WHICH GRAPH IS SAFEST? PREDICTING ADVERSARIAL CAPTURE OF LLM AGENT GROUPS FROM LOCAL KERNELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-agent LLM deliberation, a single adversarial agent can steer the group to a wrong answer, and how often this happens depends on the communication graph. We show that the deliberation dynamics are well predicted by a few local quantities: the question under deliberation, an agent's own answer, its number of neighbours, and how many of them hold the adversary's answer, with the adversary counted separately from neighbours it has converted. On this state we approximate the deliberation by a Markov chain whose transitions come from a per-question switching kernel. We study three open-weight 8B models, 600 MMLU-Pro questions and three memory protocols, with one adversary among six agents. Calibrated on eight graph–placement configurations, the kernel predicts adversarial capture, the probability that the adversary's answer becomes the group's majority, on 52 other configurations with a mean error of 0.012, and it extends to groups of eight and twelve agents, two adversaries and ten rounds. Under a new question set or new adversary messages, question-specific susceptibility changes, so absolute capture needs four to five calibration runs there, but the ranking of graphs transfers much better. Ranking all 112 connected six-agent graphs, concentrating links on one agent is worst for every model, and which graph is best depends mainly on how the group starts: counterfactually raising how often the correct answer starts as a minority moves the most accurate graph from dense to sparse for models whose agents follow the majority strongly. On a second question set, LogiQA, reading these curves at its solo-answer distribution reproduces the ranking obtained after calibrating on LogiQA with rank correlation 0.95–0.99. Our code is available on `https://anonymous.4open.science/r/which-graph-is-safest-38C3/README.md`.

1 INTRODUCTION

Background. Multi-agent LLM systems let several models exchange messages before they answer, and debate between agents is used to improve reasoning, factuality and evaluation (Du et al., 2024; Liang et al., 2024; Chan et al., 2024). The same exchange lets one agent pull the others to a wrong answer. An agent that argues confidently for a wrong option, whether adversarial or simply mistaken, can spread the wrong answer through the group (Amayuelas et al., 2024; Huang et al., 2025; Ju et al., 2026), and LLM agents conform to majorities much as people do (Weng et al., 2025; Zhu et al., 2025). How far a wrong answer spreads depends on who talks to whom, so the communication graph is a safety decision.

Motivation. Evaluating adversarial spread in communication graphs is expensive. Each candidate needs full multi-round deliberations on every question, and six agents already allow 112 connected graphs with 407 distinct adversary positions; existing studies compare a handful of architectures (Huang et al., 2025; Yu et al., 2025). We show that the group's behaviour can be predicted well from the behaviour of single agents and that an individual agent's decision is well predicted by a few local features, so the group deliberation dynamics is well approximated by a Markov chain whose transitions are local single-agent decisions, and the graph determines which local situations occur.

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.

Methodology. We place one adversary among six agents and study three open-weight 8B models, 600 MMLU-Pro questions and three memory protocols (Section 2). We identify which local features predict an agent's decisions (Section 3), compose them into a Markov approximation of the group (Section 4), test it under distribution shift (Section 5), and use it to rank all 112 connected six-agent graphs (Section 6).

Contributions. In short, our contributions are four-fold:

- We show that the group dynamics of LLM deliberation are well predicted by the question together with a few local quantities: an agent's own answer, its degree (the number of agents whose messages it receives) and how many of these neighbours hold the adversary's answer. The other features we test add little, the round matters only when agents see the full history, and message position changes single decisions, so we randomise the order.
- On this state we build a Markov approximation of the deliberation. Calibrated on eight graph–placement configurations, it predicts capture on 52 other six-agent configurations with a mean error of 0.012, and it extends to larger groups, a second adversary and longer deliberations.
- We separate what transfers across tasks from what does not. Under a new question set and new adversary messages, including an adaptive adversary, question-specific susceptibility changes, so absolute capture needs recalibration, four to five runs once the population levels are reused, whereas the ranking of graphs is far more stable.
- We rank all 112 connected six-agent graphs for three models by worst-case and average-case capture and by final accuracy. Concentrating links on one agent is worst for every model, and which graph is best depends mainly on how the group starts: reweighting the questions counterfactually moves the most accurate graph from dense to sparse as the correct answer more often starts as a minority, for models whose agents follow the majority strongly. We use this to predict the graph ranking on LogiQA from the MMLU-Pro calibration and LogiQA's solo answers alone, without any deliberation runs on LogiQA; the prediction matches the ranking obtained after calibrating on LogiQA with rank correlation 0.95–0.99.

