196$, $+0.0884$] and 0.4179 [$+0.0597$, $+0.1338$], intervals on the difference from chance, in Llama, GPT-5.6, Gemini and DeepSeek order. B.30 compares non-neighbor ranks with the tally on a different, accusation-conditioned subset, using normalized confidence ranks rather than this full-sample reciprocal-rank measure (§3.2). Source results/verify_2026-09-04/subranking_hops_hub_2026-09-04.py.

B.25 Placement against ordering: the transplant decomposition

A final ranking decomposes exactly into which slots the non-neighbors occupy and how they are ordered among themselves, and the two components can be exchanged between the model's ranking and the rule's on the same game. Averaging over both exchange orders gives each component's share of the deficit (a two-player Shapley value, which for two components is the average of the two orders). Here model order means appearance order in the final output, before confidence sorting; the shadow rule sorts by tally and then seat number. We use out-of-view games with no parser-filled entries, so that coverage is complete: 310, 196, 259, and 198 games, which is 100%, 63.2%, 83.5% and 63.9% of the out-of-view games for GPT-5.6, Gemini 3.1 Pro, DeepSeek V4 and Llama-3.1-70B. Placement takes 76% [65%, 88%], 76% [62%, 94%], 84% [77%, 93%] and 100% [92%, 111%] of a total deficit of 0.100, 0.109, 0.185 and 0.192. The contributions are signed and sum to the total deficit. Their ratio need not lie between zero and one: a negative ordering contribution permits a placement share above 100%, so the interval is printed without clipping.

Reconstructing the rule's own component from the exchange returns its actual ranking in 1,240 of 1,240 games, which is the check that the decomposition is exact rather than approximate. Source results/original_analyses_2026-09-04/splice_filtered_2026-09-04.py.

Rank-definition sensitivity. The primary MRR uses the stored confidence-sorted ranking instead. On the same complete-output games, that order differs in 2, 1, 2 and 1 games in the model order above. Repeating the decomposition with the stored order gives placement shares of 76.01%, 75.99%, 84.56% and 99.89%, against 76.01%, 75.99%, 84.22% and 99.89% in output order. Only DeepSeek's share changes; its total deficit changes from 0.184605 to 0.184026. These are post-hoc point-estimate checks, not new intervals. Thus B.25 decomposes the complete-output subset's gap under a specified order, not the full-campaign MRR gap. Source results/verify_2026-09-04/splice_rank_definition_2026-09-19.py.

Connected configurations and game-level contributions. Two further descriptive checks use the same raw-ranking definition and complete-output filter. Removing the two isolated-pairs configurations leaves 224 out-of-view games per model before filtering. This connected subset does not require an accusation naming the adversary to have arrived. Let D_i be the shadow-minus-model reciprocal-rank gap, and P_i and O_i its signed placement and ordering contributions, with $D_i = P_i + O_i$. The signed share below is $\sum_i P_i / \sum_i D_i$. The absolute share is $\sum_i |P_i| / (\sum_i |P_i| + \sum_i |O_i|)$; it describes contribution magnitudes without cancellation and is not a decomposition of the mean performance gap. The last column counts games with $P_i > O_i$ among games with $D_i > 0$, a separate outcome-conditioned frequency.

Model	Sample	Complete n	Mean D	Signed share	Absolute share	$P > O$ / $D > 0$
GPT-5.6	All	310	0.100	76.0%	68.0%	125/148
GPT-5.6	Connected	224	0.118	84.3%	73.5%	125/143
Gemini 3.1 Pro	All	196	0.109	76.0%	62.0%	63/83
Gemini 3.1 Pro	Connected	119	0.130	85.2%	67.6%	60/75
DeepSeek V4	All	259	0.185	84.2%	75.7%	134/177
DeepSeek V4	Connected	200	0.226	83.4%	75.6%	129/163
Llama-3.1-70B	All	198	0.192	99.9%	77.7%	108/157
Llama-3.1-70B	Connected	162	0.247	95.1%	84.4%	108/145

Placement remains the larger signed contribution after excluding isolated pairs. Opposing ordering contributions do reduce its mean relative to its absolute magnitude: on the full complete-output samples, placement's absolute shares are lower than its signed shares, but all remain above one half. Placement also contributes more than ordering in a majority of the positive-gap games in each model. This does not mean it dominates in the same fraction of all games, or that these contributions identify a cause or an attainable improvement. These are post-hoc descriptive summaries, without new confidence intervals or a confirmatory verdict. Source: results/verify_2026-09-04/placement_distribution_2026-09-20.py (`placedist`).

B.26 The split under alternative decision rules

The same games, two-round window and received statements are scored by six decision rules over nine non-hub configurations, including isolated pairs. Intervals resample configurations. *defmask* places defended candidates last and orders the rest by accusation count; *speaker* places defense speakers first and uses tally scores within each group. *trust* weights each statement from speaker s by $1/(1 + A_s)$, where A_s is the number of accusations against that speaker in the same window. It does not measure the speaker's consistency. *selfpen* and *selfpro* subtract or add one point to speakers who accuse the evaluated seat, whose own candidacy the tally excludes. These rules change how cues are scored but share parts of the tally construction; they are not six independent comparator families. *defmask* and *speaker* additionally exploit known emission rules of this environment. Source results/verify_2026-09-04/family_free_comparators_2026-09-10.py.

judge	cue	Δ in view	Δ out of view	split [95% CI]
tally reference	relayed accusation count	$+0.014$	-0.160	$+0.174$ [$+0.091$, $+0.259$]
defmask	defense decisive	$+0.010$	-0.160	$+0.170$ [$+0.090$, $+0.251$]
speaker	defense speaker, then tally	-0.147	-0.160	$+0.014$ [-0.050, $+0.066$]
trust	accusations against speaker	$+0.008$	-0.154	$+0.162$ [$+0.082$, $+0.248$]
selfpen	accuser of the seat penalized	$+0.178$	-0.202	$+0.380$ [$+0.249$, $+0.534$]
selfpro	accuser of the seat promoted	$+0.045$	-0.130	$+0.175$ [$+0.100$, $+0.254$]

The out-of-view model-minus-rule gap is negative with an interval excluding zero for all six rules. The audibility contrast excludes zero for five. The exception is *speaker*: its out-of-view gap is the same as the tally's, but its in-view gap is also negative, leaving a small contrast whose interval includes zero. Thus the contrast depends on which cues the comparator uses. Penalizing accusers of the evaluated seat deepens the out-of-view gap and widens the contrast; this particular extension does not remove the observed shortfall. It does not rule out other advantages supplied by the environment. The tally's interval is independently resampled here and in B.27, giving $[+0.091, +0.259]$ and $[+0.092, +0.258]$, respectively, against B.14's $[+0.093, +0.259]$; its point estimate is unchanged.

B.27 Alternative outcomes and in-view reference bands

The same nine configurations are evaluated with reciprocal rank, top-1, and log odds of top-1. These are different outcome definitions, not a proof of invariance to arbitrary rescaling. The reciprocal-rank row breaks shadow score ties by player identifier. The other two rows use expected top-1 under uniform tie-breaking: the shadow receives $1/g$ when the adversary is among g tied leaders, and zero otherwise. Log odds is computed from each model–configuration–audibility cell's rates, with p replaced by $(np + 0.5)/(n + 1)$ before taking the logit. Intervals resample configurations. All three contrast intervals exclude zero under these definitions.

The final column retains an exploratory comparison of the in-view 90% percentile interval with a symmetric reference band. Its width is based on the observed out-of-view gap: 0.080 for reciprocal rank (half the previously observed 0.160) and half the top-1 gap computed on these same games. The historical calculation holds these bands fixed while resampling the in-view mean; it does not propagate uncertainty in the out-of-view estimate. The reciprocal-rank interval lies within its band; the top-1 interval does not. We report this containment descriptively, not as confirmatory equivalence or a test that the true in-view gap is less than half the true out-of-view gap. No log-odds band was specified. Source results/verify_2026-09-04/scale_and_equivalence_2026-09-10.py.

scale	Δ in view	Δ out of view	split [95% CI]	in-view interval vs. reference band
reciprocal rank	$+0.014$	-0.160	$+0.174$ [$+0.092$, $+0.258$]	band ± 0.080, 90% [-0.028, $+0.051$]: inside
top-1	$+0.232$	-0.209	$+0.441$ [$+0.241$, $+0.669$]	band ± 0.104, 90% [$+0.038$, $+0.450$]: not inside
log odds	$+1.066$	-1.922	$+2.988$ [$+1.760$, $+4.429$]	no band specified

B.28 The split with the tie convention randomized

Ring $N = 10$ and small-world $N = 15$ provide the two relaying configurations of the tie-policy counterfactual (Appendix G). Each model and arm has 49 games per configuration: 98 pooled games, split into 23 in view and 75 out of view. The statistic is the pooled in-view minus out-of-view mean of model top-1 minus shadow top-1; the shadow averages credit over tied leaders in both arms. It is positive in the randomized arm for all four models ($+0.310$, $+0.136$, $+0.281$ and $+0.742$, in Llama, GPT-5.6, Gemini and DeepSeek order), and in the comparison arm for all four. The shadow's rate of ties for first place falls from 47.8–52.2% to 8.7–30.4% in view and from 32.0–49.3% to 14.7–28.0% out of view; each range spans the four models after pooling the two configurations. These are descriptive point estimates on two configurations. Positive contrasts in both arms do not establish equal magnitudes or exclude a contribution of the tie policy. Source results/verify_2026-09-04/tie_arm_split_2026-09-10.py.

B.29 An opening-round accusation stratum

§3.2 selects out-of-view games in which a neighbor accuses the adversary before judgment. Here we restrict that event to the opening round: an honest neighbor chooses the adversary as its random target and truthfully accuses it. The target draw is fixed by the seat's seeded random generator before it responds to any messages; the accusation is delivered to the evaluated seat at round two. This defines the same 55 configuration–seed pairs in each model. All belong to that model's received-accusation population in §3.2.

The model's stored final top suspect is compared with the shadow tally's expected top-1 under uniform tie-breaking. The shadow still scores only the two statement rounds preceding the final judgment; it does not directly score the opening message. Differences are paired within model, and percentile intervals are obtained by resampling the 55 configuration–seed pairs, without clustering a shared seed across configurations. This is an exploratory subset comparison.

Model	Model top-1	Shadow top-1	Difference [95% CI]
Llama-3.1-70B	0.091	0.641	-0.550 [-0.671, -0.428]
GPT-5.6	0.218	0.593	-0.375 [-0.502, -0.247]
Gemini 3.1 Pro	0.327	0.842	-0.515 [-0.642, -0.388]
DeepSeek V4	0.218	0.844	-0.626 [-0.747, -0.503]

All four differences are negative with intervals excluding zero. Thus the shortfall is also present when selection uses an opening accusation, before feedback-driven accusations arise. The subset still selects a correct signal generated by the scripted opening policy; it does not eliminate all selection effects or establish independence from the environment. A separate analysis in the source uses an opening accusation by any honest seat (163 games per model), whether or not it is a neighbor. That population does not require delivery to the evaluated seat and is distinct from the 55-game population above. Source results/verify_2026-09-04/stratum_seed_event_2026-09-10.py.

B.30 Where the adversary sits inside the non-neighbor block

Among out-of-view games in which a neighbor accused the adversary before the final judgment, 300 of 411 have nonuniform confidence scores among non-neighbors (89, 68, 58 and 85 by model). Games with a single shared score are excluded; partial ties are retained. Confidence-based rank uses mean ranks for ties and is normalized so that 1 is highest and 0.5 is the expected rank under a uniform ordering. On these same games, mean normalized rank is 0.889, 0.973, 0.945 and 0.939 for the model, versus 0.907, 0.942, 0.954 and 0.935 for the shadow rule. The pooled means are 0.933 and 0.932. With ties counted by expectation for both, the adversary is first within the block in 82.7% and 83.2%, respectively. Similar means do not establish equivalence or show how the model used the evidence.

The adversary's median absolute rank is fourth, third, first and third. Conditional on it not being first overall in this selected sample, every seat above it is a direct neighbor in 77%, 92%, 69% and 76% of games. Flat-score exclusions are 17%, 30%, 43% and 20% of the eligible games; 28 retained games list non-neighbors in seat order. These results describe a selected population. B.25's placement decomposition uses its own complete-output population and addresses a different question. Source: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py.

