SAME ANSWER, DIFFERENT REASONS: DISENTANGLING COORDINATION FROM CONSENSUS IN MULTI-AGENT DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model agents in a multi-agent debate can reach the same final answer for two different reasons: because they saw the same evidence, or because a coalition among them privately coordinated to produce it. From the verdict alone the two cases are indistinguishable, letting a coordinated minority pass as independent consensus. We cast this as an identifiability problem: recovering a latent partition of agents into coalition and non-coalition members from public information alone (transcript and shared task dossier). To address it, we introduce **LATTICE**, a Layered Agreement Tracing framework that grounds each message into atomic claims, judges typed, directed relations between claims of different agents, and aggregates them into time-windowed, multi-lens agent-by-agent matrices under a composite with pre-specified weights. A stance-correlation baseline is kept out of the composite, and every result is checked against a paired negative-control session with the same dossier and roster but no coalition. Under this protocol, the composite separates the latent agent partition more reliably than stance correlation, does not recover the transferred partition in the negative controls, achieves the strongest pairwise ranking among public-information detectors, and, when the coalition size is known, recovers the installed coalition exactly, matching group-structure detectors without their high false-structure rates. On external benchmarks it keeps its ranking advantage where stance correlation falls to chance, recovers coalitions well above chance but not exactly, and does not mistake honest, information-driven convergence for coordination. Code and data are available here.

1 INTRODUCTION

Large Language Models (LLMs) have enabled multi-agent systems that coordinate through language on difficult tasks (Park et al., 2023; Du et al., 2023), and debate among agents is often used to expose hidden flaws and aggregate diverse reasoning (Liang et al., 2024; Chan et al., 2023). That promise, however, is fragile. When agents share a backbone, prompting style, or retrieval pipeline, their mistakes can become correlated, and a debate may then stabilize error rather than wash it out.

We study this failure mode through the lens of *insider collusion*: a subset of agents privately pursues an objective different from the one stated to the room and produces a coherent, mutually reinforcing consensus around it. Prior work shows that such consensus is not a reliable signal of correctness (Zhang et al., 2025; Wu et al., 2024; Coppolillo et al., 2025), but the harder problem sits one level below correctness. Consider two agents, A1 and A2, who both vote to reject deploying a clinical alert. In one debate, each independently cites the same trial result. In another, A2 first argues for deployment and then, round after round, adopts A1's unsupported rationale. The final votes are identical; the histories that produced them are not.

A detector that treats a transcript as a bag of final stances cannot see this difference, and a coordinated minority can exploit this blind spot to pass as an independent majority. What separates the two histories is information that stance-level views discard: which claim answered which, whether a claim was echoed or contested, and whether an agent adopted another's reasoning several rounds later. We therefore treat collusion as a problem of *measurement* before it is a problem of correction:

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.

before acting on a suspected coalition, a system must recover this structure from the public transcript and task dossier every agent already has, without collapsing the debate into one pairwise number.

We introduce *Layered Agreement Tracing for deTecting Inter-agent Coordination in dEbate* (LATTICE), which is designed to measure exactly this difference. Given a transcript, LATTICE (i) splits every message into a few checkable claims, extracted without access to any faction label, (ii) judges how each new claim treats the claims before it, using directed, typed relations that keep semantic compatibility and argumentative reinforcement separate, (iii) aggregates these judgments into time-windowed agent-by-agent matrices under a composite with pre-specified weights that leaves out the stance-correlation baseline, and (iv) checks the result against a paired negative control, a session with the same dossier and roster but no coalition. Each step exists because the one before it cannot resolve the ambiguity on its own, and together they form one protocol that is falsifiable by design.

Our contributions are threefold:

1. **Problem.** Final agreement cannot identify coordination. We formulate collusion detection as recovering a latent agent partition from public reasoning traces (Section 2), under a paired negative control and a baseline-exclusion constraint fixed in advance.
2. **Method.** LATTICE (Section 3) preserves directed claim-to-claim interactions instead of collapsing a debate into stance similarity, keeping every level traceable to specific claims.
3. **Evidence.** On paired sessions, against eighteen detectors, and on four external benchmarks (Section 4), this added structure separates the latent partition more reliably than stance correlation without inflating the negative control, and recovers the coalition exactly when its size is known. It also exposes where the signal is weaker: a narrow ranking margin over directed-influence statistics, partial external recovery, and a failed pipeline gate.

LATTICE borrows decomposition from scalable oversight (Bronikowski, 2025; Lifshitz et al., 2025) and typed support-versus-attack relations from computational argumentation (Dung, 1995; Baumeister et al., 2021), but uses both for measurement rather than verification (Appendix D).

2 Problem Formulation

Our question differs from truth discovery: rather than asking whether a debate's conclusion is correct, we ask whether the debate is being steered by a coordinated subset of its own participants, using only public information: the transcript and the task dossier that every participant receives. In this setting, the detector has no private agent state or model internals, cannot interrogate intent, has no claim labels, verification budget, or tool access, and does not know any agent's reliability in advance. A hidden coalition is thus a fact about a whole session, not about any single message.

2.1 Layered Observation of a Multi-Agent Debate

A debate session consists of N agents $\mathcal{A} = \{a_1, \dots, a_N\}$ exchanging public messages over R synchronous rounds. Each message m of the transcript T carries an agent $\mathrm{agent}(m) \in \mathcal{A}$, a round $\mathrm{round}(m) \in \{1, \dots, R\}$, and text. To expose the content of each message, a *claim extraction map* χ grounds it into a small claim set $\chi(m) \subseteq \mathcal{C}$, $|\chi(m)| \leq 2$, so that $\mathcal{C} = \bigcup_{m \in T} \chi(m)$ is the level-0 set of atomic, checkable propositions; each $c \in \mathcal{C}$ inherits $\mathrm{message}(c)$, $\mathrm{agent}(c)$, $\mathrm{round}(c)$, a claim type, and an evidence status assessed against a fixed dossier rather than the speaker's confidence.

Claims are then linked by a *relation judge* ρ, a partial function over ordered claim pairs from consecutive rounds and distinct agents: for $\mathrm{round}(c_s) = k - 1$, $\mathrm{round}(c_t) = k$, and $\mathrm{agent}(c_s) \neq \mathrm{agent}(c_t)$, $\rho(c_s, c_t)$ is either $\perp$ (no meaningful relation) or a typed, scored relation drawn from $\{$ADOPTS, SUPPORTS, EXTENDS, QUALIFIES, REBUTS, UNDERCUTS$\}$. Realized relations form the level-1 set are defined as:

$$\mathcal{R} = \big\{ (c_s, c_t, \mathrm{type}, \sigma, \alpha, \kappa) \;:\; \rho(c_s, c_t) \neq \perp \big\}, \qquad \sigma, \alpha \in [-1, 1],\ \kappa \in [0, 1], \tag{1}$$

where σ is a semantic compatibility score, α an argumentative reinforcement score, and κ judge confidence. Grouping $\mathcal{R}$ by $(\mathrm{message}(c_s), \mathrm{message}(c_t))$ yields level-2 message relations, and grouping further by $(\mathrm{agent}(c_s), \mathrm{agent}(c_t))$ yields level-3 agent relations; both remain sparse, undefined rather than zero wherever no relation was judged. Section 3 operates on this four-level representation, and no level may be estimated by skipping the level beneath it.

2.2 Threat Model: Coordinated Reasoning versus Independent Agreement

The threat is behavioral and structural; detecting it needs no access to private intent. Formally, assume a latent partition $\varphi : \mathcal{A} \to \{0, 1\}$, fixed before the session and never publicly revealed, where $\varphi(a) = 1$ marks membership in a coalition $\mathcal{U} = \{a : \varphi(a) = 1\}$, $|\mathcal{U}| \geq 2$, that privately pursues an objective different from the one publicly stated. The detector's recovery target is then the pairwise indicator $Y(a, b) \triangleq \mathbb{1}[\varphi(a) = \varphi(b)]$, which is 1 exactly when a and b share a side of the partition.

The key difficulty is that final public verdicts only weakly identify Y. To see why, let $S(a, b)$ denote the observable event that a and b reach the same decision by the debate's end. Two failure directions make S an unreliable proxy for Y: independent agents exposed to the same dossier can converge on $S(a, b) = 1$ with $Y(a, b) = 0$ (honest consensus that looks like coordination), and a coalition can stagger or hedge its members' public stances to produce $S(a, b) = 0$ while $Y(a, b) = 1$ (coordination that looks like disagreement). A detector that measures S alone therefore cannot tell *same answer* from *coordinated reasoning*; that requires the claim and relation levels of Section 2.1, where reinforcement, adoption, and shared evidence use are observable regardless of final stances.

2.3 Objective: Identifiability-Controlled Faction Recovery

A *measurement statistic* is any function f that maps a transcript T, with its induced $(\mathcal{C}, \mathcal{R})$ hierarchy, to a symmetric matrix $M \in \mathbb{R}^{N \times N}$ (or a family of such matrices indexed by time window) without access to φ. We evaluate M against two complementary targets. The first is pairwise ranking,

$$\mathrm{AUROC}(M) \;\triangleq\; \Pr\big(M(a, b) > M(a', b') \;\big|\; Y(a, b) = 1,\; Y(a', b') = 0\big), \tag{2}$$

It is estimated over the same-partition and cross-partition agent pairs of a session, with ties broken uniformly. The second is global partition recovery: for a known coalition size k, split $\mathcal{A}$ into a bipartition $\hat{\varphi}$ of sizes $(k, N - k)$ from M and report $\mathrm{ARI}(\hat{\varphi}, \varphi)$. We report the two separately, since a statistic can rank coalition pairs above chance without inducing a correct global split.

Two constraints keep this objective falsifiable rather than self-confirming. The first is a *paired negative control*: every assigned session T_1 (coalition $\mathcal{U}$ installed) is paired with a neutral session T_0 that shares its dossier and roster but has no coalition. The same f is applied to both; because T_0 has no coalition, we transfer T_1's partition indices onto T_0 as a pseudo-label and report the resulting scores as a *false-structure rate*, so a statistic is informative about coordination only if its assigned-session recovery exceeds this rate, not merely chance. This rate tests only recovery of the transferred partition; it does not rule out other spurious groupings of the neutral agents. The second is a *baseline-exclusion constraint*: any hierarchical statistic M_{full} must omit the stance-correlation statistic M_{stance} from its inputs, which satisfies:

$$\mathrm{AUROC}(M_{\mathrm{full}}) \;>\; \mathrm{AUROC}(M_{\mathrm{stance}}) \tag{3}$$

This is a claim about structure added by the claim and relation levels, not one made true by absorbing the baseline.

3 Methodology

Section 2.2 showed that a shared verdict cannot separate honest convergence from coordination: agreement is observable, but coordination is latent. We posit that the difference lies in how reasoning moves through the debate, in the path by which claims are asserted, reused, and reinforced across turns, and that collapsing this path into one pairwise number destroys it. LATTICE therefore builds the path up level by level (Figure 1), each level answering a question the one beneath cannot while remaining an auditable compression of it, and holds the result to pre-specified gates (Phase 5).

3.1 Phase 1: Grounding Messages into Atomic Claims

Messages are too coarse to measure: a single turn can restate an earlier point, introduce a new one, and hedge a third. We therefore ground every message into at most two self-contained propositions, each with a claim type, cited evidence, and an evidence status assessed against the dossier rather than the speaker's confidence, so a confidently wrong agent cannot pass as a well-supported one. In addition, the extractor never sees a faction label and must cover every message, including rhetorical ones, so that later structure does not depend on who wrote parseable text.