2 SETUP

Protocol. Six agents on a connected graph answer a multiple-choice question. An agent's neighbours are the agents it is linked to, whose messages it receives, and its degree is the number of its neighbours. One agent is the adversary: in every round it repeats a fixed, confident argument for a designated wrong option. In round 0 each clean agent answers alone. In rounds 1 to 5 each clean agent receives the question, its own previous answer, and its neighbours' latest messages in a random order. The three memory protocols differ in what else the agent sees:

- **none**: only the letter it answered in the previous round;
- **own**: also its own previous message in full;
- **full**: every earlier round of its own and its neighbours' messages.

The protocol changes the dynamics: under full history agents become harder to move as the rounds pass (Section 3).

Models and questions. We use three open-weight instruction-tuned models (Qwen3-8B, Ministral-8B-Instruct-2410 and Llama-3.1-8B-Instruct) and 600 MMLU-Pro questions from 14 domains. Clean agents sample at temperature 0.7 and top-p 0.95, with at most 96 new tokens and Qwen3's thinking mode off. The adversary's argument for each question is generated once (Qwen3-4B, greedy, cut to 400 characters) and reused for every model, so the three models face the same attacks on the same questions. A *run* is one graph–placement configuration at one seed: all 600 questions deliberated for five rounds, which takes 2 (no memory) to 5 (full history) minutes on one H100 GPU.

Outcomes. A run is *eligible* if at least one clean agent answers correctly in round 0, so the correct answer is present in the group. The group is *captured* if, after round 5, a strict majority of the six

Table 1: Per model and memory protocol. Solo: share of clean agents that answer correctly in round 0. No attack: probability that the group's plurality answer after round 5 is correct (ties split), with six clean agents, over all questions and the calibration graphs. Under attack: the same with the adversary. Attack cost: no-attack minus under-attack accuracy. Capture rate: as defined in the text. In parentheses: standard error in units of the last digit, $0.632(16) = 0.632 \pm 0.016$; all tables use this notation.

		accuracy				
model	protocol	solo	no attack	under attack	attack cost	capture rate
Qwen3-8B	none	0.711(18)	0.745(17)	0.632(16)	+0.113(10)	0.167(12)
	own	0.711(18)	0.758(17)	0.627(16)	+0.131(11)	0.183(12)
	full	0.711(18)	0.754(17)	0.648(16)	+0.106(9)	0.149(10)
Ministral-8B	none	0.404(15)	0.476(17)	0.378(16)	+0.097(9)	0.275(13)
	own	0.404(15)	0.486(17)	0.353(15)	+0.133(10)	0.335(14)
	full	0.404(15)	0.480(18)	0.359(16)	+0.120(10)	0.308(14)
Llama-3.1-8B	none	0.336(15)	0.378(15)	0.289(13)	+0.089(7)	0.167(12)
	own	0.336(15)	0.431(16)	0.298(13)	+0.132(8)	0.237(12)
	full	0.336(15)	0.426(18)	0.374(16)	+0.052(7)	0.126(11)

agents, adversary included, hold the adversary's letter; a three–three tie counts one half. The *capture rate* is averaged over eligible runs, first within each question and then over questions. It is the quantity the kernel predicts (Appendix A.4). A small share of eligible runs is already captured at round 0 (0.3–2.7% on MMLU-Pro, up to 8% on LogiQA), and excluding them leaves the prediction error unchanged (Appendix C.2). Table 1 also reports the group's accuracy, defined in its caption.

Effect of deliberation and of the attack. Without an adversary, deliberation raises accuracy above solo accuracy by 0.035–0.096. The attack then costs 0.05–0.13 of accuracy, most with own memory and least for Llama-3.1-8B with full history, which also gains most from deliberation.

3 Which features predict the group dynamics

A small local state predicts the group dynamics well: an agent's own answer, its degree, and how many of its neighbours hold the adversary's letter, with the adversary counted separately from converted neighbours (Section 4; Figure 1 shows how adoption depends on the last two). An agent's prompt contains more than this, so for each further feature we measure its effect on a single decision, as a Mantel–Haenszel risk ratio between two arms matched on the question and the rest of the local state (Appendix A.8), and its effect on the group, as the change in prediction error when it is added to or removed from the kernel. All contrasts, the co-switching test and the ablation are in Appendix B.