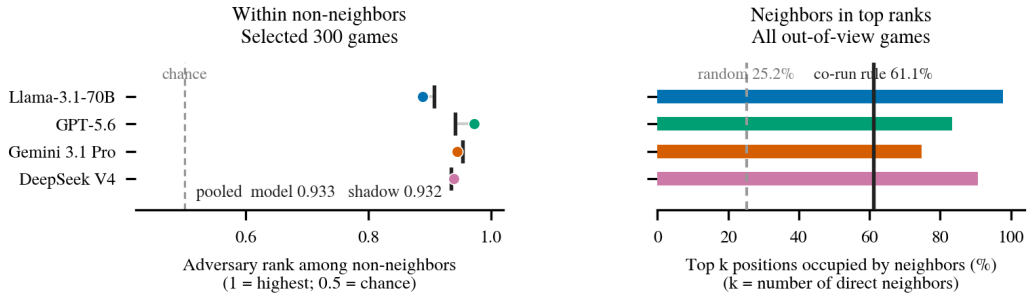


Figure B.1: Two analyses in §3.2 with different populations. **Left:** Out-of-view games with a relayed accusation and nonuniform non-neighbor confidence scores (300 of 411 games pooled across models). Confidence-based rank among non-neighbors uses mean ranks for ties and is normalized so 1 is highest and 0.5 is chance. Colored dots denote the model and black marks the shadow rule on the same selected games; similar means do not establish equivalence or internal reading. **Right:** All out-of-view games, including tied and parser-filled rankings: the share of the top k ranks filled with direct neighbors, where k is the number of direct neighbors. The solid reference is the co-run tally; the dashed reference is the expected share under a uniformly random ranking, averaged across games. These populations differ from the complete-output sample in Figure 2. Source build_fig_rank_patterns_2026-09-19.py.

B.31 Games that end on an accuser: rates and odds ratios

An accuser is a direct neighbor that accused the evaluated seat in rounds $1, \ldots, T-1$, before the final judgment at T. An event game ends with one of those accusers as top suspect. The reference group contains all other games, including those with no received accusation. Odds ratios use the Haldane correction. The first three models were held out from hypothesis formation; Llama is the discovery model.

Model	event n	wrong	reference n	wrong	odds ratio
GPT-5.6	380	0.813	110	0.300	10.01
Gemini 3.1 Pro	369	0.783	170	0.518	3.35
DeepSeek V4	362	0.751	128	0.414	4.25
Llama-3.1-70B	435	0.844	104	0.404	7.89

The held-out totals are 490, 539 and 490 games: the registered eligibility threshold, configuration-level final top-1 above 10%, excludes ring $N = 15$ for GPT-5.6 and DeepSeek. Llama uses all 539 games. The reference group splits into no-accusation and accused-but-not-selected games, with respective error rates 0.342 and 0.206, 0.551 and 0.472, 0.634 and 0.310, in model order. On the separate paired subset with a unique shadow top suspect ($n = 264, 327, 265$), the rule selects an accuser in 0.708, 0.398 and 0.434 of games. B.8 compares the model and rule on exactly that subset; ending on an accuser alone is not model-specific evidence. Source: results/verify_2026-09-04/accuser_summary_received_2026-09-20.py.

B.32 The silent-adversary arm, by model and configuration

The adversary still speaks but its statements reach no one, so the graph, the seat's position, the roster and the distances are unchanged and only audibility is removed. The quantity is the audibility effect of B.14 in the baseline minus the same effect in this arm, against the shadow tally, paired on seed within each cell: 12 in-view and 37 out-of-view games at ring $N = 10$, 11 and 38 at small-world $N = 15$.

model	configuration	reduction	95% CI
GPT-5.6	ring $N = 10$	$+0.167$	$[-0.036, +0.383]$
GPT-5.6	small-world $N = 15$	$+0.195$	$[+0.049, +0.372]$
Llama-3.1-70B	ring $N = 10$	$+0.243$	$[+0.096, +0.392]$
Llama-3.1-70B	small-world $N = 15$	$+0.214$	$[+0.035, +0.412]$

All four move in the direction of reduction and three of the four intervals exclude zero. Two models and two configurations is narrow, and the arm removes audibility without removing the adversary's presence in the tally the scripted seats keep, so this measures the change under a specific message-removal intervention rather than isolating an audibility mechanism. Sources results/original_analyses_2026-09-04/silent_arm_2026-09-04.py and results/verify_2026-09-04/silent_arm_two_models_2026-09-05.py.

B.33 Accusation of the seat: both judges lose, so the axis cannot carry a mechanism

Whether being accused is what pushes the relayed name down cannot be read off this axis. Distance-one games (the adversary among the seat's neighbors, the seat's degree at least two) split by whether any other neighbor accused the seat inside the scoring window. Twenty-seven cells over four models and 433 games; the cell in which the pattern was first seen is held out. The primary subset requires the adversary and all other neighbors to appear in the model's raw ranking text. Top-1 accuracy uses the stored scored ranking; the raw order defines the separate neighbor-displacement measure. Resampling is clustered on seed, as the four models share each seed's graph. The model ranks the adversary first in 0.776 of the games with no such accusation and 0.444 of those with one, a difference of $+0.332$ [$+0.214$, $+0.429$], positive in 22 of 27 cells and unchanged under the fill convention of B.9. The tally rule reading the same received stream falls at least as far, 0.761 to 0.381, a difference of $+0.380$ [$+0.252$, $+0.489$], and the difference of differences is -0.048 [-0.190, $+0.098$]. Both judges have lower accuracy in accused games, and the interval on their contrast includes zero. This does not establish equality or a causal effect of accusation, but it does not isolate an additional model-specific loss. At the hub, where the seat hears everyone, the model's top-1 difference between games without and with an accusation is $+0.100$ [-0.015, $+0.215$]. Reported as a null. Source results/verify_2026-09-04/accuser_displacement_allcells_2026-09-10.py.

B.34 Cells whose sign depends on the comparator mode

Per cell, the model's paired gap against the tally rule in each mode: co-run, the rule playing the seat in its own game under the same seed, and shadow, the rule replayed on exactly the statements the seat received. Under the registered treatment of ties, five of the 44 cells differ in sign between the two, all of them in reciprocal rank: isolated pairs $N = 10$ (Llama-3.1-70B), small-world $N = 10$ and cycle-heavy $N = 10$ (both GPT-5.6), hub $N = 10$ (Gemini 3.1 Pro) and hub $N = 15$ (Llama-3.1-70B). In four of the five it is the co-run gap that is negative. Using expected credit for shadow top-1 ties instead, while retaining identifier-based reciprocal rank, makes it ten cell-and-metric combinations over nine cells, five of them isolated pairs, where the shadow rule reads a stream carrying only the seat's own words echoed back and ties are pervasive. The number of such cells is therefore a statement about the tie treatment as much as about the modes. Ring at $N = 15$ is not among them: it is negative in both modes and both metrics in all four models; the reciprocal-rank gaps range from -0.0295 to -0.1686. Source results/verify_2026-09-04/ring_n15_mode_sign_2026-09-10.py.

B.35 The complete graph

The breadth limit: every seat hears every other, so no game has an out-of-view half and the split is not defined. What can be read is the level. The co-run tally rule reaches 0.6700, 0.6048 and 0.6488 MRR at $N = 10$, 15 and 20; all four models finish above it in all twelve model–size cells, by 0.048 to 0.215, mean 0.134. Forty-nine seeds per cell, deterministic re-run. This limits the breadth interpretation summarized in §4.2: it is not an isolated test of breadth: every scripted seat also hears everything here, so the comparator changes with the seat. Source: results/verify_2026-09-04/complete_graph_2026-09-04.py.

B.36 The honest-defense arm, by view

Letting honest seats defend after the opening round with probability one half destroys the signature that every later defense comes from the adversary. We use one model (GPT-5.6), seed-paired to the baseline, with the shadow comparator. The manipulation check is in Appendix E: round-one truthfulness and the adversary's own defense rate are unmoved, and honest post-opening defenses go from 0% to about 50% in all three honest variants.

configuration	view	n	model, arm − baseline	rule, arm − baseline
ring $N = 10$	in	12	+0.198 [+0.073, +0.333]	−0.008 [−0.237, +0.222]
	out	37	+0.070 [+0.020, +0.130]	−0.059 [−0.171, +0.041]
small-world $N = 15$	in	11	+0.151 [−0.038, +0.333]	−0.042 [−0.245, +0.165]
	out	38	+0.053 [+0.013, +0.104]	+0.028 [−0.043, +0.104]
hub $N = 15$	in	49	+0.050 [−0.033, +0.136]	−0.019 [−0.093, +0.058]

The hub row has no out-of-view half, so it carries no split. In the two configurations that do, the model has a larger point-estimate gain in view than out of view; each rule-change interval includes zero. This is the direction §4.2 reports; pooled over the in-view cells the model's change is +0.090 [+0.021, +0.160] against the rule's −0.020 [−0.090, +0.054]. Levels are not headlined here: the arm runs on one model and the registered script says so. Source: results/verify_2026-09-04/honest_defense_arm_2026-09-07.py.

B.37 Defense cues and final choices

A neighbor's post-opening defense is delivered before the final judgment in 205 of 539 games per model. The games are the same across models because the adversary is seed-deterministic. Arrival rates are 91.8% and 89.8% in hub N10/N15 and 8.2% to 59.2% across the other nine configurations (isolated pairs 12.2 and 8.2, ring 22.4 and 12.2, small world 40.8 and 18.4, cycle-heavy 59.2, 36.7 and 26.5). Among these games, the model ranks the defender first in 52.7%, 58.5%, 54.6% and 65.9% (Llama-3.1-70B, GPT-5.6, Gemini 3.1 Pro, DeepSeek V4). The shadow tally on the same games does so in 55.1% to 56.6% with lowest-ID ties and 48.1% to 49.5% with expected credit for top ties. A uniform random ranking of all other seats gives 9.1%. All 363 of 363 delivered post-opening defenses per model come from the adversary. These rates describe the co-occurrence of a cue and a final choice; they do not establish that either occupant recognized or used the signature. Source: results/verify_2026-09-04/timing_corrected_2026-09-20.py.

B.38 Two objections about how the rule is fed

Two objections follow from how the rule is fed (§3.1, §4.2). It might be tuned to this environment rather than the naive comparator we call it. That interpretation is partly supported by the level results, and we do not rule it out. Using the same rule in both halves does not make this dependence cancel: the streams differ with audibility, and a rule need not contain an explicit audibility variable for its relative fit to differ across them. The contrast interval excludes zero for five of six tested rules on the same statements; these rules share parts of the tally construction. The defense-speaker rule is the exception, so this check does not establish comparator independence (B.26). The second objection is ours: ties resolve toward seat 1 and the prompt calls investigators truthful, so the environment manufactures accusations of the seat and the prompt tells it those are lies (§2); randomizing that convention on both arms leaves the deficit in four of four models (Appendix G).

B.39 What the seat's own text reproduces of what it heard

§3.2 reports similar mean within-block ranks for the model and the shadow rule on a selected population. Such means do not establish statistical equivalence or faithful reading: surface mention counting can also produce an informative ordering. We therefore ask whether the model's text co-mentions actual speaker–target pairs more often than pairs that did not occur.

On the out-of-view games where a neighbor accused the adversary before the final judgment, we read the evaluated seat's stored reasoning for that judgment. For each distinct speaker–target accusation pair received before judgment, we ask whether some sentence of the text names both players. Repeated accusations of the same target by the same speaker count once per game. The control uses a neighbor who spoke before judgment and a target that neighbor had not accused by then. Absolute co-mention rises with how many names a text uses, so the quantity to read is the difference.

For Llama-3.1-70B, GPT-5.6, Gemini 3.1 Pro and DeepSeek V4, respectively, real pairs are reproduced at 0.645, 0.774, 0.159 and 0.227 against 0.156, 0.207, 0.071 and 0.052 for pairs that had not occurred before judgment. Real pairs are reproduced more often in all four models, with differences ranging from 0.088 to 0.567. Reasoning text was present in all 411 games.