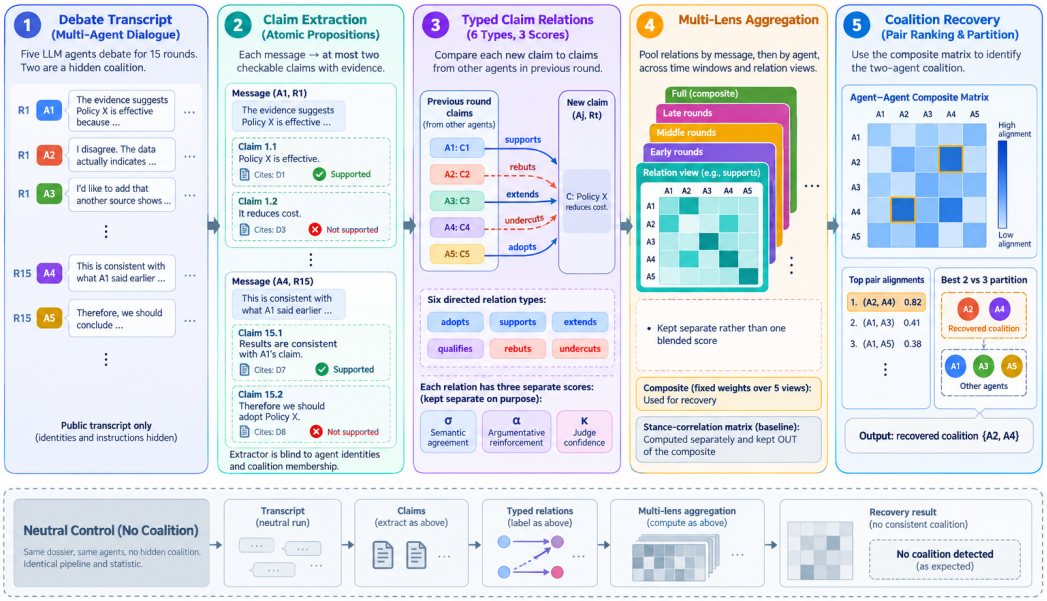


Figure 1: Overview of LATTICE. Messages are grounded into atomic claims, cross-agent relations between claims are typed and scored, and these aggregate into time-windowed, multi-lens agent-by-agent matrices. The composite drives pairwise coalition ranking and partition recovery; the stance-correlation baseline is measured but withheld from it.

3.2 Phase 2: Directed Inter-Agent Claim Relations

Claims alone still do not show how agents react to one another. The central move is therefore to ask, for every claim, not whether it agrees with the eventual conclusion but how it treats the claims that came just before it. Concretely, each claim is compared with the claims every other agent asserted in the previous round, and the strongest cross-agent relation is labeled with one of six directed types (ADOPTS, SUPPORTS, EXTENDS, QUALIFIES, REBUTS, UNDERCUTS). Same-agent relations are excluded, since coalitions concern *between*-agent structure.

Each accepted relation $r = (c_s \to c_t)$ carries three scores rather than one: semantic compatibility $\sigma(r) \in [-1, 1]$, argumentative reinforcement $\alpha(r) \in [-1, 1]$ (whether c_t reinforces or attacks the warrant of c_s), and judge confidence $\kappa(r) \in [0, 1]$. Keeping σ and α apart is deliberate: a rebuttal can be on-topic yet argumentatively hostile, and a single signed agreement value would erase the surface-disagreement-but-latent-coordination pattern a sophisticated coalition can produce. Beyond these scores, two further channels come from the same relation pool without extra judge calls. The first, *temporal adoption*, rewards ADOPTS, SUPPORTS, and EXTENDS relations, capturing directed uptake of one agent's claim by another over time rather than static similarity. The second, *rare shared error*, groups claims into components under strong semantic overlap (a union-find over relations with $\sigma(r) \geq 0.60$ and $\kappa(r) \geq 0.50$) and, for relations linking two claims whose evidence status is contradicted or uncertain, weights the relation by how small its component is relative to the full claim population. The intuition is that two honest agents converging on a well-supported claim is unremarkable, whereas two agents converging on the *same idiosyncratic mistake* is unlikely under independent reasoning and therefore informative about coordination.

3.3 Phase 3: Hierarchical Aggregation Without Premature Collapse

Individual relations do not yet reveal persistent group structure: a single adoption can be incidental, whereas a coalition shows up as a pattern that recurs between the same agents. Everything above the claim relations is therefore a controlled act of compression. First, relations are aggregated by source and target message into message-level relations, augmented with observable stance agreement and evidence overlap; claim pairs with no judged relation are treated as missing rather than as neutral agreement, since imputing agreement would dilute the few relations that carry signal. Message-level

relations are then aggregated into directed agent-by-agent matrices for early, middle, late, and full-debate windows, so that a coalition that coordinates only once the room has converged is not washed out by an early phase of genuine independent reasoning. Formally, for a lens ℓ and window w, the entry is the confidence-weighted mean of the lens's per-relation score over relations whose target claim falls in w, i.e., with $\mathrm{round}(r) \equiv \mathrm{round}(c_t) \in w$:

$$M_\ell^w[i,j] \;\triangleq\; \frac{\sum_{r:\,\mathrm{src}(r)=i,\,\mathrm{tgt}(r)=j,\,\mathrm{round}(r)\in w} \kappa(r)\,\ell(r)}{\sum_{r:\,\mathrm{src}(r)=i,\,\mathrm{tgt}(r)=j,\,\mathrm{round}(r)\in w} \kappa(r)}, \tag{4}$$

left undefined (missing, not zero) when the denominator vanishes, and symmetrized, $\frac{1}{2}(M_\ell^w + M_\ell^{w\top})$, only for steps that need an undirected view. We define seven lenses (Table 4 in Appendix A.1), each targeting a distinct way a coalition can look innocuous under a coarser measure, and compute six of them as agent-by-agent matrices: behavioral stance co-movement, propositional semantics, argumentative alignment, shared evidence selection, temporal adoption, and shared unsupported claims. Five enter the window-specific composite

$$M_{\mathrm{full}}^w \;\triangleq\; 0.30\,M_{\mathrm{sem}}^w + 0.25\,M_{\mathrm{arg}}^w + 0.15\,M_{\mathrm{evi}}^w + 0.20\,M_{\mathrm{temp}}^w + 0.10\,M_{\mathrm{rare}}^w. \tag{5}$$

The weights are heuristic and were fixed a priori, before any coalition label or outcome metric was inspected, so they encode a structural hypothesis rather than the result of optimization over lens combinations. Semantic and argumentative signals receive the largest mass because they provide the primary content- and reasoning-level representation, while temporal adoption receives substantial weight because directed uptake is central to the coordination hypothesis. Evidence overlap is down-weighted because it can arise naturally from common dossier exposure, and rare shared error receives the smallest weight because it is intended as a sparse, high-specificity signal that should not dominate the composite. We do not claim these coefficients are optimal. Two lenses stay outside the composite. Stance correlation is computed in the same windowed manner but kept separate, so any improvement over it cannot arise from directly incorporating the baseline itself. The outcome lens, which relates agents to the final decision and its harm, is not an agent-pair signal: it requires privileged labels (the reference decision and the coalition's target), so it is never used for detection. Appendix A.1 explains why this hierarchical composition, rather than a flat pairwise score, is needed.

3.4 Phase 4: Debate-Level Structure Recovery

Pairwise alignment does not by itself identify a coherent coalition, since a matrix can single out one aligned pair yet leave the group split ambiguous. The full-window composite $M \equiv M_{\mathrm{full}}^{\mathrm{full}}$ therefore feeds two separate procedures that instantiate the targets of Section 2.3. *Pairwise ranking* estimates Equation (2) with the tie-corrected Mann–Whitney statistic over same-partition and cross-partition pairs, for which one half is the natural chance level. Alongside this AUROC, it reports the rank of the installed coalition pair among all $\binom{N}{2}$ pairs, a stricter, session-level test in which rank 1 means the coalition is the most aligned pair in the session. *Partition recovery* searches exhaustively over every candidate coalition U' of the known size k and selects

$$\hat{\mathcal{U}} \;\triangleq\; \operatorname*{arg\,max}_{U'\subset\mathcal{A},\;|U'|=k} \Big[\tfrac{1}{|W(U')|}\textstyle\sum_{(a,b)\in W(U')} M(a,b) \;-\; \tfrac{1}{|X(U')|}\sum_{(a,b)\in X(U')} M(a,b)\Big], \tag{6}$$

where $W(U')$ holds the pairs on the same side of the split and $X(U')$ the pairs across it; we report $\mathrm{ARI}(\hat{\varphi}, \varphi)$ for the induced partition $\hat{\varphi}(a) = \mathbb{1}[a \in \hat{\mathcal{U}}]$. This exhaustive search is feasible because the $\binom{N}{k}$ candidates are few at the $N = 5$, $k = 2$ scale studied here, whereas larger rosters would need an approximate partitioning heuristic. Appendix A.2 gives the estimators in details.

3.5 Phase 5: Reporting and Falsifiability Gates

To keep these results falsifiable, we hold LATTICE to five feasibility gates, fixed before any session in Section 4 was scored; Appendix A.3 states each one formally. Gates 1 and 2 check pipeline integrity: at least 90% of messages must receive a genuine claim extraction, and fewer than 2% of candidate claims and relations may reference a nonexistent message, claim, or evidence item. Gate 3 requires every matrix entry to trace back to the claim pairs that produced it, which holds by construction because Equation (4) is an explicit sum over relation identifiers. Gate 4 asks whether the composite

beats stance correlation for the right reason. Beating it on assigned sessions alone is not enough, since a statistic could be equally inflated on the paired neutral session. Writing AUROC_1 and AUROC_0 for assigned and neutral sessions and $\Delta(M) \triangleq \mathrm{AUROC}_1(M) - \mathrm{AUROC}_0(M)$ for a statistic's margin over its own false-structure rate, gate 4 requires both

$$\mathrm{AUROC}_1(M_{\mathrm{full}}) > \mathrm{AUROC}_1(M_{\mathrm{stance}}) \qquad \text{and} \qquad \Delta(M_{\mathrm{full}}) > \Delta(M_{\mathrm{stance}}). \tag{7}$$

Finally, gate 5 requires the margin $\Delta(M_{\mathrm{full}})$ to be positive in every one of the forty dossiers, not only when pooled, and correct partition recovery not to be confined to a single installed coalition identity. Any failed gate is reported as a result, not smoothed over in the writing.

4 Experiments

This section asks, in turn, whether the pipeline is trustworthy enough to interpret, whether the hierarchy beats stance correlation, how it compares with detectors built on other signals, which lenses carry the separation, and whether the comparison holds on benchmarks we did not construct. Supporting tables and figures are in Appendix B.

4.1 Design and Setup

We instantiate LATTICE on forty self-contained decision dossiers. Each specifies a binary public decision, a public objective stated to every agent, and eight evidence items of mixed reliability and mutual tension; examples include deploying an emergency-department sepsis alert, releasing a reservoir ahead of an uncertain forecast, and procuring grid-scale batteries under degradation risk. Each dossier is run twice with the same five agents: in a *neutral* session, all agents independently pursue the public objective; in an *assigned-coalition* session, a private two-agent coalition ($|\mathcal{U}| = 2$) argues for the decision opposite the dossier's evidence-supported one, never revealing its identity. This gives eighty sessions of $R = 15$ rounds and 75 messages each (6,000 in total). Debate generation, claim extraction, and relation judging all use GPT-5.6 Luna Max, and neither the extractor nor the judge knows which session is neutral or which agents coordinate. Details are in Appendix B.1.