The round matters only under full history. Without full history an agent switches in rounds 3–5 at about the same rate as in round 2. Under full history it becomes harder to move as the transcript grows, switching in later rounds at as little as 0.36 times its round-2 rate (Table 2), so for this protocol the kernel adds a round term (Section 4).

Message order changes single decisions, so we randomise it. An agent with one holding neighbour among at least three adopts the letter more often when that neighbour's message comes last in the prompt, by up to 0.03 per decision for Qwen3-8B and Llama-3.1-8B; Ministral-8B responds to the first slot instead (Table 2). For Qwen3-8B the last slot multiplies the rate by 7.5, with an interior-slot rate of 0.004. We shuffle the order independently for each agent and round, so each cell of the kernel is the average over slots and adding position to the kernel cannot improve the prediction (Appendix A.6). A fixed order would change the group's outcome: we estimate that it moves a graph's worst-case capture by up to 0.065, although the ranking of graphs changes little (Appendix A.6).

Converted peers persuade less than the adversary, which matters for the group in some settings. An agent whose single holding neighbour is a converted peer adopts the letter less often than one

whose single holding neighbour is the adversary, in every model and protocol except Qwen3-8B under full history (Table 2). Separating the two in the kernel matters where a single converted peer decides the outcome: on the triangle with three leaves g_8 it cuts Qwen3-8B's prediction error by a factor of six. It also helps under full history, but averaged over the configurations outside the calibration the gain is within its standard error (Table 4). We keep the split.

Given their local states, agents decide independently. The group transition (Section 4) assumes that clean agents decide independently given their local states. Two agents that share a neighbour receive the same message, so if the message carried a common push they would switch together more often than their states imply. They do not: their excess co-switching is no larger than for two agents in the same states taken from different runs, which cannot influence each other (Table 3). This test is pairwise; that the factorisation holds for the whole group is tested by prediction (Section 4).

The binary state is enough for the group, although single decisions depend on the agent's other answer. The state records whether an agent holds the adversary's letter, not whether a non-holding agent is correct or holds another wrong answer. Pooled over questions, correct agents adopt the letter at about 0.6 times the rate of agents holding another wrong answer, but most of this gap lies between questions: on questions the agents know, they both start correct and resist the adversary. Within a question the gap is small, though significant in some settings (Appendix B.1). A kernel that also tracks correct and other wrong answers does not lower the prediction error, because the question level already captures the differences between questions (Appendix B.1).

4 A Markov approximation and its prediction

State and kernel. Write $x_i^t \in \{0,1\}$ for whether agent i holds the adversary's letter after round t. A clean agent i enters round t in the local state

$$s_i^t = (o, d, a, c), \qquad o = x_i^{t-1}, \quad d = |N(i)|, \quad a = \mathbf{1}[v^* \in N(i)], \quad c = \textstyle\sum_{j \in N(i)\setminus\{v^*\}} x_j^{t-1} : \tag{1}$$

its own answer, its degree, whether the adversary v^* is a neighbour, and how many converted neighbours hold the letter; $b = a + c$ is the number of holding neighbours. For question q the kernel is the probability that the agent's holding status changes, $\kappa_q(s) = \Pr(x_i^t \neq x_i^{t-1} \mid s_i^t = s, q)$, covering adoption ($o = 0$) and abandonment ($o = 1$).

Estimation. Many cells are sparse, so every cell is estimated by hierarchical Beta–Binomial shrinkage over three levels: pooled over questions, split by whether a holding neighbour is the adversary or a converted peer, and the question's own, each shrunk toward the level above with a prior sample size of $\lambda = 2$ fixed in advance. Every clean agent in every round of every eligible calibration run is one observation, and under full history each question's cells are shifted by a pooled round profile (equations, the round term and the sensitivity to λ in Appendices A.1 and A.2). Both kinds of level are needed: without the question level the prediction error nearly triples, and without shrinkage toward the population cells it more than triples (Appendix B). A logistic response in the share, number and degree of holding neighbours, with question offsets, predicts as well as the kernel, while simpler forms predict worse (Appendix C.1). Figure 1 shows the pooled cells: the models differ most with one or two holders among two to five neighbours, where Qwen3-8B adopts readily and the other two do not.