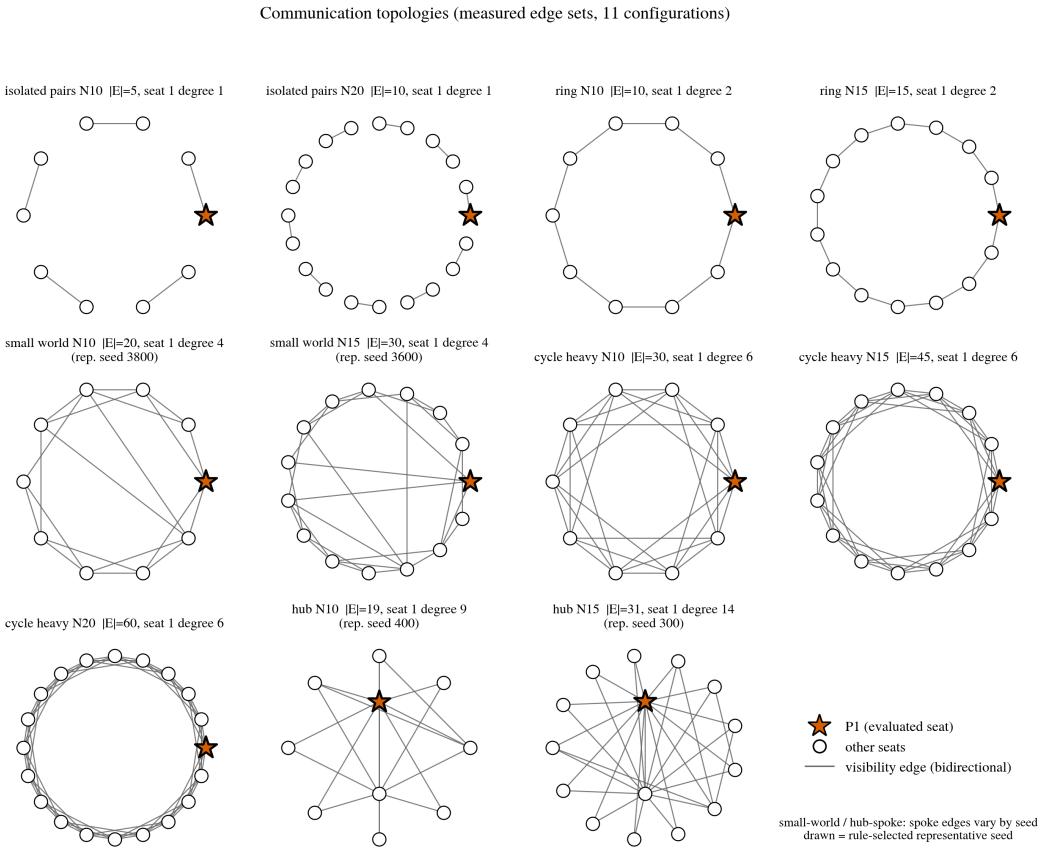


Figure C.1: The eleven communication configurations, drawn from actual edge lists (representative seeds; the seeds drawn are given in Appendix A), ordered by how much of the network the evaluated seat hears directly: isolated pairs (1 audible neighbor), ring (2), small-world (2–7, seed-dependent), cycle-heavy (6), and hub-and-spoke with the evaluated seat at the hub (9–14), with network sizes sweeping N = 10–20 within families.

The text co-mentions actual speaker–target pairs more often than absent pairs. This is evidence of relational information in the output beyond a total count per target. Co-mention neither asserts the relation nor demonstrates that it was interpreted correctly or used in the final decision. Reasoning text is not a faithful record of computation, so this statistic is not a lower bound on internal reading. The analysis is exploratory, on the same games that motivated it; its criterion was written before this measurement, not before the underlying data existed. Source results/verify_2026-09-04/text_recall_received_2026-09-20.py.

C TABLES AND FIGURES MOVED FROM THE MAIN TEXT

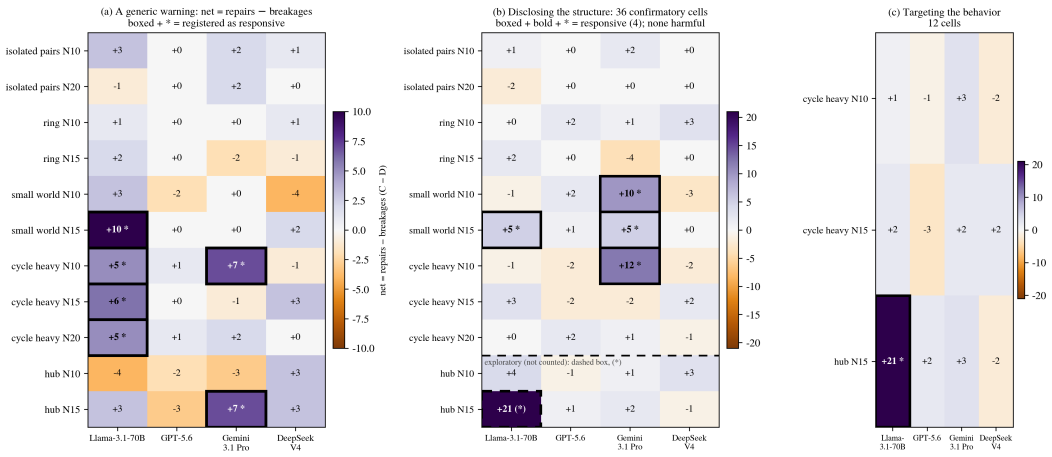


Figure C.2: Intervention landscapes under the preregistered same-seed criterion (net paired improvement ≥ 5 of 49 seeds). (a) The generic warning: responsive in 4/0/2/0 of eleven configurations across the four models; four maps, no shared pattern, one net-negative hub cell. (b) Structure disclosure: responsive in 4 of 36 confirmatory cells (two models), none harmful; hub cells (dashed) were excluded by registration and carry no verdict, * marking the one that clears the criterion (Llama, N = 15). (c) Behavioral targeting: responsive in 1 of 12 cells, none harmful.

configuration	tally rule	Llama-3.1-70B	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
isolated pairs N10	0.3917	0.3463	0.3197	0.3178	0.3310
isolated pairs N20	0.2652	0.2233	0.2164	0.2273	0.2287
ring N10	0.5646	0.3369	0.3344	0.3404	0.3944
ring N15	0.3883	0.2197	0.2426	0.2807	0.2552
small world N10	0.5824	0.4532	0.5614	0.5027	0.5300
small world N15	0.4734	0.1932	0.3119	0.3350	0.3434
cycle heavy N10	0.6601	0.6143	0.6340	0.5719	0.6273
cycle heavy N15	0.6024	0.3508	0.4368	0.4609	0.3446
cycle heavy N20	0.5661	0.3863	0.3696	0.4010	0.3886
hub N10	0.6944	**0.8082**	**0.7353**	**0.7019**	**0.7649**
hub N15	0.5687	0.5558	**0.6624**	**0.6743**	**0.7988**
k (rule above model)		10	9	9	9

Table C.1: Per-configuration final-round MRR for all four models against the co-run tally rule. Bold marks the cells where the model is above the rule: all four at hub $N = 10$ and three at hub $N = 15$, seven of the 44, and no other configuration. The last row counts, per model, the configurations in which the rule is above it. Excluding the two isolated-pairs configurations, which relay nothing, the cell count is 29 of 36 rather than 37 of 44. Ties are not counted as the rule being above. Source results/verify_2026-09-04/cells_37of44_2026-09-04.py; the underlying per-configuration means are figures/data/T1_k_map.csv.

row	probe (name in text)	verdict	status
N1	round weighting	no verdict: too few events in the terminal round (registered clause); descriptive report only	not surviving
N2	bandwagon following	no directional verdict: starved unanimity cell	not surviving
N3	confirmation asymmetry	no verdict: rebuttal supply below the registered minimum	not surviving
N4	defense discount	descriptive report with per-direction flags	not surviving (descriptive row)
N5	**audible-neighbor concentration**	**survives**	**surviving**

Table C.2: The five-row behavioral search, specified before named statistics were unsealed on existing games: screening on GPT-5.6 with a sign gate on the discovery model, confirmation sealed to a held-out model, promotion order fixed before screening. One row survives: promotion of audible neighbors (paired +0.330, bootstrap CI [+0.249, +0.414], stratified 7/7 positive); the remaining rows print without surviving verdicts, three of them empty cells of the information ecology itself.

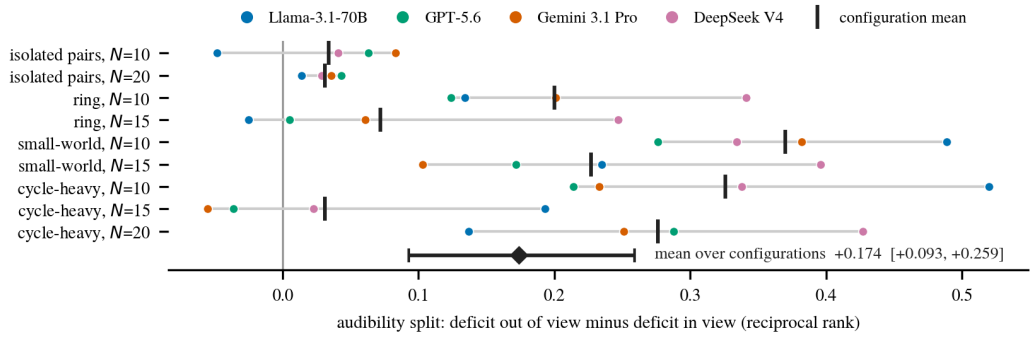


Figure C.3: The audibility split, in reciprocal rank: the model's deficit against the shadow tally out of view minus the same deficit in view, within a topology–size configuration; graph draws and role assignments can vary across seeds. Hub configurations are excluded because no game there falls outside the seat's view. All nine configuration means are positive and 32 of the 36 cells are; the four that are not are named in Appendix B, B.14, with the per-cell counts and the same values in a table.

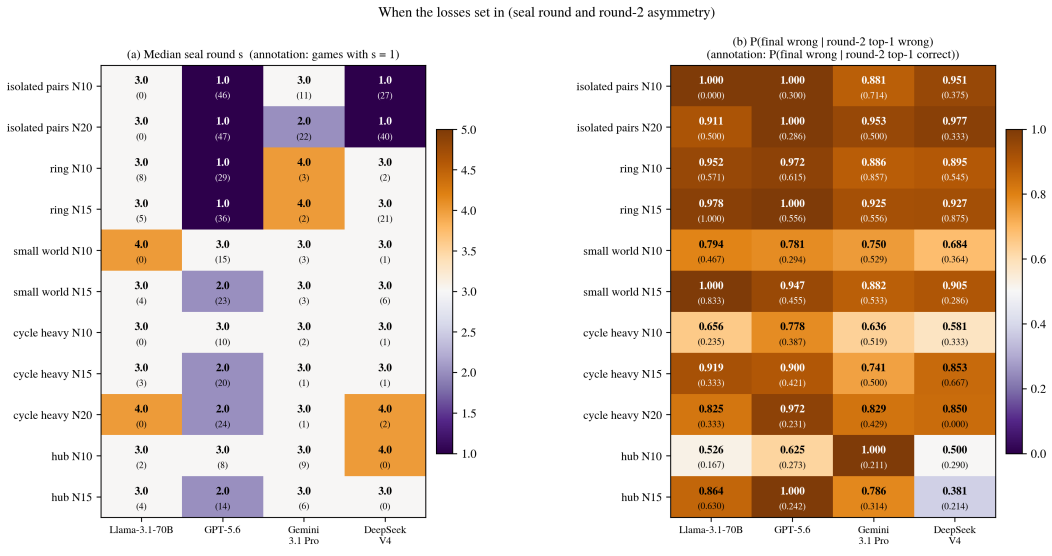


Figure C.4: When the losses set in. Two panels over 11 configurations × 4 models: (a) median seal round, the earliest round from which a game's top-1 never changes (cell annotation: games sealed at round 1); medians are 3 of 5 rounds for three models and 1 for GPT-5.6. (b) P(final wrong — round-2 leader wrong), annotated with P(final wrong — round-2 leader correct): pooled per model, 81–93% against 35–45%.