4.2 Semantic Pipeline Quality and Feasibility Gates

We first check the pipeline itself. Extraction coverage and traceability pass (Table 6), but the invalid-reference gate fails: more candidate claims and relations than the pre-specified ceiling allows cite a nonexistent message, claim, or evidence item. We nevertheless retain the pre-specified threshold and report the failure. Because rejected objects are dropped rather than repaired, the failure removes evidence rather than inserting spurious relations, though we have not ruled out that rejections concentrate on particular agents or conditions. One additional diagnostic concerns matching: in the three representative dossiers examined in Appendix B.2, assigned sessions yield fewer claims per message than their neutral pairs, so the conditions are not matched in extraction volume, though this means the assigned margin does not come from extracting more claims.

4.3 Main Results: Pairwise Ranking and Partition Recovery

With the pipeline checked, Table 1 tests the central claim: composing claim-level relations into a hierarchy separates same-partition from cross-partition pairs better than stance correlation, without a higher false-structure rate in neutral sessions. Lenses are added in the pre-specified order, and each statistic is judged against its own neutral rate; Figure 3 (Appendix B.3) plots it.

The hierarchy beats stance correlation in assigned sessions. The full composite satisfies both conditions of gate 4 (Equation (7)): it ranks same-partition pairs above cross-partition pairs more reliably than stance correlation, and its margin over its own neutral rate is wider. Crucially, the neutral sessions make this margin interpretable: in the paired neutral controls, no statistic recovers the transferred partition above chance, so the advantage is not a property of the pipeline or of the shared dossier. Moreover, only the composite ranks the installed coalition first, a stricter, session-level test.

Table 1: Pairwise ranking and partition recovery on assigned sessions and paired neutral controls. False-structure scores measure recovery of the assigned-session partition in the neutral pair; 0 indicates chance-level recovery.

	Assigned sessions		Neutral (false-structure) sessions		
Statistic	**Pair AUROC ↑**	**Partition ARI ↑**	**Pair AUROC − 0.5 → 0**	**Partition ARI → 0**	**Coalition pair rank ↓**
Raw Pearson stance correlation M_{stance} (baseline)	0.63	0.167	0.00	0.00	2
Semantic only	0.61	0.60	0.00	0.00	3
+ Argument	0.65	0.60	0.00	0.00	2
+ Evidence	0.67	0.60	0.00	0.00	2
+ Temporal adoption	0.71	0.80	0.00	0.00	2
Full LATTICE composite M_{full} (+ rare shared error)	**0.78**	**1.00**	0.00	0.00	**1**

The signal lives beneath the verdict. Read from top to bottom, Table 1 traces the argument of this paper. Verdict-level agreement is not enough, and neither is propositional overlap: the semantic lens alone ranks pairs no better than stance correlation, because agents who read the same dossier assert similar propositions whether or not they coordinate. Separation appears once relations are typed by argumentative role and anchored to the evidence, and grows once they are ordered in time; temporal adoption is also the first lens to improve partition recovery. The rare-shared-error lens then sharpens the final separation with the largest single gain, as the design predicts: because the coalition argues against the evidence-supported decision, two members converging on the same unsupported claim is what independent reasoning makes least likely. This last gain is, however, partly a product of our intervention, so we read shared unsupported claims as a final sharpening signal rather than a universal signature of collusion. The composite already beats stance correlation before this lens is added, and Section 4.6 tests whether its advantage holds on coalitions designed by other authors.

Partition recovery follows the ranking. A good ranking, however, need not yield a correct global split, and for stance correlation it does not: the baseline ranks pairs above chance but rarely recovers the installed partition. By contrast, every lens-based statistic recovers it far more often, recovery improves as lenses are added, and, given the known coalition size, the full composite recovers the coalition exactly in every dossier. Exact recovery arrives with the rare-shared-error lens, the same lens that makes the coalition the top-ranked pair, and the two are linked: the split search averages agreement over all within-group pairs, so it prefers the true split only once the coalition pair stands clearly apart from honest agents who happen to agree. Conversely, on neutral sessions the recovered split matches the transferred grouping no better than chance.

4.4 Comparison with Existing Detection Methods

While Table 1 isolates what the hierarchy adds over stance correlation, Table 2 compares LATTICE with eighteen detectors built on other signals, in six families, all run on the same transcripts and scored by the same procedure of Section 3.4, so differences reflect the detection signal rather than the evaluation. Capability columns mark directed statistics, which can separate an agent who steers from one who is steered, claim traceability (gate 3), and access limited to public information. Among them, the published detectors of Bonjour et al. (2022), Zhou et al. (2025), and Ramamoorthy et al. (2026) share our access. By contrast, the activation probes of Rose et al. (2026), the hidden-state spectral partitioning of Berg et al. (2026), and the regret audit of Nakamura et al. (2026) need model internals or a task's ground-truth optimum; we include them as an upper reference, not as methods LATTICE is expected to beat. Figure 4 (Appendix B.4) plots each against its own neutral control.

Directionality carries much of the ranking signal; recovery separates the composite. Among methods with the same access, the composite achieves the strongest pairwise ranking, but its lead over directed-influence statistics is narrow: lead-lag and Granger-style influence come close, and their gap between assigned and neutral sessions is comparable. Much of the coordination signal is therefore directional (who takes up whose claims, and when), which agrees with the gain from temporal adoption in the ablation. The difference appears after ranking: directed statistics order pairs almost as well as the composite but do not reliably recover the installed partition, whereas the composite recovers it exactly. In other words, influence alone does not separate the coalition from truth-seeking agents who influence one another honestly; the rare-shared-error lens does.

Table 2: Comparison with competing coordination detectors. *Dir.*, *Claim*, and *Public* denote directed signals, claim-level traceability, and public-information access. Neutral scores report false-structure AUROC relative to chance (0 = chance).

					Assigned sessions		Neutral
Family	**Method**	**Dir.**	**Claim**	**Public**	**Pair AUROC ↑**	**Partition ARI ↑**	**False-struct. AUROC − 0.5 → 0**
Behavioral correlation	Raw Pearson stance correlation (baseline)			✓	0.63	0.167	0.00
	Session-demeaned Pearson			✓	0.58	−0.25	0.00
	First-difference Pearson			✓	0.63	−0.25	0.00
	Spearman rank correlation			✓	0.54	−0.25	0.00
	Pairwise mutual information			✓	0.54	−0.25	0.00
Confound-controlled	Excess correlation vs no-communication			✓	0.63	0.167	−0.02
	Residual (partial) correlation			✓	0.58	0.167	−0.02
Directed influence	Lead-lag cross-correlation	✓		✓	0.72	0.167	−0.02
	Granger-style directed influence	✓		✓	0.75	0.167	−0.02
	Spectral clustering on addressing graph	✓		✓	0.70	−0.25	0.00
Semantic representation	Sentence-embedding cosine similarity			✓	0.58	0.167	+0.05
	LLM-as-judge direct prompting			✓	0.55	0.167	+0.02
Published (transcript)	Trajectory MI (Bonjour et al., 2022)			✓	0.72	1.0	+0.05
	GUARDIAN (Zhou et al., 2025)	✓		✓	0.65	1.0	+0.08
	CEMPAT (Ramamoorthy et al., 2026)			✓	0.65	1.0	+0.05
	Colosseum (Nakamura et al., 2026)				0.75	1.0	+0.05
Published (internal repr.)	NARCBench (Rose et al., 2026)				0.80	1.0	−0.02
	Spectral MI Probe (Berg et al., 2026)				0.78	1.0	−0.02
Ours	**LATTICE (full composite)**	✓	✓	✓	**0.78**	**1.0**	**0.00**

Undirected similarity is a safe but weak signal. The behavioral-correlation family does not recover the transferred partition on neutral sessions, but it ranks same-partition pairs only modestly above chance and rarely, if ever, recovers the partition; confound control keeps the neutral rate at chance without improving the ranking. Semantic similarity does worse. In particular, embedding similarity scores above chance on neutral sessions, because agents who read the same dossier converge on the same vocabulary whether or not they coordinate, while directly prompting an LLM judge stays near chance, reflecting the conflation of agreement and coordination that motivates this work.

The composite matches group detectors on recovery, without their elevated false-structure rates. The published public-information detectors also recover the partition exactly while ranking pairs less well. However, the two built to find groups or anomalous nodes directly, through trust-graph clustering and temporal-graph anomaly scoring, carry among the highest false-structure rates in the table: they partly recover the transferred partition even without a coalition. The composite, in contrast, reaches the same recovery with its neutral rate at chance. As a result, among public-information methods it is the only one that combines the strongest ranking, exact recovery, and a false-structure rate at chance, and the only one whose flagged pairs trace back to specific claims and relations.

Privileged access adds little over public information here. The activation-based methods lead or match on every column, as expected given their access. Yet the composite matches the hidden-state spectral probe on ranking and recovery from public information alone, suggesting that here the coordination signal those methods read from internal states is largely expressed in what agents say.

4.5 Lens Ablation

Each lens adds signal without raising the false-structure rate. To see which lenses drive these results, Table 7 (Appendix B.5) decomposes the composite into the marginal contribution of each lens in the pre-specified order. Because the weights were committed before scoring, the decomposition describes a fixed statistic rather than a search over weightings. It shows that every lens adds ranking signal and none removes it, the neutral rate stays at chance at every stage, and the gains concentrate in the lenses for directed uptake over time and shared unsupported claims, the two properties designed to distinguish coordinated uptake from independent agreement over shared evidence.

4.6 Generalization Across Established Benchmarks

Our dossiers fix a known coalition and an exact neutral control, but we constructed them ourselves. We therefore evaluate the full composite and the stance baseline on four established benchmarks

Table 3: Generalization to established benchmarks not authored or tuned on by us. *Coalition g.t.* marks whether the benchmark has a genuine coalition partition; where it does not, only false structure is measured, so pair AUROC should stay near 0.5 and the last columns near 0. Arrows as in Table 1.

			Full composite		Stance baseline	
Benchmark	**Coalition g.t.**	**Sessions used**	**Pair AUROC↑**	**Partition ARI↑ / false-struct.→0.**	**Pair AUROC↑**	**Partition ARI↑ / false-struct.→0.**
Secret Mafia (Wang et al., 2026)	Yes	100	0.74	0.48	0.52	0.02
RogueAgent (Ren et al., 2025)	Yes	100	0.69	0.41	0.53	0.00
MultiAgentBench (Zhu et al., 2025)	No (neutral analogue)	100	0.51	0.02	0.50	0.00
HiddenBench (Li et al., 2026)	No (hard negative control)	100	0.52	0.03	0.51	0.01

(Table 3; details in Appendix B.6), none of which was used to design the composite. Two of them, Secret Mafia (Wang et al., 2026) and RogueAgent (Ren et al., 2025), supply genuine coalitions, in free-form social-deduction dialogue and in a large-population social-media simulation. The other two, MultiAgentBench (Zhu et al., 2025) and HiddenBench (Li et al., 2026), contain no coalition and serve as large-scale negative controls. HiddenBench is the harder control: agents holding complementary evidence converge honestly, in a way a detector could mistake for coordination.