Propagation. Clean agents decide independently given their local states (Section 3) and the adversary never changes, so the group is approximated as a Markov chain on the 2^5 configurations of the five clean agents with a factorised transition matrix,

$$T_t(x, y) = \prod_{i \in C} p_i(x)^{y_i} \big(1 - p_i(x)\big)^{1-y_i}, \qquad p_i(x) = \begin{cases} \kappa_q(s_i) & x_i = 0 \\ 1 - \kappa_q(s_i) & x_i = 1, \end{cases} \tag{2}$$

and the probability of capture after R rounds is

$$F_q = \sum_x \big(\mu_0 T_1 \cdots T_R\big)(x)\, \varphi(x), \qquad \varphi(x) = \mathbf{1}[h(x) > 3] + \tfrac{1}{2}\,\mathbf{1}[h(x) = 3], \tag{3}$$

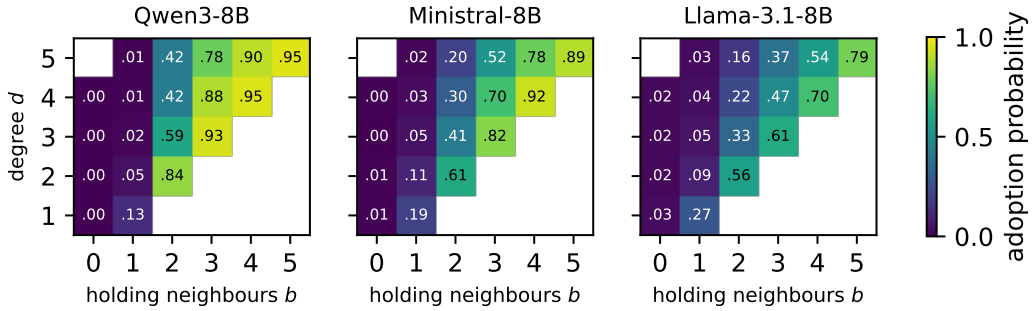

Figure 1: The kernel under own-memory, pooled over questions: probability that an agent that does not hold the adversary's letter adopts it, by degree d and number of holding neighbours b. Empty cells are unreachable ($b > d$) or have fewer than five observations.

with $h(x)$ the number of holders including the adversary (Appendices A.3 and A.4). For large groups the probability can be estimated using sampling instead of direct computation (Appendix A.9). A graph's score averages F_q over questions; its worst-case score is the maximum over adversary positions and its average-case score the mean.

The kernel generalises to graphs covered by the calibration. A graph is *covered* by the calibration when the local states its agents pass through during the deliberation appear in the calibration data, each with at least five observations (Appendix A.7). Which states a graph passes through follows from the kernel itself, so coverage can be checked before the graph is run. Prediction error falls sharply with coverage, to a median of 0.011 when all visits are covered (Appendix C). Eight calibration configurations (the star with the adversary at the hub and on a leaf, the ring, the prism, the octahedron, the complete graph, g_{19} and g_6) cover every placement of every six-agent graph and every larger graph whose agents have at most five neighbours. Outside this calibration the kernel accurately predicts capture rates on new graphs, larger groups and longer deliberations as long as coverage is good (Figure 2).

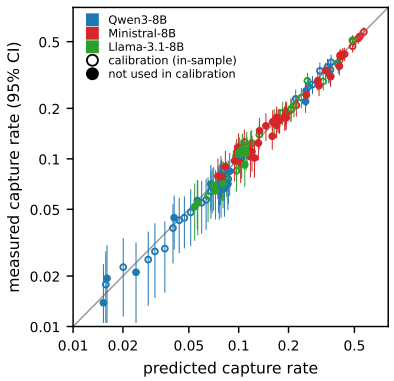

Figure 2: Predicted against measured capture rate for every measured configuration (log axes). Hollow: calibration; filled: not used in calibration.

New graphs. On 52 six-agent configurations (15 graphs never used in calibration and new placements on g_6) the mean error is 0.012, only slightly above the error in calibration (Table 7). A regression of capture on the adversary's degree and the number of edges, fitted to the same configurations, has twice the kernel's error; it predicts capture directly from graph features instead of composing local responses, so the comparison measures the value of the local-to-global decomposition (Table 9).

More agents, more adversaries, more rounds. We ran these tests without memory, at two seeds of the 600 questions, and predicted them with the kernel calibrated on six agents and five rounds (Table 8). With eight agents and one adversary, on the ring and the 4-regular circulant $C_8(1, 2)$, every prediction lies inside the measured 95% interval. With twelve agents and two adversaries on opposite sides of $C_{12}(1, 2)$, the kernel predicts the share of agents holding the letter for all three models and capture for Qwen3-8B, and under-predicts capture by about 0.02 for the other two. Run for ten rounds, it predicts Qwen3-8B and Llama-3.1-8B but over-predicts Ministral-8B by 0.05, whose groups nearly stop changing after round 5.