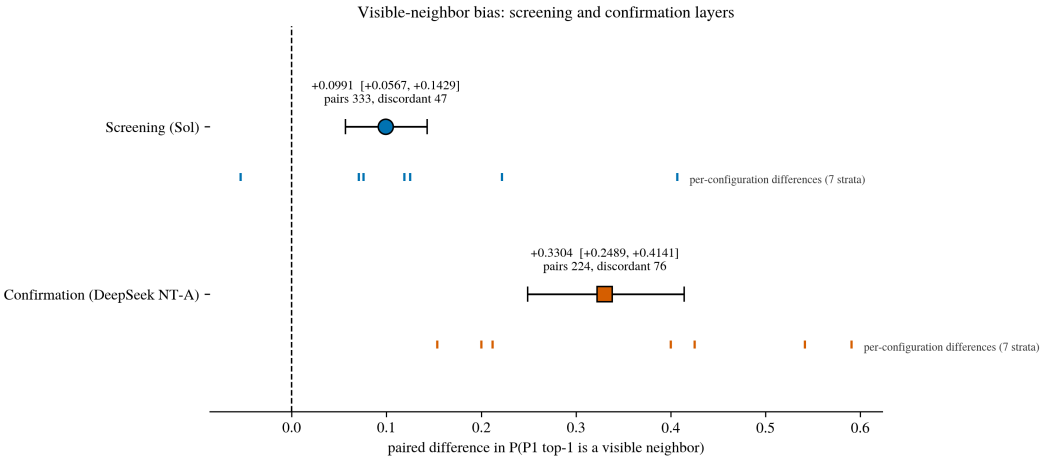


Figure C.5: Confirmation of the surviving row: paired audible-neighbor promotion effect +0.330 (bootstrap CI [+0.249, +0.414]), positive in 7 of 7 strata, against the screening estimate +0.099 (CI [+0.057, +0.143]) that promoted it.

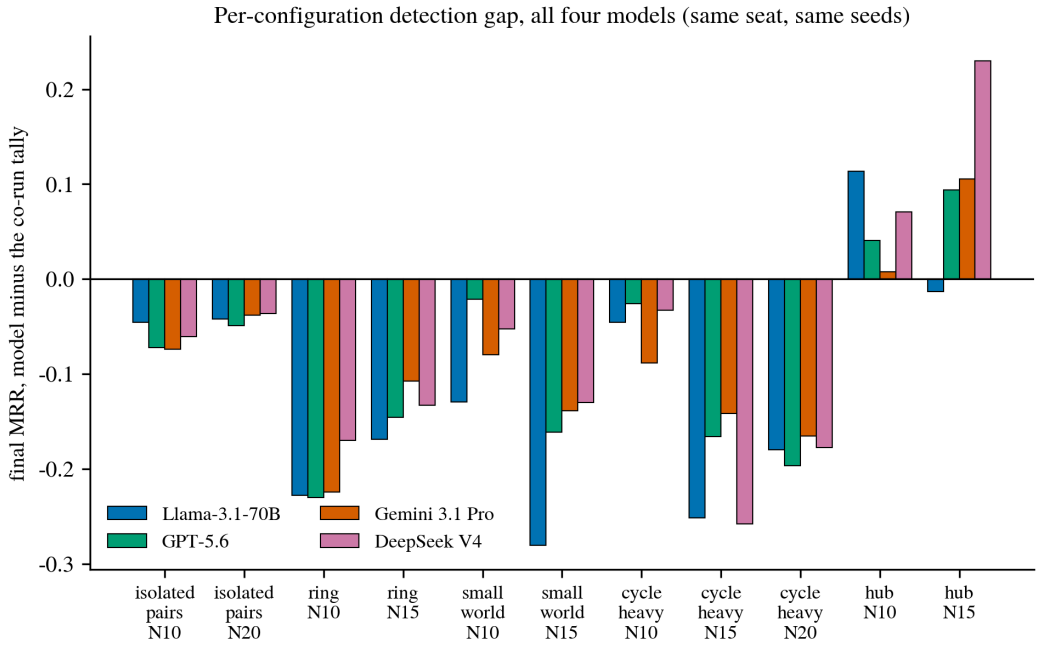


Figure C.6: Per-configuration detection gap for all four models: final-round MRR of each model minus the co-run tally in the same seat under the same seeds (values in Table C.1). Every model falls below the rule in nine or ten configurations; every cell above zero that a re-run settles is one of the two hub seats. Annotated cell in the text, ring N = 10: the baseline reaches 0.565 against a chance MRR of 0.314 while the models reach 0.334 to 0.394; this is the active replacement comparison. §3.1 separately compares rankings on the model's observed stream.

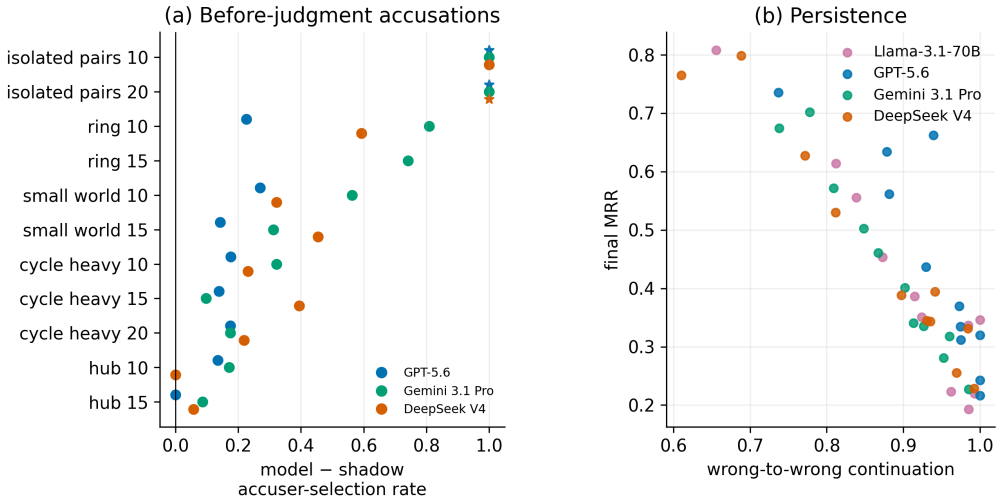


Figure C.7: (a) Per-configuration paired excess in selecting an accuser over the shadow tally, for the three held-out models. Accusers are neighbors that accused the evaluated seat before final judgment. Only eligible configurations and games with a unique shadow top suspect enter ($n = 264, 327, 265$ for GPT-5.6, Gemini and DeepSeek). The models contribute 10, 11 and 10 configurations, respectively; ring N15 contributes only a Gemini point. Stars mark cells with fewer than five paired games. Llama appears only in panel (b); colors have the same model meanings across panels. (b) Per-configuration wrong-to-wrong continuation against final MRR, 11 configurations per model. Panel (b) retains the original scatter data.

D PROTOCOL-LAYER FEASIBILITY PILOT

Appendix D: Protocol-layer feasibility pilot. D.2 below reports the protocol family's main run; the registered pilot that preceded it is reported here, a pilot of its most direct instance (a query channel by which the evaluated seat may ask any other player for its current top suspect, with the reply delivered privately in the next round), run to de-risk the follow-up protocol, whose outcome analysis is registered in Appendix A. Design: two arms (channel present / absent) × two configurations (isolated pairs and hub-and-spoke, N = 10) × 15 seeds = 60 games, Llama-3.1-70B, all settings frozen and hashed before the first game log. Feasibility: across 30 smoke-test and 150 main-run probe-arm opportunities, the parser logged no out-of-range targets in recognized query commands. Commands not matched by its query expression do not increment this counter, so it does not establish complete format compliance. The model selected the query action in 8 of the 150 probe-arm opportunities (5 in isolated pairs, 3 in hub-and-spoke), naming its target each time; one reply named the evaluated seat itself. Validity: five independent contamination checks (prompt reconstruction, game composition, received streams, static code installation, runtime branch counters) found no probe-arm code or text in the control arm (30/30 games, counter 0; positive self-test 14). Model serving is nondeterministic at temperature zero (§6): same-instance reruns reproduced the final metric in 1 of 5 games, and the baseline runs predate this serving instance, so the pilot supports within-run comparison only. Table D.1 reports the complete per-seed final MRR for all four cells, as logged. We draw no directional conclusion from it: the registered same-seed criterion is specified for forty-nine seeds (§2), and at fifteen seeds per-seed differences cannot be separated from the run-to-run variation documented above. An initial regression gate that assumed bit-level reproducibility against the older instance failed and was replaced, before unsealing, by the direct contamination checks above. The instability is not a matter of settings drifting between campaigns. The discovery campaign and the counterfactual re-runs were served from the same vLLM build (0.18.1) under byte-identical non-default arguments: eager execution, four-way tensor parallelism, prefix caching and chunked prefill both disabled, one sequence per iteration, and seed 0. Even with these settings held fixed, re-execution changes per-game outcomes. Reproducing our serving configuration is therefore nec-

essary but not sufficient: the figures we report are cell means precisely because individual games do not reproduce.

Table D.1: Per-seed final MRR for the four pilot cells (arm × configuration; 15 seeds each), as logged.

seed	probe_off isolated	probe_on isolated	probe_off hub-and-spoke	probe_on hub-and-spoke
100	0.1667	0.2000	1.0000	1.0000
200	1.0000	1.0000	0.3333	1.0000
300	0.2000	0.1429	1.0000	1.0000
400	0.2000	0.2000	1.0000	1.0000
500	0.3333	1.0000	1.0000	1.0000
600	0.2000	0.3333	0.2500	0.2000
700	0.3333	0.2500	0.1111	0.5000
800	0.1111	0.1111	1.0000	0.5000
900	0.2000	0.2000	0.3333	0.5000
1000	0.1429	0.1429	1.0000	1.0000
1100	1.0000	1.0000	0.5000	0.5000
1200	0.1667	0.1667	0.1111	1.0000
1300	0.1667	0.1250	1.0000	1.0000
1400	0.2500	0.2500	0.5000	0.3333
1500	0.2000	0.2000	1.0000	1.0000
mean	0.3114	0.3548	0.6759	0.7689

D.2 Protocol-layer main run (registered). The main run followed the pilot's method: its preregistration (sha256 a0e88df8…96db) fixed question, cells, seeds, and verdict rule, was frozen and committed before the first game log, and was not edited afterwards. Design: two arms (channel present / absent) × two configurations (isolated pairs and hub-and-spoke, N = 10) × 34 seeds per configuration, drawn deterministically from the registered seed list after excluding the pilot's 15, giving 136 games on the same serving instance and settings as the pilot. Registered verdict rule, per cell: count the same-seed pairs in which the channel-present arm scores higher final MRR (Cm) and lower (Dm), ties uncounted; the cell is labeled "responsive" if —Cm − Dm— ≥ 4 of 34 and otherwise "no detected response under our design". The registration records that this threshold is an analogy to the §2 same-seed criterion (≥ 5 of 49) on a different metric, rounded in the conservative direction, and that the two are not directly comparable in stringency.