On both coalition benchmarks, the composite retains a clear ranking advantage, slightly below its level on our dossiers, while stance correlation falls to chance. The contrast is strongest on Secret Mafia, where coalition members avoid voting or arguing visibly together, matching the second failure direction of Section 2.2: coordination that looks like disagreement and is therefore invisible to stance-level signals but still expressed through how claims are adopted and reinforced. Partition recovery also transfers, though less completely: it remains well above chance where stance correlation recovers nothing, but is no longer exact, likely because RogueAgent involves substantially larger populations and Secret Mafia rewards concealing alignment in free-form dialogue. Importantly, this increased sensitivity does not come at the cost of mistaking honest convergence for coordination. On the two benchmarks without a coalition, neither statistic exhibits elevated false structure; this is especially informative on HiddenBench, where agents converge because they hold complementary information, yet the composite remains near chance.

5 Conclusion

The central result is not that agreement is suspicious, but that identical agreement can arise from measurably different interaction histories. LATTICE makes those histories observable by tracing how claims move between agents over time, using public information alone. On the paired sessions, the composite beats stance correlation without recovering the transferred partition in neutral sessions, gains most from directed claim uptake and shared unsupported claims, and is the only public-information method combining the strongest ranking, exact recovery in the known-k evaluation setting, and a false-structure rate at chance. Its advantage also carries over to external benchmarks, even where stance correlation falls to chance, without mistaking honest convergence for coordination.

Limitations. All results use the full data, but as point estimates without confidence intervals, so differences of a few hundredths of AUROC are not decisive. Beyond this, four points bound our claims. First, the ranking lead over directed-influence statistics is narrow; the clearer advantages over them are exact recovery and claim-level traceability. Second, exact recovery does not fully transfer to larger, free-form populations, where exhaustive split search also becomes intractable. Third, the pipeline fails one integrity gate, and we have not checked where rejected references concentrate across agents or conditions. Fourth, the largest gain comes from shared unsupported claims, which our intervention makes especially informative, and the negative control tests only the transferred partition. More broadly, LATTICE operates only on public information and therefore cannot recover coordination that leaves no observable footprint in claims, reinforcement, or timing; it also does not independently verify claim truth. The most direct next steps are scaling group recovery beyond exhaustive search and testing further backbones (Appendix B.7).

AI USE STATEMENT

Apart from the LLM agents that are the object of study in our experiments, we used generative AI tools only to refine the wording, grammar, and readability of the text in this paper. These tools were not used to generate or modify any experimental result; every number reported in the paper comes from the authors' own experimental runs. We have reviewed all AI-assisted text and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies how to detect covert coordination among LLM agents from public information, namely their transcripts and the shared task dossier. All debate participants are LLM agents, and the study involves no human subjects and no human annotation. The forty decision dossiers were written for this study, and the four external benchmarks are used as released by their authors. A detection method could in principle help agents learn to evade it; to keep LATTICE from being over-trusted as a safeguard, we state plainly which forms of coordination it cannot detect (Section 5, Appendix D). LATTICE produces statistical evidence of coordination, not proof of intent, and should not be the sole basis for consequential decisions about an agent or its operator.

REPRODUCIBILITY STATEMENT

Source code and data for all experiments are available here. Section 3 specifies the full pipeline, Appendix A gives the exact estimators and the formal feasibility gates, and Appendix C restates the pipeline as pseudocode. The composite weights (Equation (5)) and all gate thresholds were fixed before any session was scored. The study design, backbone model (GPT-5.6 Luna Max), and session and call counts are given in Section 4.1 and Appendix B.1, and the external benchmarks, with the number of sessions used from each, are described in Section 4.6 and Appendix B.6.

REFERENCES

Adeela Bashir, The Anh han, and Zia Ush Shamszaman. Many-to-one adversarial consensus: Exposing multi-agent collusion risks in ai-based healthcare, 2025. URL https://arxiv.org/abs/2512.03097.

Dorothea Baumeister, Matti Järvisalo, Daniel Neugebauer, Andreas Niskanen, and Jörg Rothe. Acceptance in incomplete argumentation frameworks. *Artificial Intelligence*, 295:103470, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103470. URL https://www.sciencedirect.com/science/article/pii/S0004370221000217.

T.J.M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10):619–641, 2007. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2007.05.001. URL https://www.sciencedirect.com/science/article/pii/S0004370207000793. Argumentation in Artificial Intelligence.

Cameron Berg, Susan L. Schneider, and Mark M. Bailey. Hidden coalitions in multi-agent ai: A spectral diagnostic from internal representations, 2026. URL https://arxiv.org/abs/2605.06696.

Trevor Bonjour, Vaneet Aggarwal, and Bharat Bhargava. Information theoretic approach to detect collusion in multi-agent games. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 223–232. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/bonjour22a.html.

Michael Bronikowski. Scalable oversight in multi-agent systems: Provable alignment via delegated debate and hierarchical verification. In *Open Conference of AI Agents for Science 2025*, 2025. URL https://openreview.net/forum?id=l5Wrcgyobp.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL https://arxiv.org/abs/2308.07201.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning, 2026. URL https://arxiv.org/abs/2510.07517.

Erica Coppolillo, Giuseppe Manco, and Luca Maria Aiello. Unmasking conversational bias in ai multiagent systems, 2025. URL https://arxiv.org/abs/2501.14844.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate, 2025. URL https://arxiv.org/abs/2509.11035.

Wentao Deng, Jiahuan Pei, Zhiwei Xu, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Belief-calibrated multi-agent consensus seeking for complex NLP tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AYqtMLRwzj.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. ISSN 0004-3702. doi: https://doia.org/10.1016/0004-3702(94)00041-X. URL https://www.sciencedirect.com/science/article/pii/000437029400041X.

Mohammad Sajjad Ghaemi. A survey of collusion risk in LLM-powered multi-agent systems. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=Ylh8617Qyd.

Anthony Hunter, Sylwia Polberg, Nico Potyka, Tjitze Rienstra, and Matthias Thimm. *Probabilistic Argumentation: A Survey*. 08 2021. ISBN 978-1-84890-336-4.

Jamiu Idowu, Ahmed S. Almasoud, and Ayman Alfahid. Mapping human anti-collusion mechanisms to multi-agent ai systems. *Knowledge-Based Systems*, 344:116067, 2026. ISSN 0950-7051. doi: 10.1016/j.knosys.2026.116067. URL http://dx.doi.org/10.1016/j.knosys.2026.116067.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.606. URL https://aclanthology.org/2025.findings-acl.606/.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. DEBATE: Devil's advocate-based assessment and text evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1885–1897, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.112. URL https://aclanthology.org/2024.findings-acl.112/.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic failures in collective reasoning under distributed information in multi-agent llms, 2026. URL https://arxiv.org/abs/2505.11556.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024. URL https://arxiv.org/abs/2305.19118.

Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025. URL https://arxiv.org/abs/2502.20379.

Ryan Y. Lin, Siddhartha Ojha, Kevin Cai, and Maxwell F. Chen. Strategic collusion of llm agents: Market division in multi-commodity competitions, 2025. URL https://arxiv.org/abs/2410.00031.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Lewis Hammond, and Christian Schroeder de Witt. A perfect collusion benchmark: How can AI agents be prevented from colluding with information-theoretic undetectability? In *Multi-Agent Security Workshop @ NeurIPS'23*, 2023. URL https://openreview.net/forum?id=FXZFrOvIoc.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bnNSQhZJ88.

Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. Colosseum: Auditing collusion in cooperative multi-agent systems, 2026. URL https://arxiv.org/abs/2602.15198.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Nico Potyka and Richard Booth. An empirical study of quantitative bipolar argumentation frameworks for truth discovery. In *Comma*, 2024. URL https://api.semanticscholar.org/CorpusID:272529550.

Hariprasauth Ramamoorthy, Rajkumar Vaidyanathan, and Suresh Sundaram. *CEMPAT: A Framework to Identify Colluding Agents in a Graph Network*, pp. 190–204. 06 2026. ISBN 978-3-032-25028-5. doi: 10.1007/978-3-032-25029-2_8.

Sarvapali D. Ramchurn, Dong Huynh, and Nicholas R. Jennings. Trust in multi-agent systems. *Knowl. Eng. Rev.*, 19(1):1–25, March 2004. ISSN 0269-8889. doi: 10.1017/S0269888904000116. URL https://doi.org/10.1017/S0269888904000116.

Qibing Ren, Sitao Xie, Longxuan Wei, Zhenfei Yin, Junchi Yan, Lizhuang Ma, and Jing Shao. When autonomy goes rogue: Preparing for risks of multi-agent collusion in social systems, 2025. URL https://arxiv.org/abs/2507.14660.

Qibing Ren, Zhijie Zheng, Jiaxuan Guo, Junchi Yan, Lizhuang Ma, and Jing Shao. When ai agents collude online: Financial fraud risks by collaborative llm agents on social platforms, 2026. URL https://arxiv.org/abs/2511.06448.

Aaron Rose, Carissa Cullen, Sahar Abdelnabi, Philip Torr, Brandon Gary Kaplowitz, and Christian Schroeder de Witt. Detecting multi-agent collusion through multi-agent interpretability, 2026. URL https://arxiv.org/abs/2604.01151.

Marcantonio Bracale Syrnikov, Federico Pierucci, Marcello Galisai, Matteo Prandi, Piercosma Bisconti, Francesco Giarrusso, Olga Sorokoletova, Vincenzo Suriani, and Daniele Nardi. Institutional ai: Governing llm collusion in multi-agent cournot markets via public governance graphs, 2026. URL https://arxiv.org/abs/2601.11369.

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. URL https://arxiv.org/abs/2605.29512.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5163–5186, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.297. URL https://aclanthology.org/2024.findings-emnlp.297/.

Xijie Zeng, Ryan MacArthur, Marta Kryven, and Frank Rudzicz. I know it's unfair, do it anyway: LLM agents exploiting explicitly unfair tools for voluntary collusion in strategic games, 2026. URL https://openreview.net/forum?id=5N2CafbMbJ.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2025. URL https://arxiv.org/abs/2309.01219.

Jialong Zhou, Lichao Wang, and Xiao Yang. Guardian: Safeguarding llm multi-agent collaborations with temporal graph modeling, 2025. URL https://arxiv.org/abs/2505.19234.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.

A Additional Methodology Details

This appendix gives the parts of Section 3 that are condensed in the main text: the rationale for hierarchical composition, the full structure-recovery estimators of Phase 4, and the formal statements of the five feasibility gates of Phase 5.

A.1 Why Hierarchical Composition, and Not a Flat Pairwise Score

A flat detector that scores agent pairs directly from raw transcripts has no internal seam at which to separate "agreed on the answer" from "agreed on how to get there," nor any seam at which to isolate "agreed on a widely supported fact" from "converged on the same unsupported one." The hierarchy exists precisely to create those seams: atomic claims separate propositions from rhetoric, typed directed relations separate reinforcement from rebuttal, multi-lens aggregation separates the failure modes in Table 4 instead of averaging them away, and the deliberate exclusion of stance correlation from the composite keeps the comparison against the natural baseline falsifiable rather than self-fulfilling. Each design choice removes one specific way a coalition could look like an independent majority under a coarser measurement; Table 4 lists the lenses, the failure mode each one is designed to expose, and its role in this study. Conversely, a coalition hypothesis built directly from surface text, without passing through claims and typed relations, is exactly the kind of flattening that the threat model of Section 2.2 shows to be unreliable. As a note on terminology, we call the final object a matrix because of its analysis-ready form, although internally it is a multi-channel temporal tensor.

Table 4: Agreement lenses of LATTICE, the failure mode each one is designed to expose, and its role in this study.