5 THE KERNEL UNDER DISTRIBUTION SHIFT

The population-level dynamics transfer across distribution shifts; question-level susceptibility does not. Section 4 fixes the questions and the attack. Under a new question set and new adversary messages (Qwen3-8B and Ministral-8B, own memory; Appendix D), the kernel's population levels carry over and its question level does not.

New questions. On 600 LogiQA questions capture is two to three times as high as on MMLU-Pro, yet the pooled cells of the two question sets correlate at 0.98–0.99. The MMLU-Pro kernel therefore under-predicts every LogiQA test graph, by 0.10 on average, while a hybrid that keeps its population levels and takes only the question cells from LogiQA runs is as accurate as LogiQA's own kernel (Table 12). The hybrid needs few runs: four to five LogiQA calibration configurations bring its median error to 0.03 (Figure 5), and one seed per configuration does as well as four.

New adversary messages. On the MMLU-Pro questions we reran the calibration and three test graphs under a second fixed argument, a bare assertion of the option, and an adaptive adversary that writes a new argument every round. Capture changes with the attack, almost doubling for Qwen3-8B from the bare assertion to the adaptive adversary. The population cells again agree, and a hybrid with the paper's population levels and the attack's question cells is off by at most 0.014 (Table 14).

What transfers. A question cell records how easily a question is flipped by a given message: per-question capture under another attack correlates only moderately with that under the paper's argument. Predicting absolute capture on a new task or attack therefore needs calibration runs there, about five once the population levels are reused (Appendix D.1). Choosing a graph needs less information: Section 6 shows that the ranking of graphs depends mainly on how the group starts, which can be measured from solo answers.

6 APPLICATION: CHOOSING A TOPOLOGY

We score all 112 connected six-agent graphs at each of their 407 distinct adversary placements. A graph's worst-case score is its capture rate at the worst placement, and its average-case score is the mean over the six seats, which applies when the adversary's position is random. The first three findings rank the graphs on the MMLU-Pro questions as they were run, by worst-case capture, average-case capture and final accuracy; the last reweights those questions to show how the best graph moves when groups start differently (Figure 3). Full rankings with predicted and measured values are in Appendix E, and the graphs are drawn in Figure 6.

For worst-case adversary placement, concentrating links on one agent is worst for every model, and the safest graph depends on the model. The star is the riskiest graph in eight of the nine settings, and the near-star g_6 and the tree g_{19} are among the five riskiest in all nine. At the safe end several dense graphs are almost equally safe. The kernel's own error on the difference between two graphs is about 0.014, so graphs closer than about 0.03 are not resolved (Appendix E.1). The complete graph is most likely the safest for Qwen3-8B, the ring for Llama-3.1-8B, and $K_{3,3}$ or the ring for Ministral-8B (Table 18). The predicted safest graph is within 0.015 of its measurement in every setting.

For random adversary placement, the graph matters less and Ministral-8B's ranking inverts. Averaging over seats shrinks the gap between the safest and the riskiest graph to at most 0.08. For Ministral-8B the ranking nearly inverts (Figure 7): hub-and-leaf graphs become the safest, because most of their seats are leaves from which the adversary reaches few agents, and the tree g_3 is its safest graph in all three protocols. Hub-and-leaf graphs are thus the worst choice when the adversary picks its seat and, for Ministral-8B, among the best when it does not.

Ranked by accuracy under attack on MMLU-Pro, dense graphs are best for every model. Capture counts only whether the adversary's letter wins. Extending the state to three answers per agent (correct, another wrong answer, the adversary's letter), propagated exactly on 3^5 configurations, predicts the group's final accuracy on the 52 configurations outside the calibration with a mean error of 0.015 (Appendix E.4). Ranked by accuracy, the complete graph is the most accurate graph, or within 0.015 of it, in every setting, while the ring, Llama-3.1-8B's safest graph by capture, falls to