Primary count (registered):

Table D.2: Per-seed final MRR for the four main-run cells (arm × configuration; 34 seeds each) and the paired difference (channel present minus absent), as logged.

cell	n	Cm (on > off)	Dm (on < off)	tie	abs(Cm − Dm)	label
isolated pairs N = 10	34	5	7	22	2	no detected response under our design
hub-and-spoke N = 10	34	4	7	23	3	no detected response under our design

Registered descriptive quantities (no bearing on the label): mean paired ΔMRR -0.0248 (isolated pairs) and -0.0463 (hub-and-spoke); pooled over 68 pairs, Cm 9, Dm 14, tie 45; top-1 net count, C (on hit, off miss) minus D (off hit, on miss), -1 and -2 (this is the registration's "Cm/Dm-style" count applied to top-1: ties uncounted); query-action selection rate 0.0824 (isolated pairs, 14 of 170 opportunities) and 0.0294 (hub-and-spoke, 5 of 170). For reference only, and excluded from the verdict by registration, the pilot's 15 seeds give Cm 3, Dm 3, tie 9 (isolated pairs) and Cm 4, Dm 3,

tie 8 (hub-and-spoke). Table D.2 reports the complete per-seed final MRR and paired difference for all four cells, as logged.

seed	probe_off isolated	probe_on isolated	Δ	probe_off hub-and-spoke	probe_on hub-and-spoke	Δ
1600	1.0000	1.0000	+0.0000	1.0000	0.5000	-0.5000
1700	0.1429	0.1429	+0.0000	1.0000	1.0000	+0.0000
1800	0.1667	0.1667	+0.0000	0.1667	0.1111	-0.0556
1900	1.0000	1.0000	+0.0000	1.0000	0.3333	-0.6667
2000	0.1250	0.1250	+0.0000	0.5000	1.0000	+0.5000
2100	0.2000	0.1667	-0.0333	0.2000	0.1111	-0.0889
2200	0.1667	0.1667	+0.0000	1.0000	1.0000	+0.0000
2300	0.1111	0.1111	+0.0000	1.0000	1.0000	+0.0000
2400	0.3333	0.3333	+0.0000	0.5000	0.5000	+0.0000
2500	1.0000	1.0000	+0.0000	1.0000	1.0000	+0.0000
2600	0.1111	0.3333	+0.2222	0.1250	0.3333	+0.2083
2700	0.2000	0.2000	+0.0000	1.0000	1.0000	+0.0000
2800	0.5000	0.5000	+0.0000	1.0000	1.0000	+0.0000
2900	0.1667	0.2000	+0.0333	0.5000	0.2500	-0.2500
3000	0.2500	0.1250	-0.1250	0.5000	0.5000	+0.0000
3100	0.5000	0.5000	+0.0000	1.0000	1.0000	+0.0000
3200	0.3333	0.5000	+0.1667	1.0000	1.0000	+0.0000
3300	0.5000	0.5000	+0.0000	1.0000	1.0000	+0.0000
3400	0.1250	0.1250	+0.0000	1.0000	1.0000	+0.0000
3500	0.1429	0.1429	+0.0000	1.0000	1.0000	+0.0000
3600	1.0000	0.2000	-0.8000	1.0000	1.0000	+0.0000
3700	1.0000	1.0000	+0.0000	0.5000	1.0000	+0.5000
3800	0.1667	0.2000	+0.0333	0.1111	0.1111	+0.0000
3900	0.5000	0.3333	-0.1667	1.0000	0.5000	-0.5000
4000	0.1429	0.1429	+0.0000	0.5000	0.5000	+0.0000
4100	0.1667	0.1429	-0.0238	0.3333	0.3333	+0.0000
4200	0.1111	0.1111	+0.0000	1.0000	1.0000	+0.0000
4300	0.1111	0.1111	+0.0000	1.0000	1.0000	+0.0000
4400	1.0000	1.0000	+0.0000	1.0000	1.0000	+0.0000
4500	0.1111	0.1111	+0.0000	1.0000	0.1111	-0.8889
4600	0.2000	0.2500	+0.0500	1.0000	1.0000	+0.0000
4700	0.5000	0.3333	-0.1667	1.0000	1.0000	+0.0000
4800	0.1250	0.1250	+0.0000	1.0000	1.0000	+0.0000
4900	0.2000	0.1667	-0.0333	0.3333	0.5000	+0.1667
mean	0.3650	0.3402	−0.0248	0.7726	0.7263	−0.0463

E PREREGISTERED ABLATION ARMS

This appendix reports the ablation campaign in full: six arms on GPT-5.6 with the tally rule run under the same manipulation and two cells on Llama-3.1-70B, all preregistered before any data contact, all cells 49 seeds, seed-paired to the main campaign. Later arms are reported here too and are not part of that registration: the seat swap across all four models, the roster-only runs on DeepSeek V4 and Llama-3.1-70B, the roster-plus-premise run on DeepSeek V4, the LLM-neighbor cell on GPT-5.6, and the corrected-premise relay extensions for DeepSeek V4, Llama-3.1-70B and Gemini 3.1 Pro analyzed in E.6. The silent arm's two Llama cells are in B.32.

Preregistration lineage. Five revisions, each frozen before the data it governs existed: v1 sha 181530ae (2026-08-31 02:04 KST; opening-truth dilution design, verdict clause, and the rejection of a p = 0.5-only design) → v2 9c1e69ce (02:39; silent-adversary arm, naive-collapse reporting clause, defense-cue clause) → v3 2d22a754 (02:46; honest-defense arm with its manipulation check, honest-replacement arm) → v4 692764c5 (02:49; corrected-premise arm with branch verdicts fixed in advance, cross-arm metrics) → v5 9ccfe405 (03:02; cross-model differential prediction, minimal cell set, batch execution path). The naive arms already produced at the time of v5 were sealed unread. The judgment script (sha 04a3a72a) ran once, after every arm completed; its bootstrap intervals use 10,000 resamples, seed 20260831. E.1 uses separately retained table generators, specified below.

Arm definitions. roster-only (R): the line naming the seat's direct neighbors by number is replaced by two lines, one listing every other player and stating that exactly one of them is the criminal, the other stating that the seat hears statements from only some of them each round and learns of the rest through what others say; the original premise is retained. The modification therefore supplies the full candidate list and withdraws the explicit neighbor identities together, and is not a one-sided addition; the round messages name their speakers, so those identities are observable again from the second round. roster-plus-premise (C): the same interface modification together with the corrected premise, hence a three-part change; corrected-premise: the two premise lines of the system prompt replaced as reproduced in Appendix A, nothing else changed; the two lines are replaced together, so the arm is a two-line bundle rather than a premise-only contrast, in the same sense in which thinking mode is a protocol bundle (Limitations); no arm corrects one line alone; p = 0.75 and p = 0.5: honest opening statements true with probability p; honest-defense q = 0.5: honest seats defend their lowest-ranked seat with probability q per round after the opening, erasing the defense-speaker signature of the base environment; honest-replacement (directory name noadv): the agent in the criminal seat replaced by an honest naive rule, with the scored seat, roles, seeds, and graphs unchanged; silent: the adversary's statements are generated and logged but delivered to no one.

E.1 Final MRR by arm and configuration

GPT-5.6 against the tally rule: the co-run tally arm where the manipulation applies to the rules (p, honest-defense, honest-replacement, silent), the registered p = 1.0 per-game baseline where it does not touch them (baseline row, corrected-premise). Paired gap over the 49 shared seeds, bootstrap 95% CI. The 25 GPT-5.6 rows use 10,000 paired-seed resamples from a shared NumPy generator seeded at 20260831, with interpolated percentile endpoints (47b2daa7_L26225_cce58abec2.py). The four Llama rows use the later reproducible calculation in premise_arm_coverage_2026-09-07.py: 10,000 paired-seed resamples, a cell-specific CRC32 seed, and sorted bootstrap endpoints at zero-based indices 250 and 9750. These conventions reproduce the retained values; they are not a single shared draw sequence.

Arm	Configuration	GPT-5.6	naive	gap	95% CI
baseline (random adversary)	isolated pairs N10	0.3197	0.3917	-0.0720	[-0.1443, -0.0179]
baseline (random adversary)	hub N10	0.7353	0.6944	+0.0410	[-0.0558, +0.1391]
baseline (random adversary)	ring N10	0.3344	0.5646	-0.2302	[-0.3302, -0.1341]
baseline (random adversary)	small world N15	0.3119	0.4734	-0.1615	[-0.2601, -0.0677]
baseline (random adversary)	hub N15	0.6624	0.5687	+0.0937	[-0.0010, +0.1930]
corrected-premise	isolated pairs N10	0.3771	0.3917	-0.0145	[-0.0643, +0.0260]
corrected-premise	hub N10	0.7122	0.6944	+0.0178	[-0.0827, +0.1205]
corrected-premise	ring N10	0.4019	0.5646	-0.1626	[-0.2445, -0.0820]
corrected-premise	small world N15	0.3290	0.4734	-0.1444	[-0.2409, -0.0529]
corrected-premise	hub N15	0.6640	0.5687	+0.0954	[-0.0067, +0.2012]
p = 0.75	ring N10	0.3574	0.4151	-0.0577	[-0.1503, +0.0367]
p = 0.75	small world N15	0.2495	0.3000	-0.0504	[-0.1078, +0.0018]
p = 0.75	hub N15	0.5140	0.3170	+0.1969	[+0.0988, +0.2988]

Arm	Configuration	GPT-5.6	naive	gap	95% CI
honest-defense q = 0.5	ring N10	0.4354	0.4285	+0.0069	[-0.0769, +0.0983]
honest-defense q = 0.5	small world N15	0.3868	0.4256	-0.0388	[-0.1262, +0.0503]
honest-defense q = 0.5	hub N15	0.7121	0.5248	+0.1873	[+0.0745, +0.2970]
honest-replacement	ring N10	0.4021	0.5918	-0.1897	[-0.2790, -0.1039]
honest-replacement	small world N15	0.2988	0.5139	-0.2151	[-0.3150, -0.1231]
honest-replacement	hub N15	0.5369	0.6543	-0.1174	[-0.1861, -0.0571]
silent adversary	ring N10	0.2934	0.5714	-0.2779	[-0.3715, -0.1901]
silent adversary	small world N15	0.2668	0.5088	-0.2420	[-0.3359, -0.1538]
silent adversary	hub N15	0.5934	0.6543	-0.0609	[-0.1439, +0.0161]
p = 0.5	ring N10	0.3317	0.3461	-0.0144	[-0.1033, +0.0727]
p = 0.5	small world N15	0.2370	0.2356	+0.0014	[-0.0510, +0.0551]
p = 0.5	hub N15	0.4572	0.2686	+0.1886	[+0.0859, +0.2902]

Llama-3.1-70B under the corrected premise, the two cells the differential prediction required, against the same tally rule:

Arm	Configuration	Llama-3.1-70B	naive	gap	95% CI
baseline (random adversary)	isolated pairs N10	0.3463	0.3917	-0.0453	[-0.0983, +0.0041]
corrected-premise	isolated pairs N10	0.3396	0.3917	-0.0520	[-0.1169, +0.0022]
baseline (random adversary)	hub N10	0.8082	0.6944	+0.1138	[-0.0183, +0.2453]
corrected-premise	hub N10	0.7535	0.6944	+0.0591	[-0.0716, +0.1910]

Registered decision rules and observed outcomes: (1) At p = 0.75 the model-minus-naive gap keeps its baseline point-estimate sign in all three decision configurations, meeting the v1 threshold of two of three; the naive-above-chance precondition is also met. At ring N = 10 the gap narrows from -0.230 to -0.058 as the tally rule's score falls from 0.565 to 0.415 and the model's changes from 0.334 to 0.357. Both baseline-deficit configurations have intervention-arm gap intervals that include zero. (2) Under honest-defense q = 0.5 the manipulation check passed (the adversary's share of post-opening defenses fell from 1.00 to 0.07–0.11). Point-estimate signs are retained in small-world N15 and hub N15, meeting v3's two-of-three threshold. The hub was a model advantage at baseline; both baseline-deficit configurations have intervention-arm gap intervals that include zero. These sign-count criteria are satisfied, but do not establish that the opening or defense conventions made no contribution to the deficit. (3) The corrected-premise arm follows v4 branch (ii) in the three original decision configurations: the ring N10 and small-world N15 deficits narrow and remain negative with intervals excluding zero, while hub N15 retains a positive point estimate. This preserves the observed sign pattern without establishing unchanged hub performance. (4) The p = 0.5 arm is descriptive by registration; its scores are reported alongside the other arms.

Across the seven cells rerun under the corrected premise, five on GPT-5.6 and two on Llama-3.1-70B, the sign of the gap against the tally rule is unchanged in every one. These seven are the registered arm; the extension to 35 cells that §4 cites is reported in B.22 and is outside this registration.

E.2 CORRECTED-PREMISE ARM: FINISHING ON AN ACCUSER

Rate of finishing with a neighbor who accused the seat before its final judgment as top suspect, among games with such an accusation (registered P5 definition, neighbor-limited), baseline against corrected-premise, paired delta over games accused in both. The registered differential prediction (v5) is a larger decrease in premise-citing models than in Llama, whose reasoning rarely matches those patterns (B.11). The first table uses the original campaign arms and does not by itself evaluate the contrast at matched temperature.