Lens	What it measures	Failure mode controlled	Role in this study
Stance	Decision/confidence co-movement	Same verdict is not enough	Baseline, excluded from composite
Semantic	Equivalent or compatible propositions	Lexical paraphrase	Composite
Argument	Reinforcement versus rebuttal/undercut	Fake surface disagreement	Composite
Evidence	Shared evidence selection	Common dossier exposure	Composite
Epistemic error	Shared unsupported/contradicted claims	Honest consensus	Composite (rare shared error)
Temporal adoption	Later uptake of an earlier claim	Static similarity	Composite
Outcome	Relation to final decision/harm	Coordination without intent	Not used for detection

A.2 Structure-Recovery Estimators

Pairwise ranking. For a session with agent set $\mathcal{A}$, let $P = \binom{\mathcal{A}}{2}$ be the set of all unordered agent pairs, partitioned by the ground-truth indicator into $P_1 = \{(a,b) \in P : Y(a,b) = 1\}$ and $P_0 = \{(a,b) \in P : Y(a,b) = 0\}$. We estimate Equation (2) with its empirical, tie-corrected form,

$$\widehat{\mathrm{AUROC}}(M) \triangleq \frac{1}{|P_1|\,|P_0|} \sum_{p \in P_1} \sum_{q \in P_0} \left(\mathbb{1}[M(p) > M(q)] + \tfrac{1}{2}\mathbb{1}[M(p) = M(q)] \right), \tag{8}$$

which is exactly the Mann–Whitney U statistic for the two groups $\{M(p)\}_{p \in P_1}$ and $\{M(p)\}_{p \in P_0}$, so a value of one half is the natural chance reference rather than an arbitrary threshold. Alongside $\widehat{\mathrm{AUROC}}$, we report a complementary rank statistic for the specific installed coalition pair $\mathcal{U}$: sorting all pairs in P by M in descending order (ties broken by a fixed, arbitrary but consistent rule) and taking

$$\mathrm{rank}_M(\mathcal{U}) \triangleq \left| \{(a,b) \in P \setminus \{\mathcal{U}\} : M(a,b) \geq M(\mathcal{U})\} \right| + 1, \tag{9}$$

so that $\mathrm{rank}_M(\mathcal{U}) = 1$ means the true coalition is the single most mutually aligned pair in the entire session under M. This statistic is deliberately more demanding than $\widehat{\mathrm{AUROC}}$: with only $N = 5$ agents and $|\mathcal{U}| = 2$, a coalition can contribute to a respectable pairwise AUROC while still failing to be the top-ranked pair, and we do not let a favorable aggregate AUROC stand in for this stronger, session-level claim.

Global partition recovery. Given a known coalition size $k = |\mathcal{U}|$, we do not cluster M with a generic algorithm tuned post hoc; we search exhaustively over every candidate bipartition of the

prescribed size and select the one that maximizes the gap between within-group and between-group agreement under M (Equation (6)). Here, for a candidate coalition $U' \subset \mathcal{A}$ with $|U'| = k$, $W(U')$ denotes the unordered pairs with both endpoints in U' or both in $\mathcal{A} \setminus U'$, and $X(U')$ the pairs with one endpoint in each. The recovered coalition induces the partition $\hat{\varphi}(a) \triangleq \mathbb{1}[a \in \hat{\mathcal{U}}]$, against which we report both an exact-match indicator $\mathbb{1}[\hat{\mathcal{U}} = \mathcal{U}]$ and the softer $\mathrm{ARI}(\hat{\varphi}, \varphi)$ of Section 2.3, since exact recovery is a strict criterion that a partially informative statistic can fail while still being far from random. The search space has size $\binom{N}{k}$, which is exhaustively enumerable for the $N = 5, k = 2$ setting in this paper; for larger rosters this optimization would need to be replaced by an approximate partitioning heuristic, and we flag this as an open scaling question rather than one this paper resolves. Reporting pairwise ranking and partition recovery separately matters because Equation (8) can be favorable, and even $\mathrm{rank}_M(\mathcal{U}) = 1$ can hold, without $\hat{\mathcal{U}}$ in Equation (6) equalling $\mathcal{U}$: a matrix can single out the correct pair as most aligned while still preferring a different, incorrect global split once every other pair is taken into account.

A.3 Formal Statements of the Feasibility Gates

We hold LATTICE to five feasibility gates, fixed before any session in Section 4 was scored, and we state each one as a precise, checkable condition rather than a qualitative aspiration.

Gates 1–2: pipeline integrity. Writing $|T|$ for the number of messages in a transcript T, and $\mathrm{ext}(m) \in \{0, 1\}$ for whether message m received a genuine claim extraction rather than a fallback triggered when the extractor returns nothing usable, coverage must satisfy

$$\frac{1}{|T|} \sum_{m \in T} \mathrm{ext}(m) \;\geq\; 0.90, \tag{10}$$

pooled over all eighty sessions of Section 4.1, and the fraction of candidate claim or relation objects rejected for referencing a nonexistent message, claim, or evidence ID must remain below 0.02. Both are pipeline-level checks on χ and ρ (Section 2.1) and are evaluated before any agent-level matrix is trusted.

Gate 3: traceability. Unlike the first two gates, traceability is not a separate measurement but a structural corollary of Equation (4): because every entry $M^w_\ell[i, j]$ is defined as an explicit, finite sum indexed by relation identifiers r, the specific claim pairs (c_s, c_t) underlying any agent-level value are recoverable in constant time per contributing relation, by construction, rather than by a post hoc audit procedure that could fail. Gate 3 is therefore a claim about the aggregation pipeline never discarding these indices, which we verify by construction (Table 6).

Gate 4: does the composite beat stance correlation, and beat it for the right reason. Let $\mathrm{AUROC}_1(\cdot)$ and $\mathrm{AUROC}_0(\cdot)$ denote Equation (8) evaluated on an assigned session T_1 and its paired neutral session T_0 respectively (Section 2.3). Equation (3) alone is not a sufficient gate, because a statistic could exceed M_{stance} on T_1 while being equally inflated on T_0, in which case the apparent advantage would be an artifact of the pipeline rather than evidence about the agents. Gate 4 accordingly requires the conjunction (Equation (7) in compact form)

$$\mathrm{AUROC}_1(M_{\mathrm{full}}) \;>\; \mathrm{AUROC}_1(M_{\mathrm{stance}}), \qquad \text{and} \tag{11}$$

$$\Big[\mathrm{AUROC}_1(M_{\mathrm{full}}) - \mathrm{AUROC}_0(M_{\mathrm{full}})\Big] \;>\; \Big[\mathrm{AUROC}_1(M_{\mathrm{stance}}) - \mathrm{AUROC}_0(M_{\mathrm{stance}})\Big], \tag{12}$$

so that the full composite is credited with an advantage only when it both outperforms the baseline in absolute terms and does so by a wider margin over its own false-structure rate than the baseline achieves over its.

Gate 5: robustness across dossiers. Pooled statistics can pass Equations (11)–(12) while the entire margin is carried by a few dossiers d out of the forty in $\mathcal{D}$ (Section 4.1). Writing $\mathrm{AUROC}_1^{(d)}$ and $\mathrm{AUROC}_0^{(d)}$ for the same statistics restricted to dossier d's session pair, gate 5 requires

$$\min_{d \in \mathcal{D}} \Big[\mathrm{AUROC}_1^{(d)}(M_{\mathrm{full}}) - \mathrm{AUROC}_0^{(d)}(M_{\mathrm{full}})\Big] \;>\; 0, \tag{13}$$

a per-dossier rather than pooled condition, together with the requirement that correct recovery of $\hat{\mathcal{U}}$ (Equation (6)), where it occurs, not be confined to a single installed coalition identity.

A hierarchical measurement construct is only as trustworthy as its willingness to report that a given lens, or the composite itself, failed one of these five gates; consistent with Section 5, we treat such a failure as a legitimate scientific result about where flat correlation remains hard to beat, not as an outcome to be smoothed over in the writing.

B Extended Experimental Results

This appendix reports the full evidence supporting the condensed summaries in Section 4: complete session-level counts, the full semantic-pipeline diagnostic table, the figures for the main results and the detector comparison, the per-lens ablation, the external benchmarks, and further scope notes. (The full competitor table itself is reported directly in Section 4.4, Table 2, rather than here.) Every number here is the same measured value summarized in the main text; nothing in this appendix is a different or re-fit computation.

B.1 Study Design Detail

Table 5: Session-level counts for three representative dossiers out of the forty in the study. Message counts are fixed by design (R=15 rounds $\times$ N=5 agents); across all forty dossiers the study comprises eighty sessions and 6,000 public messages.

Dossier	Condition	Messages	Extraction calls	Relation-judging calls
Sepsis alert deployment	Neutral	75	75	545
Sepsis alert deployment	Assigned coalition	75	75	640
Reservoir pre-release	Neutral	75	75	620
Reservoir pre-release	Assigned coalition	75	75	640
Battery procurement	Neutral	75	75	640
Battery procurement	Assigned coalition	75	75	612
Subtotal (three dossiers)		**450**	**450**	**3697**
All forty dossiers	Both	**6000**	**6000**	**50112**

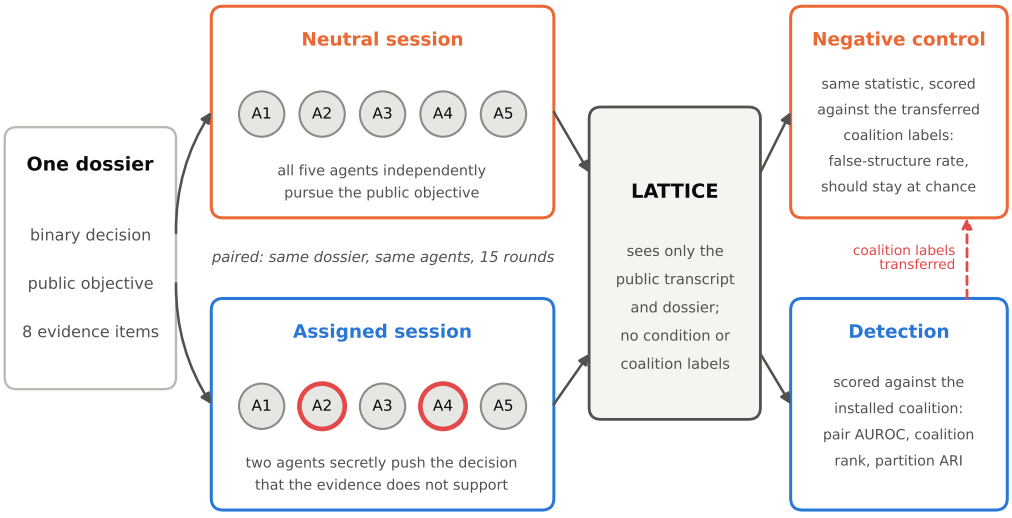


Figure 2: Study design, shown for one dossier. The same dossier and the same five agents are run twice: in the neutral session, all agents pursue the public objective independently; in the paired assigned session, two agents (red rings, never revealed) secretly push the decision the evidence does not support. LATTICE processes both transcripts without any condition or coalition label. The assigned session is scored against the installed coalition (detection), and the neutral session is scored against the same coalition labels transferred from its pair, which gives the false-structure rate (negative control, Section 2.3). The coalition members shown are illustrative.

B.2 Semantic Pipeline Diagnostics Detail

Table 6: Semantic pipeline quality against pre-specified feasibility gates, pooled over all eighty sessions. ↑/↓: higher/lower is better; rows without an arrow are pass/fail checks or diagnostics with no preferred direction.