Model	Configuration	baseline	corrected	paired delta	95% CI
GPT-5.6	isolated pairs N10	1.000 (n = 49)	0.755 (n = 49)	-0.2449	[-0.3673, -0.1224]
GPT-5.6	hub N10	0.846 (n = 39)	0.789 (n = 38)	-0.0526	[-0.1579, +0.0526]
GPT-5.6	ring N10	1.000 (n = 46)	1.000 (n = 45)	+0.0000	[+0.0000, +0.0000]
GPT-5.6	small world N15	0.929 (n = 42)	0.929 (n = 42)	+0.0000	[-0.0714, +0.0714]
GPT-5.6	hub N15	0.889 (n = 36)	0.865 (n = 37)	-0.0278	[-0.1389, +0.0556]
Llama-3.1-70B	isolated pairs N10	1.000 (n = 49)	0.957 (n = 46)	-0.0435	[-0.1087, +0.0000]
Llama-3.1-70B	hub N10	0.733 (n = 45)	0.775 (n = 40)	+0.0000	[-0.1250, +0.1250]

Populations and temperature. The first table uses the original campaign arms, with accusations restricted to those received before the final judgment. Intervals use 10,000 paired-seed bootstrap resamples with a shared generator seeded at 20260831. Source: results/verify_2026-09-04/accuser_received_2026-09-20.py. A zero-width interval reflects an all-zero empirical difference vector, not a population equivalence test. Its rates use each arm's own accused population, whereas its differences use seeds accused in both arms; these are different denominators. The Llama rows also compare a corrected arm at temperature 0.7 with a baseline at 0.0. They must not be read as matched-temperature premise effects. The retained reruns at 0.0 give a Llama paired change of -0.143 in isolated pairs and $+0.073$ at the hub. Movement in Llama alone does not contradict v5: the prediction concerns relative reductions, not an unchanged Llama.

Direct evaluation of the relative prediction. We use the retained temperature-matched arms in the two registered configurations. Within each model–Llama comparison, a seed enters only if the seat received an accusation before its final judgment in both arms of both models. Define Δ as corrected minus baseline accuser-first outcome; the contrast is $\Delta_{\text{held-out}} - \Delta_{\text{Llama}}$, so negative values follow the registered prediction. Graphs, roles and opening statements match at each paired seed. This reanalysis implements v5's same-cell, same-seed comparison with an explicit common accused population and 10,000 paired-seed bootstrap draws; that population rule and bootstrap implementation are post-hoc analysis choices, not parts of v5. Intervals are unadjusted for multiple comparisons.

Held-out model	Configuration	Paired seeds	Contrast [95% CI]
GPT-5.6	isolated pairs $N = 10$	49	-0.102 [-0.265, $+0.041$]
Gemini 3.1 Pro	isolated pairs $N = 10$	10	-0.600 [-0.900, -0.300]
DeepSeek V4	isolated pairs $N = 10$	44	-0.045 [-0.182, $+0.091$]
GPT-5.6	hub $N = 10$	36	-0.111 [-0.278, $+0.028$]
Gemini 3.1 Pro	hub $N = 10$	35	-0.200 [-0.343, -0.057]
DeepSeek V4	hub $N = 10$	37	$+0.027$ [-0.189, $+0.243$]

Five of the six point estimates follow the predicted direction. Only the two Gemini intervals exclude zero; both GPT-5.6 intervals and both DeepSeek intervals include zero, and the DeepSeek hub estimate has the opposite sign. Support for the relative prediction is limited and varies across models and configurations. Accusation is affected by the intervention, so conditioning on it limits this comparison; in particular, the Gemini isolated-pairs estimate uses only 10 seeds. Neither this contrast nor a reasoning-text pattern establishes mediation by premise use. Source: results/verify_2026-09-04/matched_received_2026-09-20.py.

E.3 THE COST OF ONE DISHONEST MEMBER

Registered interpretation fixed in advance (v2/v3): the cost is the same-seed MRR change from replacing the random adversary with an honest rule (honest-replacement, primary) or silencing it (silent, auxiliary). A positive delta means the agent does better without the dishonest member. Innocent top-1 is the arm's rate of ending on an innocent seat.

Arm	Agent	Configuration	delta vs baseline	95% CI	innocent top-1
honest-replacement	GPT-5.6	ring N10	+0.0677	[+0.0260, +0.1184]	0.816
honest-replacement	GPT-5.6	small world N15	-0.0131	[-0.0491, +0.0174]	0.857
honest-replacement	GPT-5.6	hub N15	-0.1255	[-0.1965, -0.0570]	0.612
honest-replacement	naive	ring N10	+0.0272	[+0.0000, +0.0612]	0.531
honest-replacement	naive	small world N15	+0.0405	[-0.0181, +0.1067]	0.612
honest-replacement	naive	hub N15	+0.0857	[+0.0194, +0.1637]	0.408
silent	GPT-5.6	ring N10	-0.0410	[-0.0970, +0.0060]	0.918
silent	GPT-5.6	small world N15	-0.0451	[-0.0967, -0.0043]	0.878
silent	GPT-5.6	hub N15	-0.0689	[-0.1544, +0.0139]	0.592
silent	naive	ring N10	+0.0068	[+0.0000, +0.0204]	0.571
silent	naive	small world N15	+0.0354	[+0.0012, +0.0841]	0.612
silent	naive	hub N15	+0.0857	[+0.0199, +0.1647]	0.408

The rule's two hub $N = 15$ rows carry the same delta and the same innocent top-1 (their intervals differ in the fourth decimal, which is the bootstrap draw) because its final ranking is unchanged in 48 of the 49 seeds whether the adversary is made honest or silenced; the two arms are separate runs and the adversary seat's statements differ in every seed (Source results/verify_2026-09-04/armtwin_2026-09-14.py). The tally rule has positive point changes in every cell. Under honest replacement, the model improves in the ring, has an interval including zero in small-world, and deteriorates at the hub. Under silencing, all three model point changes are negative, with only the small-world interval excluding zero. The registered clause interprets a nonpositive model change as evidence against the adversary being the main obstacle. The interventions establish a narrower result: removing its random dishonest policy does not consistently improve the rank of the designated seat. That seat remains the scoring target even when it follows an honest rule, so these comparisons do not identify the relative causal contributions of deception and honest-seat errors. The adversary contributes 4.5–7% of post-opening false statements in the base environment; a share of false statements is not a share of performance loss.

Statement composition of the base environment, by round, pooled over the eleven configurations and the four models: the adversary's share of the false statements a seat hears.

Round	Llama-3.1-70B	GPT-5.6	Gemini 3.1 Pro	DeepSeek V4
2	0.045	0.045	0.045	0.045
3	0.057	0.060	0.061	0.058
4	0.063	0.067	0.070	0.066
5	0.065	0.070	0.072	0.067

Source: results/peer1_analyses_2026-08-30/r2c_premise_split.json (stream block).

With the scored seat emitting no distinguishing behavior, this arm measures who tracks the opening round's truthful signal better, which is the tally rule's own decoding task.

THE SEAT SWAP, PER CELL

Hub-to-spoke within hub-and-spoke, same seeds, same graph; the rule occupies the same two seats.

Occupant	N	hub MRR	spoke MRR	cost	rule cost	n
Llama-3.1-70B	15	0.556	0.340	0.215	0.029	49
GPT-5.6	15	0.662	0.430	0.233	0.029	49
Gemini 3.1 Pro	15	0.674	0.381	0.294	0.029	49
DeepSeek V4	15	0.799	0.489	0.309	0.029	49
Llama-3.1-70B	20	0.775	0.417	0.357	0.070	49
GPT-5.6	20	0.803	0.429	0.374	0.070	49
DeepSeek V4	20	0.893	0.445	0.448	0.070	49
Gemini 3.1 Pro	20	0.798	0.343	0.455	0.070	49

The rule's cost is one number per size because the rule's game does not depend on which model it is compared against; it is generated by replaying the scripted environment with the rule in each seat. Every cell costs the model more than the rule, in eight of eight.

THE ROSTER INTERVENTIONS

Question. Whether the neighbor bias of §3 is a property of the model's reasoning or an artifact of the prompt, which in every arm above names the seat's direct neighbors by number and gives no roster for anyone else.

Bundled arm C: design. One run, DeepSeek V4, two configurations at the ends of the audibility range used throughout (ring $N = 10$, small-world $N = 15$), the same 49 seeds as the main campaign, 98 games and 980 model calls in a single window (2026-09-04, 04:54–06:09 UTC; the run manifest records the requested alias, the served model string, the sampling parameters and every request id). The prompt change is the bundle defined above. Everything else (graph draws, adversary placement, scripted seats, scoring) is the main campaign's.

Decision rule, fixed before any game was read. The comparison is the audibility effect against the shadow tally, the same quantity and the same comparator as §3.5, measured on these two configurations for this model, where the original condition gives 0.369 (0.341 and 0.396). Four branches were written down and committed: the headline stands if the symmetric condition reaches at least half of that, 0.184, with an interval excluding zero; the headline moves to the relay result alone if the interval excludes zero below that threshold; the interface account wins and the paper is rebuilt around it if the point estimate falls below the threshold with an interval spanning zero; and the verdict is withheld pending two further configurations if the point estimate clears the threshold but the interval does not. The fourth branch was added while the run was in progress and before any output was read, because the first three did not separate no effect from too few games. Auxiliary quantities (neighbor occupancy, roster coverage) were declared as description and excluded from the decision.

Outcome. 0.310 with a 95% interval of $[0.083, 0.553]$, resampled over seeds within each configuration since there are only two of them: the first branch. The paired difference against the original condition is -0.059, interval $[-0.220, +0.105]$. Two registered scripts resample this arm and their intervals differ in the third decimal, as the split's do for the reason B.26 gives: the figures here are the roster arm's own script, and the shadow recomputation of the same condition returns $[0.081, 0.553]$ and $[-0.224, +0.103]$. Sources results/verify_2026-09-04/roster_split_2026-09-14.py and results/verify_2026-09-04/condition_C_shadow_2026-09-04.py. Computed instead against a co-run tally the arm gives 0.549 against 0.494, a paired difference of $+0.055$ with interval $[-0.049, +0.156]$; the sign of the difference is not stable across comparators and neither is distinguishable from zero, which is the finding. The shadow figures use the comparator of §3.5.

Scope of C. Passing this registered threshold establishes a remaining audibility effect under the bundle. It does not separate roster and premise effects, and the paired interval does not rule out a substantial reduction. A premise-only result on another model cannot identify which part of this bundle matters.

Roster-only arm R: design and registration. R retains the original premise and changes only the roster line. DeepSeek V4 and Llama-3.1-70B each ran 98 games: ring $N = 10$ and small-world

$N = 15$, 49 seeds per configuration. The retained graphs and roles match the original condition A at every paired seed. The DeepSeek registration preceded R; its later Llama amendment followed the DeepSeek readout and preceded the Llama run. The supplementary snapshots preserve both stages. The primary readout is the change in the probability that top-1 is a neighbor, restricted to games where the adversary is not a neighbor. This differs from the share of top-k slots used elsewhere. There are 75 paired eligible games per model. Secondary seat-mention coverage is the fraction of other player IDs appearing anywhere in the final raw Call2 text, over all 98 pairs; it does not measure whether those seats were evaluated as candidates. Intervals below resample paired game differences pooled over the two configurations (8,000 draws).

Model	Readout	A	R	R minus A [95% CI]
DeepSeek V4	Neighbor top-1	0.867	0.933	$+0.067$ [-0.013, $+0.147$]
Llama-3.1-70B	Neighbor top-1	1.000	0.987	-0.013 [-0.040, 0.000]
DeepSeek V4	Seat-mention coverage	0.964	1.000	$+0.036$ [$+0.012$, $+0.066$]
Llama-3.1-70B	Seat-mention coverage	0.965	0.999	$+0.034$ [$+0.010$, $+0.064$]

Registered interpretation and its limits. DeepSeek's registration explicitly withholds a directional verdict when the interval includes zero; that branch applies here. A positive point estimate is insufficient to select the alternative branch rejecting the interface account. The Llama amendment expected a coverage increase exceeding DeepSeek's, based on a low-coverage estimate from a broader configuration set. Its actual baseline in these two configurations is already 0.965, and the increase of 0.034 misses the registered threshold of more than 0.036. We retain that failed manipulation check, and do not treat Llama as confirmation under the registered gate. Neighbor preference remains high under a full roster, but these results do not exclude an interface contribution or isolate a reasoning mechanism. Sources: results/verify_2026-09-04/roster_arm_R_2026-09-05.py and its registered Llama entry, roster_arm_R_llama_2026-09-19.py. The scripts also report the registered top-3 and raw-mention-subset checks; the latter is a proxy for parser nonintervention, not proof of it.