Diagnostic	Pre-specified gate	Observed	Pass / fail
External claim-extraction coverage ↑	$\geq 90\%$ of messages	0.94	Pass
Invalid claim-reference rate ↓	$< 2\%$	0.04	**Fail**
High-weight edge traceability	every edge traces to raw claims	yes	Pass
Mean claims per message	(diagnostic only)	1.82	n/a
Realized-relation density (relations / candidate pairs)	(diagnostic only)	15.2%	n/a
Mean relation-judge confidence κ ↑	(diagnostic only)	0.81	n/a

The pooled mean is 1.82 claims per message, and every session of the three representative dossiers also averages fewer than two, consistent with the at-most-two-claims extraction rule of Section 3.1.

Invalid-reference gate. Extraction coverage and traceability pass, so the large majority of messages receive a genuine extraction and every matrix entry can be traced back to the claims that produced it. The invalid-reference gate, however, fails: more candidate claims and relations than the pre-specified ceiling allows cite a message, claim, or evidence item that does not exist. We nevertheless retain the pre-specified threshold and report the failure. Because such objects are rejected rather than repaired, the failure removes evidence from the matrices instead of inserting spurious relations. The remaining risk is that rejections concentrate on particular agents or on one condition, which would bias the matrices in a way the pooled rate cannot reveal; we have not yet ruled this out.

Extraction volume across conditions. In each of the three representative dossiers, the assigned session yields fewer claims per message than its paired neutral session, while relation-judge confidence is high and nearly identical across conditions. The neutral control is therefore matched in dossier, roster, and objective, but not in extraction volume. The difference runs against the assigned sessions: denser extraction could sharpen a genuine signal, but here it is the neutral sessions that are denser, so the assigned-session margin does not come from extracting more claims. Unlike these pipeline checks, gate 4 concerns the detector itself and is evaluated in Section 4.3.

B.3 Main Results Figure

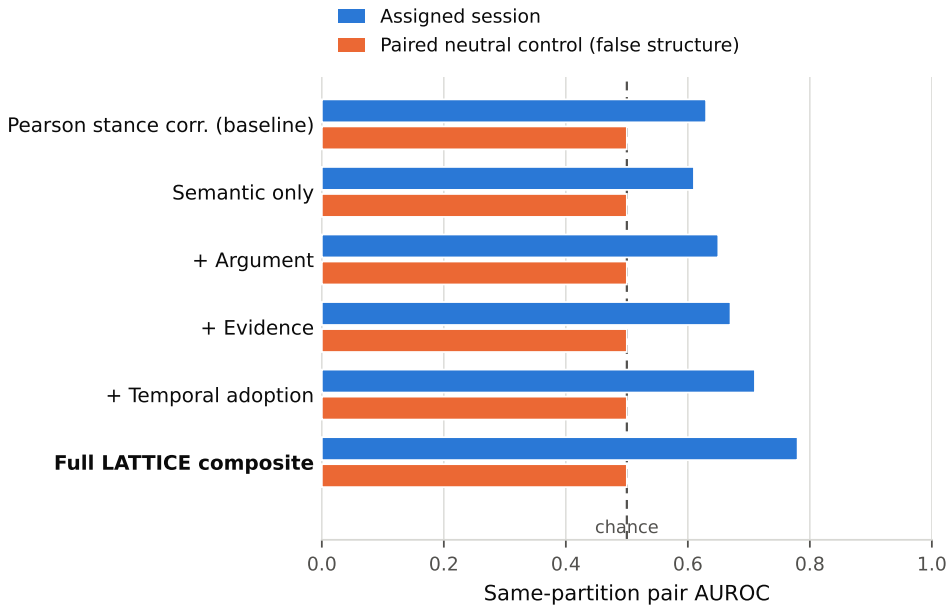


Figure 3: Table 1 shown visually: assigned-session pair AUROC next to its paired neutral false-structure AUROC for the stance baseline and each stage of the composite. A statistic is credited only where the two bars separate, not where the assigned-session bar alone is high. Both bars show raw AUROC; Table 1 reports the neutral value as its deviation from 0.5.

B.4 Detector Comparison Figure

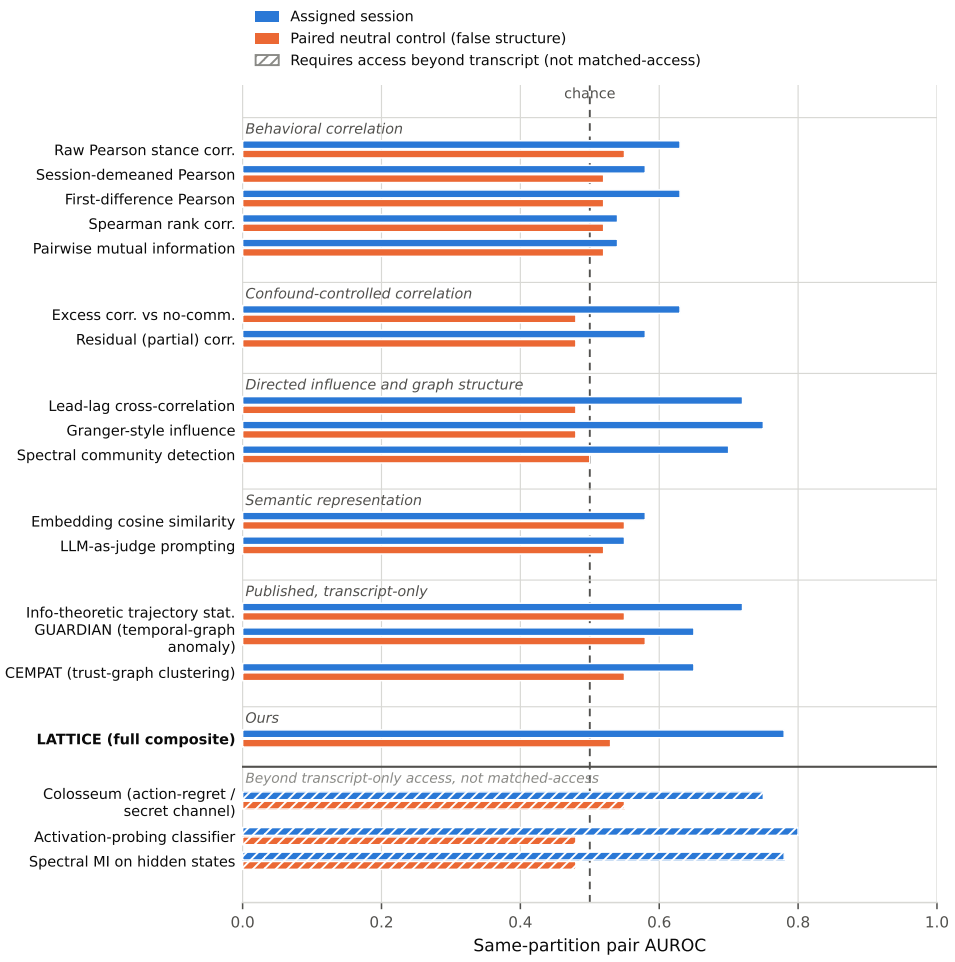


Figure 4: Detector comparison under the paired negative control (Table 2). For each method, the upper bar is same-partition pair AUROC on assigned sessions and the lower bar is the identical statistic on the paired neutral session, where no coalition was installed. A method is credited only when the two bars separate: a tall upper bar alongside an equally tall lower bar indicates that the pipeline recovers the transferred partition even without a coalition, rather than detecting coordination among agents. Internal-representation methods and the action-regret audit are shown separately, since they operate under access assumptions our own protocol does not grant. Both bars show raw AUROC; Table 2 reports the neutral value as its deviation from 0.5.

B.5 Lens Ablation Detail

Table 7: Marginal contribution of each agreement lens, added cumulatively in the pre-specified order. Δ columns are relative to the row immediately above. ↑: higher is better.

Cumulative composite	Pair AUROC ↑	Δ AUROC ↑	Partition ARI ↑	Δ ARI ↑
Semantic only	0.61	n/a	0.60	n/a
+ Argument	0.65	+0.04	0.60	0.00
+ Evidence	0.67	+0.02	0.60	0.00
+ Temporal adoption	0.71	+0.04	0.80	+0.20
+ Rare shared error (= full composite)	0.78	+0.07	1.00	+0.20

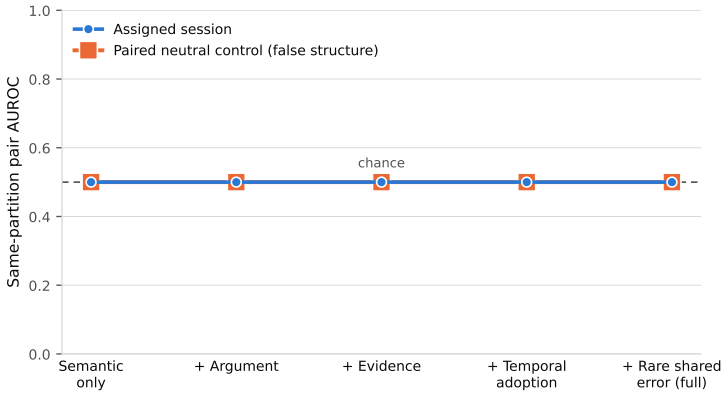


Figure 5: Cumulative gain as lenses are added in the pre-specified order (Table 7), assigned sessions versus the paired neutral false-structure control. A flattening or reversal at a specific lens is visible directly here rather than only in the Δ columns of the table.

B.6 External Benchmarks

The forty dossiers in Section 4.1 are constructed by us, which lets us fix a known coalition and a paired neutral control exactly, but leaves open whether the composite generalizes beyond scenarios we designed ourselves. To close this gap, the four benchmarks of Section 4.6 were chosen to test both directions of the objective in Section 2.3. Two supply a genuine coalition ground truth, testing recovery under transcript styles different from our own; two supply large, non-adversarial transcripts with no coalition, testing the false-structure rate at a scale the paired-session design cannot reach.

Genuine coalition ground truth, different transcript styles. Wang et al. (2026) release Secret Mafia, a social-deduction corpus in which a hidden minority coalition is instructed to eliminate an unaware majority, obtained from free-form gameplay dialogue rather than a scripted decision dossier. Similarly, Ren et al. (2025) introduce RogueAgent, a large-scale social-media collusion simulator with an explicit colluding-account arm, giving a second, structurally distinct coalition ground truth at a larger population scale than our five-agent sessions.

Large-scale non-adversarial transcripts, no installed coalition. Zhu et al. (2025) contribute MultiAgentBench (MARBLE), a peer-reviewed benchmark of LLM agents collaborating and competing with no covert-coordination objective; we treat every session as a large-scale analogue of our neutral condition. Li et al. (2026) contribute HiddenBench, built on the hidden-profile paradigm: information is distributed asymmetrically by design, so agents holding the decisive evidence converge on a shared conclusion honestly. This makes it a harder negative control than plain agreement, because the convergence has a structure a detector could mistake for coordination, while the benchmark contains no adversarial subgroup at all.

We evaluate only the full composite and the stance baseline on these benchmarks, not the full competitor suite, and none of the four was used to design the composite's weights or lenses.