Exploratory placement check in A, R and C. We apply B.25's raw-order decomposition to the retained DeepSeek games in these two configurations. Simulation settings, graphs, roles, adversary identities and other seats' opening statements match at each paired seed; later messages may differ. Each arm has 75 out-of-view games. The first three rows use each arm's complete outputs; the last three use the 65 games complete in all three arms. Gap is shadow-minus-model mean reciprocal rank, with the shadow scoring the two statement rounds before judgment.

Sample	Arm	Games	Gap	Placement share [95% CI]
Arm complete	A: original	67	0.183	86% [72, 105]
Arm complete	R: roster	74	0.188	103% [97, 113]
Arm complete	C: roster + premise	73	0.228	91% [81, 101]
Common complete	A: original	65	0.181	85% [71, 105]
Common complete	R: roster	65	0.199	102% [93, 113]
Common complete	C: roster + premise	65	0.221	90% [81, 99]

Placement contributes most of the mean gap in the bundled arm on both samples. Shares are signed and can exceed 100% when ordering favors the model. The intervals use 8,000 resamples of the 49 shared seed clusters across the two fixed configurations (NumPy seed 20260919). These are arm-level intervals, not tests of unchanged effects or paired differences. Selection on complete outputs, including the common sample's dependence on all three arms, remains a limitation. On all 75 out-of-view C games, the stored-order gap is 0.222 [0.151, 0.294]; that estimate uses a different rank definition and population from the decomposition. The R/C run-level manifests record temperature zero and thinking off, but sent prompts and a common dated model checkpoint are not established. This later reanalysis is separate from C's registered audibility decision and does not establish transfer beyond one model and two configurations. Source: results/verify_2026-09-04/interface_placement_2026-09-19.py (`interfaceplace`); arm provenance and sampling details are in the supplementary analysis notes.

E.4 The LLM-neighbor arm

Design. The objection this answers is that the relayers are scripted, so the deficit might be a property of scripted relay rather than of relay. We test one cell (ring $N = 10$, GPT-5.6, 49 seeds, batch), with the seats whose words reach the evaluated seat directly (86 of them across

the 49 games) occupied by language models instead of rules. The graph, the seed, the adversary's placement and behavior, and the evaluated seat itself are unchanged; a gate checks each game's generated graph and adversary against the baseline cell's game at the same seed and refuses to run on any mismatch, and all 49 match, so the comparison is paired. 1,350 calls. Sources results/verify_2026-09-04/llm_neighbors_split_2026-09-10.py and, for the two readings below, llm_neighbors_channel_2026-09-10.py and neighbor_accusation_by_variant_2026-09-10.py.

Outcome, and what it does and does not extend. The out-of-view deficit keeps its sign and its interval: -0.171 [-0.266, -0.095] against -0.204 [-0.320, -0.098] in the baseline cell, which was the criterion fixed before the arm ran. The audibility split, however, is not evaluated by this arm, and reporting it as reproduced or as failed would both be wrong. Decomposing the difference shows why: out of view the seat's top-1 is exactly zero in both arms, so the movement from -0.204 to -0.171 is entirely the rule's; and in view the seat falls from 0.500 to 0.000 while the rule falls from 0.538 to 0.078, both below the 0.111 a random pick would give, so that half has no working reading on either side. The narrow interval that half produces is degeneracy, not precision.

Changes in the statements. In the final two-round statement window, each arm contains 196 neighbor statements across 49 games. Statements targeting the evaluated seat rise from 100 to 160, while those naming the adversary fall from 33 to 8; the adversary is mentioned in 19 games with rule neighbors and 5 with model neighbors. Among the 37 out-of-view games, the tally rule ties all candidates in 5 games under rule neighbors and 21 under model neighbors. These window-specific counts do not describe the LLM's entire carried state. Source: results/verify_2026-09-04/matched_received_2026-09-20.py.

Across all five generated rounds, the rules the arm displaced accuse the seat in 45.0%, 32.9% and 25.0% of their statements against the model seats' 86.3%, so filling every displaced slot with the highest-accusing rule would reach 209.5 accusations against the 387 observed. The adversary's own seat, untouched, reads 26.7% in both arms. So the arm extends the level claim past scripted relayers, and separately shows that model relayers spend the channel on the listener. One model in one configuration, and the arm changes which rules are present as well as who holds the seats, so both readings are directional.

E.5 The split with the opening convention diluted

We split the opening-truth arm at $p = 0.75$ by audibility within each configuration, using GPT-5.6 and the shadow comparator; the hub is excluded for lack of an out-of-view half. The gate reproduces the registered per-cell splits at $p = 1.0$ to three decimals. At ring $N = 10$ the out-of-view level goes -0.183 to -0.071 while the split goes $+0.124$ to $+0.135$; at small-world $N = 15$ the level goes -0.108 to -0.070 and the split $+0.172$ to $+0.096$. The sign of the split survives in two of two configurations and no interval excludes zero in either.

This arm is printed because it does *not* support one argument we had considered making. That the level's magnitude is environment-dependent while the split is not cannot be read off it: the split retains 109% and 56% of its value where the level retains 39% and 65%, which is not a consistent difference in either direction. The split is assessed through separate sensitivity checks across specified comparators (B.26), scales (B.27), tie policy (B.28) and stratum definition (B.29), with limits and exceptions reported in each analysis. This arm does not establish environment-invariance. One model, two configurations. Source results/verify_2026-09-04/audibility_p075_2026-09-10.py.

E.6 The corrected premise, component by component and model by model

§3.4 reports the GPT-5.6 comparison. Applying the same paired decomposition to the corrected-premise arm in every model with retained relay cells gives the table below. Each cell is paired with its baseline by seed after asserting that the graph, roles and criminal placement match; the population is out-of-view games whose raw ranking is complete in both arms, and intervals resample seed clusters. These are observed differences between arms, not causal effects: the shadow rule reads each arm's own stream, so its score moves as well.

Provenance. Every row uses retained corrected-premise relay runs paired to their corresponding main-campaign baselines. For DeepSeek, Llama and Gemini these are the later extensions, which ran at temperature 0.0 and record sampling, model and run metadata in a per-game manifest; they

are not the earlier temperature-mismatched cells discussed in B.22, which lie outside the relay configurations. The GPT-5.6 row uses the batch runner's corrected-premise cells, which that runner pins at 0.0 and which carry per-game manifests in five of the seven relay cells, and is the comparison reported in §3.4.

model	cells	complete n	base	corr.	Δ placement	Δ ordering
GPT-5.6	7	224/224	0.118	0.092	$+0.002$ [$-.019$, $+.023$]	-0.028 [$-.047$, $-.011$]
DeepSeek	7	200/224	0.226	0.194	$+0.006$ [$-.033$, $+.043$]	-0.038 [$-.058$, $-.019$]
Llama	7	131/224	0.220	0.190	-0.019 [$-.049$, $+.010$]	-0.011 [$-.031$, $+.010$]
Gemini	3	49/107	0.150	-0.004	-0.161 [$-.234$, $-.097$]	$+0.007$ [$-.071$, $+.089$]

GPT-5.6 and DeepSeek show lower ordering contributions without detectable placement changes; neither component change is clear for Llama. In Gemini, the placement contribution decreases and the mean shadow-minus-model gap changes from 0.150 to -0.004. These arm-specific profiles are heterogeneous descriptive evidence, not an identified mechanism or a formal test of between-model differences.

The four rows are not read on comparable populations. Complete-output coverage runs from every out-of-view game for GPT-5.6 to fewer than half for Gemini, which also has three retained relay cells against seven for the others because a spend cap left two of its cells short (B.22). The model whose result differs most is therefore the model whose sample is most heavily selected, and the magnitude and the generality of that row should be read with that in mind.

Source results/verify_2026-09-04/intervention_components_2026-09-21.py. Exploratory and post-hoc: the arm's own registration governs its registered endpoints, not this decomposition.

The arm in which the environment is not the comparator

Every scripted honest seat in the main campaign aggregates the way the comparator does: it counts accusations and subtracts defenses. A reader may then take the rule's lead to be a property of the environment's generator rather than of the occupant (§3.1). Appendix B, B.26 answers that for the *judge*, by reading the same statements with six decision rules, three outside the tally's family. This arm answers it for the *environment*.

The scripted honest seats speak by a rule the tally cannot be written as: an accusation counts for as much as its speaker has been steady, one over one plus the number of distinct names that speaker has used, and defenses are not read at all. Nothing else moves. The adversary is the same coin-flip script, the evaluated seat is the same model, the opening round is unchanged so the environment's truth injection is the same, the statement grammar is unchanged so the shadow tally reads the stream it always did, and the seeds and criminal placement are identical cell by cell. The comparator stays the tally.

Two configurations, 49 seeds each, Llama-3.1-70B in the evaluated seat. The split is $+0.597$ against the baseline's $+0.505$ on the same cells, and it is larger in both of them rather than smaller: $+0.641$ against $+0.489$ at small-world $N = 10$ and $+0.553$ against $+0.520$ at cycle-heavy $N = 10$. A neighbor named the adversary in 25 of the out-of-view games against 27 in the baseline. The new rule therefore retained games with a delivered accusation naming the adversary, satisfying the event-supply criterion written before the run. This does not establish that the available claims uniquely identified the adversary.

What this is worth, and what it is not. The interval quoted for an arm of two configurations is a bootstrap over two points and excludes zero for the trivial reason that both are positive; it is not an interval in the sense the rest of this paper uses one. One model. Outside the frozen preregistration, so it is *exploratory* and is not entered into the registry of §2. It narrows the objection rather than closing it: both rules still accumulate over accusations, so what was tested is whether the split needs the comparator to share the environment's weighting, not whether it needs the environment to accumulate at all. Why the arm's split is larger than the baseline's is not in the record and is not guessed at here.

The implementation's design note records two earlier discarded candidates: scoring accusations against otherwise unsuspected seats, and scoring the number of distinct targets a speaker named. Their pilot outcomes cannot be reconstructed from retained records, so they supply no quantitative

evidence here. The note also records the decision to exclude defenses: after the opening round, only the adversary defends, giving defense-reading rules the diagnostic cue examined in B.5 and B.37. Source for the retained steady-rule analysis: results/verify_2026-09-04/arm_steady_split_2026-09-17.py; the arm's runner is run_arm_steady_llama_2026-09-17.py and the seat is agents/f_i_steady.py.

F THE INTERVENTION PROBE

This probe measures how three warning prompts change paired outcomes across models and configurations. Variation in those outcomes does not by itself identify a mechanism that enables or blocks an intervention.

F.1 A generic warning. The generic warning cautions against treating inconsistency across a neighbor's statements as evidence of deception[4]. Its label, warn_a, comes from the design-time hypothesis described in this appendix. Under the registered net-five top-1 criterion, four Llama cells and two Gemini cells are responsive; none is harmful. GPT-5.6 and DeepSeek have no responsive cells. These 44-cell maps include the discovery model and are exploratory (Figure C.2).

The conditional calibration in Appendix A gives an expected 3.98 responsive-or-harmful cells, a 95% reference interval of 1 to 8, and $P = 0.190$ for at least the observed six. Counting only responsive cells gives $P = 0.011$; the signed-count calibration gives $P = 0.006$. The latter summaries suggest a directional improvement under the calibration assumptions, while the total count does not exceed its reference interval. All are post-hoc, and cell independence is an assumption.

Twenty-four of the 44 cells have enough discordant pairs to cross the criterion; six of these are responsive. On the 36 non-hub cells, the corresponding counts are 17 reachable and five responsive. These outcome-selected denominators describe where the criterion could fire; they are not a new confirmatory sample. The fixed threshold also does not determine a universal minimum detectable effect: power depends on the probabilities of both improvement and deterioration.

F.2 Targeted warnings. Structure disclosure (warn_s) states that the model hears only part of the network; behavioral targeting (warn_c) states that an accusation of oneself does not identify its speaker as the adversary. The registered disclosure comparison excludes the two hub configurations, where the limited-view statement is false. Four of its 36 non-hub cells are responsive, among 15 reachable cells; GPT-5.6 supplies none of the reachable cells. On the same 36 cells, disclosure has a larger net improvement than the generic warning in 11 cells, a smaller one in 19, and a tie in six (two-sided sign test treating cells as independent, $P = 0.200$). This does not support the predicted superiority of disclosure. Variation across cells does not establish that structure controls intervention effectiveness.