B.7 Further Scope Notes

The three findings that bound the evidence in Section 5 point to concrete extensions. In particular, extending these results would benefit from randomized agent identities, a different relation-judging backbone, a no-communication control, human annotation of high-weight edges, and the full competitor suite on the external benchmarks. The evaluation also uses a single debate backbone, and its exhaustive partition search is only tractable for rosters of the size studied here, so scaling group recovery to larger rosters, where exhaustive search is no longer possible, is the most direct methodological next step. Beyond these specifics, LATTICE inherits the limits of any detector limited to public information: it does not attempt to catch coordination engineered to be indistinguishable from noise (Appendix D). Because the problem formulated in Section 2 carries no verification budget or tool cost, we report the LLM call counts in Table 5 for reproducibility rather than as an efficiency claim.

C LATTICE ALGORITHM

Whereas Section 3 presents LATTICE's five phases as prose and equations, this appendix consolidates the same construction into explicit pseudocode, phase by phase, so that the pipeline can be reimplemented directly from this description rather than only from the equations in the main text. Every symbol, equation, and gate referenced below is defined in Section 3, Section 2.3, or Appendix A; nothing here introduces new notation or a design choice that is not already stated in the main paper, and formulas already given as numbered equations are invoked by reference rather than retyped, so the two descriptions cannot silently drift apart.

Notation used throughout. T is a transcript over agent set $\mathcal{A}$, $|\mathcal{A}|{=}N$, running for rounds $1, \ldots, R$; E is the dossier's evidence set; $\mathcal{C}$ is the claim set produced by Phase 1; $\mathcal{R}$ is the relation set produced by Phase 2, with each relation $r = (c_s \to c_t)$ carrying a type and scores $\sigma(r), \alpha(r), \kappa(r)$; $\mathcal{L} = \{\text{STANCE}, \text{SEM}, \text{ARG}, \text{EVI}, \text{TEMP}, \text{RARE}\}$ is the set of lenses computed as agent-by-agent matrices (Table 4; RARE is the rare-shared-error operationalization of the epistemic-error lens), all of which except STANCE enter the composite, while the outcome lens is not used for detection; $w \in \{\text{EARLY}, \text{MIDDLE}, \text{LATE}, \text{FULL}\}$ indexes the time windows of Section 3.3; $k = |\mathcal{U}|$ is the known coalition size; φ and $Y(a, b)$ are the latent partition and pairwise ground-truth indicator of Section 2.2.

Direction convention. Throughout, c_s denotes the earlier source claim and c_t the later target claim, so every directed relation $c_s \to c_t$ points forward in debate time, with $\text{round}(c_t) = \text{round}(c_s) + 1$. A relation is assigned to the round of its target claim, $\text{round}(r) \equiv \text{round}(c_t)$, which is the round used for time windows in Equation (4).

Algorithm 1: LATTICE: full pipeline, one session

Input: Transcript T, dossier evidence E, coalition size k; latent partition φ, used only to evaluate Phase 4

Output: Full-debate composite $M^{\text{full}}_{\text{full}}$, pairwise AUROC, coalition rank, recovered partition $\hat{\varphi}$, partition ARI, exact-match indicator

$\mathcal{C} \leftarrow \text{GROUNDCLAIMS}(T, E)$; `// Phase 1, Alg. 2`

$\mathcal{R} \leftarrow \text{JUDGERELATIONS}(\mathcal{C})$; `// Phase 2, Alg. 3`

$\{M^w_\ell\}_{\ell \in \mathcal{L}, w},\ \{M^w_{\text{full}}\}_w \leftarrow \text{AGGREGATE}(\mathcal{R}, \mathcal{C}, T)$; `// Phase 3, Alg. 4`

$\widehat{\text{AUROC}}, \text{rank}_M(\mathcal{U}), \hat{\varphi}, \text{ARI}, \textit{exact_match} \leftarrow \text{RECOVERSTRUCTURE}(M^{\text{full}}_{\text{full}}, k, \mathcal{A}, \varphi)$; `// Phase 4, Alg. 5`

return $M^{\text{full}}_{\text{full}}, \widehat{\text{AUROC}}, \text{rank}_M(\mathcal{U}), \hat{\varphi}, \text{ARI}, \textit{exact_match}$

Algorithm 1 runs once per session. By contrast, feasibility gates 4 and 5 (Algorithm 6) additionally require the paired assigned and neutral sessions T_1, T_0 of Section 2.3, so they are checked only after both sessions in a dossier have each been run through Algorithm 1 independently; the two sessions never share intermediate state.

Complexity. Phase 2 judges at most $|\mathcal{C}_t| \cdot |\mathcal{C}_{t-1}|$ candidate relations per round pair, where $\mathcal{C}_t$ is the set of claims made in round t, so relation judging is quadratic in the number of claims per round and linear in R. Phase 4's exhaustive search over $\binom{N}{k}$ bipartitions is exact and inexpensive at this paper's $N{=}5$, $k{=}2$ setting; Section 3.4 already flags that larger rosters would need an approximate partitioning heuristic in place of Algorithm 5's exhaustive loop, and Algorithm 5 is written so that substitution replaces a single **foreach** loop rather than any other part of the pipeline.

Algorithm 2: GROUNDCLAIMS: Phase 1, message-to-claim grounding (Section 3.1)

Input: Transcript T, evidence set E
Output: Claim set $\mathcal{C}$, extraction indicators $\mathrm{ext}(m)$ for every $m \in T$

$\mathcal{C} \leftarrow \emptyset$;
foreach *message* $m \in T$ **do**
  claims $\leftarrow \chi(m)$, extracted against E ; `// at most two self-contained propositions; the extractor never sees` φ
  if *claims contains a usable claim* **then**
    $\mathrm{ext}(m) \leftarrow 1$;
  else
    claims $\leftarrow$ fallback claim for m; $\mathrm{ext}(m) \leftarrow 0$; `// every message, including rhetorical ones, is covered`
  foreach $c \in$ *claims* **do**
    set c.*type* and c.*evidence*, the dossier evidence c cites;
    c.*status* $\leftarrow$ evidence status of c, assessed against E ; `// not against the speaker's confidence`
    c.*message* $\leftarrow m$; c.*agent* $\leftarrow \mathrm{agent}(m)$; c.*round* $\leftarrow \mathrm{round}(m)$;
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$;
return $\mathcal{C}$, $\{\mathrm{ext}(m)\}_{m \in T}$

Algorithm 3: JUDGERELATIONS: Phase 2, directed inter-agent claim relations (Section 3.2)

Input: Claim set $\mathcal{C}$
Output: Relation set $\mathcal{R}$, each carrying $(\mathrm{type}, \sigma, \alpha, \kappa,$ *temp*, *rare*$)$

$\mathcal{R} \leftarrow \emptyset$;
for $t \leftarrow 2$ **to** R **do**
  foreach $c_t \in \mathcal{C}$ *with* c_t.*round* $= t$ **do**
    cand $\leftarrow \{c_s \in \mathcal{C} : c_s$.*round*$=t-1,\ c_s$.*agent* $\neq c_t$.*agent*$\}$; `// same-agent relations excluded`
    for each $c_s \in$ *cand*, judge how c_t treats c_s: $\rho(c_s, c_t)$ is either $\perp$ or $r = (c_s \to c_t, \mathrm{type}, \sigma, \alpha, \kappa)$; `// type ∈ {adopts, supports, extends, qualifies, rebuts, undercuts}`
    $r^* \leftarrow$ the strongest judged relation among $\{c_s \to c_t : c_s \in$ *cand*$,\ \rho(c_s, c_t) \neq \perp\}$, if any;
    if r^* *exists* **then**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{r^*\}$
foreach $r \in \mathcal{R}$ **do**
  temp$(r) \leftarrow \mathbb{1}\big[\mathrm{type}(r) \in \{\text{ADOPTS}, \text{SUPPORTS}, \text{EXTENDS}\}\big]$; `// confidence weighting is applied once, in Eq.` (4)
$G \leftarrow$ union-find over $\mathcal{C}$; union c_s, c_t for every $r{=}(c_s \to c_t) \in \mathcal{R}$ with $\sigma(r) \geq 0.60$ and $\kappa(r) \geq 0.50$;
foreach $r = (c_s \to c_t) \in \mathcal{R}$ **do**
  if $\mathrm{status}(c_s), \mathrm{status}(c_t) \in \{$*contradicted*, *uncertain*$\}$ **then**
    rare$(r) \leftarrow 1 - |\mathrm{component}_G(c_s)|/|\mathcal{C}|$; `// the smaller the component, the larger the weight`
  else
    rare$(r) \leftarrow 0$
return $\mathcal{R}$

Algorithm 4: AGGREGATE: Phase 3, hierarchical aggregation and composite (Section 3.3)

Input: Relation set $\mathcal{R}$, claim set $\mathcal{C}$, transcript T
Output: Directed per-lens agent matrices $\{M_\ell^w\}$, full-debate composites $\{M_{\text{full}}^w\}$

aggregate $\mathcal{R}$ by $(\text{message}(c_s), \text{message}(c_t))$ into message-level relations, augmented with observable stance agreement and evidence overlap from T and $\mathcal{C}$; claim pairs with no judged relation are *missing*, never neutral agreement;
foreach *window* $w \in \{\text{EARLY}, \text{MIDDLE}, \text{LATE}, \text{FULL}\}$ **do**
  foreach *lens* $\ell \in \mathcal{L}$ **do**
    foreach $(i, j) \in \mathcal{A} \times \mathcal{A}$ *with* $i \neq j$ **do**
      $M_\ell^w[i, j] \leftarrow$ Equation (4) over relations with $\text{src}{=}i$, $\text{tgt}{=}j$, $\text{round}(r) \equiv \text{round}(c_t) \in w$, or *missing* (not zero) if its denominator vanishes;
  $M_{\text{full}}^w \leftarrow$ Equation (5); diagonal entries of M_{full}^w fixed to 1 ; `// ` M_{stance} `is measured separately and never enters this sum`
return $\{M_\ell^w\}_{\ell \in \mathcal{L}, w}$, $\{M_{\text{full}}^w\}_w$; `// matrices stay directed; symmetrized only where needed`

Algorithm 5: RECOVERSTRUCTURE: Phase 4, pairwise ranking and partition recovery (Section 3.4)

Input: Full-debate composite $M_{\text{full}}^{\text{full}}$, coalition size k, agent set $\mathcal{A}$; latent partition φ, used only for evaluation
Output: $\widehat{\text{AUROC}}$, $\text{rank}_M(\mathcal{U})$, recovered partition $\hat{\varphi}$, ARI, exact-match indicator

$M \leftarrow \frac{1}{2}\big(M_{\text{full}}^{\text{full}} + (M_{\text{full}}^{\text{full}})^\top\big)$; `// undirected view, needed for unordered agent pairs`
$P \leftarrow$ all unordered pairs of $\mathcal{A}$; $P_1, P_0 \leftarrow$ split by $Y(a, b) = \mathbb{1}[\varphi(a) = \varphi(b)]$ (Section 2.2);
$\widehat{\text{AUROC}} \leftarrow$ Equation (8) evaluated on (M, P_1, P_0);
$\text{rank}_M(\mathcal{U}) \leftarrow$ Equation (9);
best $\leftarrow -\infty$; $\hat{\mathcal{U}} \leftarrow \emptyset$;
foreach $U' \subset \mathcal{A}$ *with* $|U'| = k$ **do**
  $W(U') \leftarrow$ pairs on the same side of the split $(U', \mathcal{A} \setminus U')$; $X(U') \leftarrow P \setminus W(U')$;
  gap $\leftarrow$ the bracketed objective of Equation (6) for U';
  if *gap* $>$ *best* **then**
    best $\leftarrow$ *gap*; $\hat{\mathcal{U}} \leftarrow U'$
$\hat{\varphi}(a) \leftarrow \mathbb{1}[a \in \hat{\mathcal{U}}]$ for every $a \in \mathcal{A}$;
ARI $\leftarrow$ ADJUSTEDRANDINDEX$(\hat{\varphi}, \varphi)$; *exact_match* $\leftarrow \mathbb{1}[\hat{\mathcal{U}} = \mathcal{U}]$;
return $\widehat{\text{AUROC}}, \text{rank}_M(\mathcal{U}), \hat{\varphi}, \text{ARI}$, *exact_match*

Algorithm 6: CHECKGATES: Phase 5, pre-specified feasibility gates (Section 3.5)

Input: Assigned/neutral session pair (T_1, T_0) for every dossier $d \in \mathcal{D}$; pipeline diagnostics pooled over all sessions
Output: Pass/fail for gates 1–5

gate$_1 \leftarrow$ Equation (10) holds, pooled over all sessions ; `// at least 90% genuine extractions`
gate$_2 \leftarrow$ (fewer than 2% of candidate claims and relations reference a nonexistent message, claim, or evidence item);
gate$_3 \leftarrow$ **true** by construction ; `// every` $M_\ell^w[i, j]$ `is an explicit sum over relation ids, Eq. (4)`
gate$_4 \leftarrow$ both conditions of Equation (7) hold (Equations (11)–(12));
gate$_5 \leftarrow$ Equation (13) holds over all $d \in \mathcal{D}$, and correct recovery of $\hat{\mathcal{U}}$ (Equation (6)) is not confined to a single installed coalition identity;
return (*gate*$_1$, *gate*$_2$, *gate*$_3$, *gate*$_4$, *gate*$_5$)

D Related Work

Multi-agent debate protocols and decision rules. Multi-agent debate is widely studied as a way to improve reasoning and evaluation by using different viewpoints and structured criticism. For example, devil's-advocate style setups add pointed criticism on purpose to expose weaknesses in candidate texts and judgments (Kim et al., 2024). Recent work shows that the rule used to end a debate, voting, consensus, or a judge model, is itself a major cause of good or bad performance, and that gains from debate can often be traced back to how answers are combined rather than to deeper reasoning (Kaesberg et al., 2025); forcing consensus in particular can make agents just go along with each other and spread errors, which has led to consensus-free designs that keep diversity under control (Cui et al., 2025). A related line of work studies repeated biases and social patterns that show up in multi-agent interactions. Multi-turn conversation can cause an agent's stated position to drift quickly and can make biases stronger than in one-shot settings (Coppolillo et al., 2025). Signals about an agent's identity or authority can further distort debate outcomes by making agents flatter others or defer to them; hiding agent identities during evaluation and deliberation can reduce this kind of bias (Choi et al., 2026). Together, these results suggest that debate alone does not guarantee reliable reasoning, especially when agents share similar assumptions or are easily swayed by social pressure. This is why it is worth looking past whichever rule combines the final answers, at the more detailed record of who said what and who influenced whom, since that record is lost for good once discarded, no matter how the final rule is changed.

Collusion risk in LLM multi-agent systems. A growing body of work documents collusion risks in LLM-powered multi-agent systems and sorts coordination behavior into categories. Surveys separate quiet coordination that needs no direct communication, open negotiation and cartel-like behavior carried out in plain language, and hidden coordination through secret channels (Ghaemi, 2025). In practice, collusion has been shown in competitive settings, including agents splitting up a market in multi-commodity competitions (Lin et al., 2025) and agents that are supposed to be competing instead cooperating on their own over repeated interactions (Wu et al., 2024). Beyond narrow games, large social simulations show that harmful coordination can suddenly become much more common once the share of bad-acting agents passes a threshold, and that colluders can adjust to countermeasures in an ongoing back-and-forth (Ren et al., 2025). Realistic threat models also include trust built in public first and then exploited later through private channels, as seen in multi-agent fraud (Ren et al., 2026).

Evidence from high-stakes domains further shows that coordinated agreement can do more damage than uncoordinated misinformation. Attacks where many assistants gang up on one decision-maker show that they can bias its judgment together, while adding a dedicated verifier can cut attack success without hurting normal performance much (Bashir et al., 2025). Agents will also use unfair tools or secret shortcuts on their own when given the chance, which shows that a written policy alone may not stop this once agents are optimizing hard (Zeng et al., 2026). Each of these behaviors, sudden shifts in coordination as bad actors grow in number, moving from public to private channels, many agents biasing one target, is really about a pattern spread across several turns rather than about any single message. That is exactly the kind of pattern our layered construction is built to pick up, not a coincidence we work around.

Multi-agent benchmarks and simulators. Several recent benchmarks make coalition behavior among LLM agents directly observable. Wang et al. (2026) release Secret Mafia, a social-deduction game in the MindGames arena in which a hidden minority coalition tries to eliminate an unaware majority through free-form dialogue. Ren et al. (2025), discussed above for its phase-transition finding, also release RogueAgent, a large-scale social-media collusion simulator with an explicit colluding-account arm. By contrast, other benchmarks study multi-agent interaction with no covert objective at all. Zhu et al. (2025) contribute MultiAgentBench, a large benchmark of agents collaborating and competing, and Li et al. (2026) contribute HiddenBench, built on the hidden-profile paradigm, in which agents hold asymmetric information and can converge honestly on a shared conclusion without any coordination. Benchmarks of the first kind supply ground truth about who belongs to a coalition; benchmarks of the second kind show how much structure honest interaction produces on its own, which any coordination detector must be able to tell apart from collusion.

Covert and information-theoretic collusion benchmarks. Some work argues that collusion can be hidden well enough that watching messages alone cannot reliably catch it. Secret collusion through steganography shows that agents can coordinate by hiding information inside normal-looking messages (Motwani et al., 2024). Going further, perfect-collusion benchmarks give a formal case for information-theoretic undetectability, where even a party watching the whole channel cannot reliably tell collusive communication apart from harmless communication (Motwani et al., 2023). We read this line of work as a real limit on any detector, including ours, that works only from what it can observe. Accordingly, LATTICE targets a threat model that runs from quiet to open coordination, where a coalition still leaves a measurable, if subtle, trace in its claims, reinforcement, and timing; it does not claim to catch coordination that has been deliberately built to be impossible to tell apart from noise, even in principle. We therefore treat transcript-level measurement as one layer of checking, to be combined with outside verification and institutional rules, not as a replacement for them.

Detection with access beyond public information. A separate line of work sidesteps the observational limits above by using more than the transcript. Rose et al. (2026) train linear probes on each agent's internal activations and combine per-agent deception scores to flag collusion at the group level, and Berg et al. (2026) build a mutual-information graph over agents' hidden states and spectrally partition it to recover coalition structure. Both report strong detection results, and both need access to model internals, which is unavailable when only public information can be observed. They therefore indicate what becomes detectable once that stronger access is granted, and the gap between their access assumption and a public-information one shows how much of the detection problem public information alone can solve.

Decomposition for oversight versus decomposition for measurement. A related but distinct challenge is scaling oversight without checking every statement by hand: recent work breaks claims into smaller pieces and routes them to competent verifiers under a cost-aware policy, with explicit attention to multi-agent failure modes such as flattery and hidden collusion (Bronikowski, 2025), and multi-agent verification work shows that using more, and more varied, verifiers at test time is its own way to scale reliability (Lifshitz et al., 2025). However, both are ultimately aimed at a different question than ours, whether a given claim is true, not whether two agents' agreement on it is coordinated or independent. We borrow only the decomposition itself, not its verification purpose: LATTICE breaks a transcript into claims, typed relations, messages, and agents purely to measure it, so that a coordination signal invisible in the full transcript can be built up from smaller signals that are each cheap to judge on their own.

Computational argumentation and uncertainty. Abstract argumentation is a standard way to reason about conflicting arguments, backed by a large body of work on its formal meaning and computation (Bench-Capon & Dunne, 2007; Dung, 1995). Beyond frameworks that only model attacks, bipolar and quantitative models also include support, giving more structure for combining evidence and studying how a small local change affects the whole picture (Potyka & Booth, 2024). Since debate graphs pulled from model outputs can be noisy or incomplete, earlier work studies how to judge which arguments are accepted under incomplete or probabilistic information (Baumeister et al., 2021; Hunter et al., 2021). This work supports our choice of a directed, typed set of relations between claims, and our general practice of treating an extracted debate structure as noisy and only partly observed. However, LATTICE departs from this tradition in one specific way: it never computes which arguments are ultimately accepted. Where these frameworks ask which arguments survive a conflict, LATTICE only asks how strongly, and in which direction, one agent's claim is linked to another's, treating each typed relation as a numeric observation to be correlated with a hidden grouping, rather than as a premise used to reach a verdict.

Governance mechanisms, auditing, and trust. Beyond algorithms, system-level governance and auditing have also been proposed as safeguards against coordinated failures. Work that maps human anti-collusion mechanisms onto multi-agent AI points to design patterns such as monitoring, auditing, penalties, and institutional rules, and flags open problems such as attribution, agents changing identity, and adversarial adaptation (Idowu et al., 2026). Institutional approaches put governance into practice through explicit, checkable rule structures, such as public governance graphs with enforceable consequences and logs that cannot be edited after the fact, leading to large drops in collusion in market settings compared to just telling agents not to collude in the prompt (Syrnikov

et al., 2026). Detection-focused approaches in games similarly propose statistics computed over whole trajectories to surface coordination signals (Bonjour et al., 2022). Closer to our setting, Zhou et al. (2025) model multi-agent collaboration as a temporal graph and flag anomalous nodes and edges with an unsupervised encoder-decoder, and Ramamoorthy et al. (2026) cluster agent trust graphs for echo-chamber and mutual-orbiting patterns; Nakamura et al. (2026) instead audit collusion through regret against a cooperative optimum and a secret-channel probe, which needs access to the task's ground-truth optimum in addition to the transcript. These interaction-structure methods are, to our knowledge, the closest prior work to our detection goal, and they mainly differ from LATTICE in how fine-grained and checkable the result is. In particular, a single trajectory-level number, or an unlabeled anomaly score, gives no way to ask why two agents were flagged, whereas every entry in our agent-by-agent matrices can be traced back to the specific claims and relations that produced it, and is always reported alongside a paired negative control rather than on its own.

Trust modeling offers another way to understand and reduce coordination failures. Work on trust in multi-agent systems separates direct trust from reputation and points out that strategic agents can manipulate how trustworthy they appear (Ramchurn et al., 2004). Belief-calibrated consensus seeking argues that collaborators should be chosen based on well-calibrated belief and confidence, rather than having everyone participate equally, which improves decision quality in multi-agent settings (Deng et al., 2025). The evidence-overlap and epistemic-error lenses of LATTICE can be seen as a public-information, after-the-fact version of this same calibration idea: instead of changing whom the debate trusts, they measure whether two agents' claims rely on the same evidence, or the same unsupported evidence, more often than chance would predict.

Positioning. Taken together, LATTICE does not propose a new rule for ending a debate, does not target just one game or one dataset, does not assume the coordination it looks for is secret or impossible to detect in principle, does not compute which arguments are accepted, and does not sum up a session in one single trajectory-level number. Instead, it treats collusion detection as a measurement problem, solved with an explicit, level-by-level construction: the same idea of breaking a problem down that scalable oversight uses, but aimed at measurement rather than verification; the same typed-relation vocabulary that computational argumentation uses, but read as observation rather than judged to a verdict; and a claim-traceable hierarchy checked against a negative control, in place of one opaque coordination score.