The behavioral warning has one responsive cell, Llama at hub $N = 15$, among 12 tested cells; none is harmful. There are 71 improved and 43 deteriorated top-1 outcomes, averaging 9.5 discordant pairs per cell. For context, the generic warning averages 6.6 across 44 cells and disclosure 5.2 across its 36 non-hub cells; these are different configuration sets. Eleven behavioral-warning cells have net changes within three games of zero, while the responsive cell has a net gain of 21. A cell can change many outcomes without crossing the net-five threshold.

An exploratory reading of the behavioral target counts only accusations from neighbors generated before the final judgment. The rate of choosing an accuser among accused games decreases in 10 of 12 cells and increases in two, with differences ranging from -0.307 to $+0.068$. There is no preregistered numerical threshold for this metric. The accused populations can change across arms, so these conditional differences do not establish a causal effect among a fixed group of games. Source: results/verify_2026-09-04/warning_probe_2026-09-20.py.

Changes relative to the rule. Warning effects are not confined to the net-five labels. Using each warning's mean MRR against the retained co-run rule benchmark, the counts of configurations below the rule are:

[4]The full texts of the generic warning, structure disclosure and accusation warning are reproduced in Appendix A.

model	original	generic warning	structure disclosure
Llama-3.1-70B	10/11	8/11	9/11
GPT-5.6	9/11	9/11	8/11
Gemini 3.1 Pro	9/11	9/11	7/11
DeepSeek V4	9/11	9/11	9/11

These are exploratory point-estimate sign counts, including the two hub cells for disclosure; they do not replace the original campaign's registered k. The behavioral warning covers only three configurations, so it does not define an eleven-configuration count. Warnings change some signs, while every model remains below the rule in most configurations under the two full-coverage warnings.

A separate ablation corrects two false premise lines in the prompt (§2). In the original GPT-5.6 minimal cells, accuser targeting falls by 24.5 and 5.3 percentage points under the receipt-time definition; the second interval includes zero (Appendix E). The warnings also change behavior in some cells. These experiments do not establish a general ranking of premise correction and warnings.

F.3 Possible interventions. The results motivate further tests of seat assignment, prompt information and communication protocols. Seat assignment could change which agents receive information directly, but its effect has not been tested here. Disclosure and behavioral warnings have mixed effects, including some improvements; their scattered responsive cells do not identify why an intervention succeeds in one configuration and fails in another. A channel carrying reasoning or source information would also differ from the verdict-only statements used here.

For the protocol family we tested its most direct instance, a query channel by which the evaluated seat may ask any other player for its current top suspect, with the reply delivered privately in the next round. A registered pilot established that the model selects the query action on its own (existence only; Appendix D). A separately preregistered main run then asked whether the channel changes same-seed paired MRR in two configurations, isolated pairs N = 10 and hub-and-spoke N = 10, with 34 seeds each, disjoint from the pilot's. In both, the registered verdict is "no detected response under our design": in isolated pairs 5 pairs improved, 7 deteriorated, and 22 tied; in hub-and-spoke 4, 7, and 23; the registered threshold was 4 of 34. Ties dominate (45 of 68 pairs; Appendix D, Table D.2). The registration records that its threshold is an analogy to §2's five-of-forty-nine rule, which counts top-1 outcomes rather than MRR, so the two are not directly comparable in stringency.

The query result is specific to that channel, model and pair of configurations. The proposed changes to seat assignment and richer message content remain untested.

G THE TIE-POLICY COUNTERFACTUAL

The main campaign resolves rule ties toward the lowest player number. Because the evaluated seat is player 1, other seats often accuse it when their tallies tie. This appendix tests how changing that convention changes the environment and the model's performance relative to a shadow tally.

Design. In hub-and-spoke $N = 10$ and isolated pairs $N = 20$, comparison runs retain the main campaign's tie policy and treatment runs randomize tied choices uniformly. The patch applies to the honest rules and the parser's completion of omitted candidates. Each arm uses three repetitions of the same seeds. These runs are separate from the main campaign. There are 1,199 treatment games and 1,200 comparison games; excluding seed 0 leaves 1,175 and 1,176. Treatment rep 2 for Gemini at the hub lacks seed 4900, as recorded in the protocol.

Registered odds-ratio test. Protocol v7 fixes the decision rule before the recorded inspection of comparison-arm results, after treatment runs had completed. For each game, an event is a final top-1 choice of a neighbor that accused the evaluated seat at any generated round. The outcome is whether that final choice is wrong. All non-event games, including games without an accuser, form the reference group. This is a whole-game association; it is not restricted to accusations delivered before the final choice and does not measure a response to a received accusation.

Within each repetition, the two configurations are pooled and the odds ratio uses the Haldane correction, adding 0.5 to each contingency-table cell. The comparison point is the mean of its three repetition-level odds ratios. The treatment interval is its mean plus or minus 4.303 times the standard error across three repetitions. A model is judgeable if the treatment interval's full width does not exceed the comparison point. The registered prediction requires at least three models' treatment

intervals to contain their comparison points; fewer than two judgeable models makes the overall result undecided. These are the protocol's operational criteria, not a test of equivalence.

model	comparison mean OR	treatment mean OR [95% CI]
Llama-3.1-70B	6.182	0.127 [+0.096, +0.157]
GPT-5.6	6.170	0.095 [+0.081, +0.110]
Gemini 3.1 Pro	0.208	0.039 [−0.004, +0.082]
DeepSeek V4	4.249	0.061 [−0.045, +0.167]

All four models pass the width criterion, and none of the treatment intervals contains its comparison point. The registered prediction that the odds ratio would be maintained is therefore not supported. All four treatment means are lower. The negative lower endpoints arise from the registered arithmetic t interval on the OR scale with only three repetitions; they do not denote negative odds ratios.

Exploratory MRR comparison. The shadow tally scores the two statement rounds preceding the final judgment. To compare the streams under a common scoring convention, both arms average the reciprocal rank over a uniform ordering of tied candidates. With a candidates above the adversary's tie group and t candidates in that group, its score is $t^{-1}\sum_{i=1}^{t}(a+i)^{-1}$. In the comparison arm this tie-averaged shadow differs from the realized rule with seat-number tie breaking. No rule occupied the evaluated seat in these runs.

For each seed, scores are averaged over its available repetitions, then over the 49 seeds. The same weighting is used for the absolute levels and the model-minus-shadow gaps. These comparisons are exploratory and carry no preregistered verdict.

	hub N10		isolated pairs N20	
model	model MRR	shadow MRR	model MRR	shadow MRR
Llama-3.1-70B	0.808 → 0.764	0.683 → 0.758	0.223 → 0.220	0.203 → 0.181
GPT-5.6	0.739 → 0.837	0.682 → 0.799	0.220 → 0.220	0.214 → 0.192
Gemini 3.1 Pro	0.749 → 0.861	0.710 → 0.803	0.204 → 0.210	0.201 → 0.195
DeepSeek V4	0.761 → 0.738	0.697 → 0.801	0.227 → 0.224	0.214 → 0.201

Each entry gives comparison → treatment. The shadow score rises at the hub and falls in isolated pairs in all four models. The model-minus-shadow difference decreases at the hub for three models and increases for Gemini; it increases in all four isolated-pairs cells. Thus the effect depends on configuration. These data do not support a single claim that the convention uniformly helps either the model or the rule.

The gap intervals below resample seeds after averaging their repetitions (5,000 bootstrap draws, 49 seeds per cell). This preserves each seed's repeated observations as a cluster.

model	configuration	comparison	treatment
Llama-3.1-70B	hub N10	+0.125 [+0.038, +0.219]	+0.006 [−0.025, +0.035]
GPT-5.6	hub N10	+0.057 [−0.020, +0.134]	+0.038 [−0.021, +0.099]
Gemini 3.1 Pro	hub N10	+0.039 [−0.031, +0.109]	+0.058 [−0.004, +0.124]
DeepSeek V4	hub N10	+0.064 [−0.019, +0.154]	−0.064 [−0.124, −0.001]
Llama-3.1-70B	isolated pairs N20	+0.020 [−0.051, +0.103]	+0.039 [−0.035, +0.124]
GPT-5.6	isolated pairs N20	+0.005 [−0.083, +0.104]	+0.028 [−0.065, +0.126]
Gemini 3.1 Pro	isolated pairs N20	+0.003 [−0.087, +0.097]	+0.015 [−0.071, +0.104]
DeepSeek V4	isolated pairs N20	+0.013 [−0.062, +0.096]	+0.022 [−0.056, +0.109]

At the hub, the randomized-policy intervals include zero for Llama, GPT-5.6 and Gemini; DeepSeek's interval is below zero. Only Llama's comparison-arm interval is above zero. The paired treatment-minus-comparison changes are -0.120 $[-0.217, -0.026]$ for Llama and -0.128 $[-0.218, -0.043]$ for DeepSeek; the other two hub intervals include zero. All isolated-pairs gap and arm-difference intervals include zero. Inclusion of zero does not establish equal performance. The comparison here uses a shadow tally, whereas §3.1 uses a co-run tally; hub $N = 15$ was not tested under the randomized policy.

Accusations generated. Across all generated rounds, the share of games in which a neighbor accuses the evaluated seat falls from 0.959 to 0.850 and from 1.000 to 0.299 (Llama), 0.857 to 0.721 and 1.000 to 0.265 (GPT-5.6), 0.878 to 0.795 and 0.313 to 0.116 (Gemini), and 0.946 to 0.830 and 1.000 to 0.224 (DeepSeek), at the hub and in isolated pairs respectively. These are means over

three repetitions. They describe generated accusations, including the final round, and do not show that the model read or acted on them. The OR, MRR and accusation-rate calculations are in results/verify_2026-09-04/tie_policy_verified_2026-09-20.py.

THE SAME POLICY IN THE CONFIGURATIONS THAT RELAY

The hub hears every seat, and isolated pairs have no relay path. Ring $N = 10$ and small-world $N = 15$ therefore provide a separate test of whether the out-of-view deficit remains under randomized ties. Both arms use the same 49 seeds per configuration and the same schedule, giving 98 games per model per arm and 75 paired out-of-view games. The outcome is the model's top-1 minus the shadow tally's top-1, averaging uniformly over tied first places in both arms. The intervals resample configuration–seed game pairs (8,000 draws); they do not cluster a seed across configurations.

model	registered ties	randomized ties	difference [95% CI]
Llama-3.1-70B	-0.252 [-0.338, -0.171]	-0.322 [-0.422, -0.228]	-0.070 [-0.163, $+0.017$]
GPT-5.6	-0.158 [-0.231, -0.091]	-0.184 [-0.268, -0.109]	-0.027 [-0.112, $+0.056$]
Gemini 3.1 Pro	-0.204 [-0.290, -0.125]	-0.194 [-0.287, -0.113]	$+0.009$ [-0.057, $+0.071$]
DeepSeek V4	-0.317 [-0.414, -0.228]	-0.304 [-0.400, -0.213]	$+0.012$ [-0.080, $+0.107$]

Every randomized-policy interval lies below zero. The out-of-view deficit therefore persists under both tested policies in all four models. Every arm-difference interval includes zero, which leaves the size of the policy's contribution uncertain; it does not establish equivalence. This meets the protocol's branch requiring a below-zero treatment interval in at least three judgeable models. The retained relay protocol records an amendment to the branch-specific power clause before the recorded inspection of arm outputs. Source: results/verify_2026-09-04/tie_relay_arm_2026-09-09.py.

Serving windows. The design pairs arms run on the same schedule, so the windows are recorded. All times UTC. Gemini 3.1 Pro, both arms 2026-09-08 17:32 to 18:47, batch. GPT-5.6, both arms 2026-09-08 22:40 to 2026-09-09 02:40, batch. DeepSeek V4, both arms 2026-09-08 17:32 to 20:01, in real time during the provider's off-peak window. Llama-3.1-70B, local, both arms 2026-09-08 17:32 to 2026-09-09 05:53. For each model the two arms therefore sit inside one window of a few hours, which is the only checkpoint defense available for cells whose provider exposes no dated snapshot; the relay-configuration arms summarized in §4.2 carry the same record.