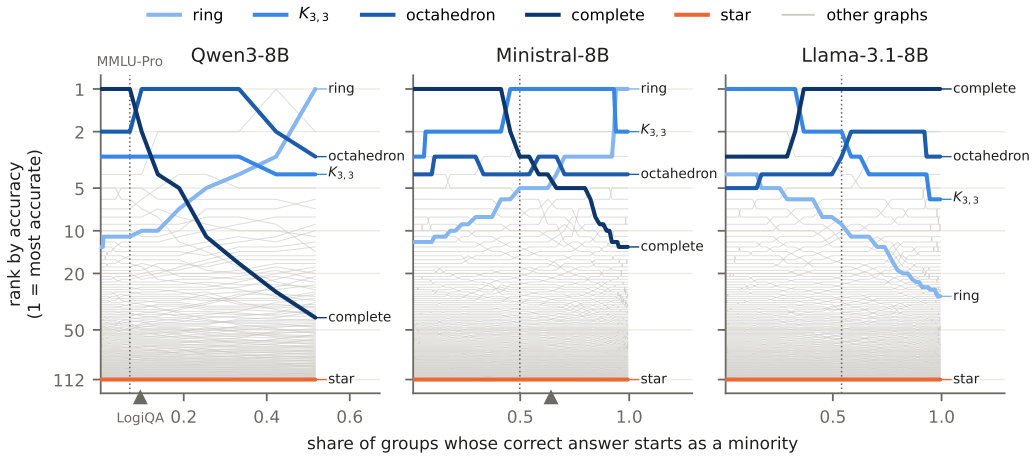

Figure 3: How the most accurate graph moves with the way groups start (own memory, worst-case placement). Each line is a graph's rank by final accuracy among all 112 (log scale, 1 = most accurate) as the question set is reweighted toward questions whose correct answer more often starts as a minority, held by one or two of the five clean agents (Appendix E.3). Dotted line: MMLU-Pro as run; triangle: LogiQA. Graphs near the top are often within 0.015 of each other, so swaps among the first few ranks are not resolved. The same analysis for capture is in Figure 8.

rank 9–24 and the star is the least accurate for every model. A dense graph spreads the adversary's letter but also the correct answer; on MMLU-Pro, where the correct answer usually starts as the majority whenever an agent has it, the second matters more. The next finding shows where this stops holding.

The best graph depends mainly on how the group starts. Absolute capture depends on how susceptible individual questions are, but the ranking of graphs depends more on how the group starts. Re-ranking the graphs with one part of a model's kernel removed or taken from another model, the largest change comes from the round-0 answers: giving Qwen3-8B the round-0 answers of either other model moves the complete graph from safest to as low as rank 35 (Appendix E.2). Figure 3 follows the most accurate graph as the start changes. We emulate question sets that start differently by reweighting each model's questions: a question's weight grows exponentially with one property of its round-0 answers, shown on the x-axis. Each question keeps its own start distribution and its own kernel cells. At each weighting we rank all 112 graphs with the three-state chain under worst-case placement and plot each graph's rank on the y-axis, highlighting the regular graphs from sparse to dense (ring, $K_{3,3}$, octahedron, complete) and the star (Appendix E.3). Because the start is a property of solo answers, the x-coordinate of a new task can be measured before any deliberation is run.

The property that matters most for accuracy is how often the correct answer starts as a minority. When it often does, the complete graph falls to about rank 40 for Qwen3-8B and the ring becomes the most accurate graph, because in a dense graph a wrong majority overwrites a correct minority before it can persuade anyone, the kind of confident wrong cascade that Han et al. (2026) find becomes more likely as connectivity grows. The ring overtakes the complete graph at a minority share of about 0.2 for Qwen3-8B and 0.6 for Ministral-8B (Table 20), earlier for Qwen3-8B because its lone correct agents persuade about twice as often. Llama-3.1-8B moves the other way, and the complete graph stays the most accurate. Its agents are noisy: an agent whose five neighbours hold the correct answer, as it does itself, leaves it in 6.6% of rounds against at most 0.2% for the other two models, and a dense graph corrects such slips because the drifting agent sees five correct neighbours rather than two (Table 19).

For capture the corresponding property is the share of agents that start on the adversary's letter: as it grows, Ministral-8B's complete graph falls to about rank 20 and the ring becomes the safest (Figure 8). LogiQA, whose four options produce more minority starts, falls where the curves predict: on its calibration runs the ring is both more accurate and safer than the complete graph for Ministral-8B,

and tied with it for Qwen3-8B (Appendix E.3). Across question sets and attacks the ranking of graphs keeps a rank correlation of at least 0.91 with MMLU-Pro's. Reading Figure 3 and its capture version at LogiQA's start, measured from its round-0 answers alone, gives rankings that match LogiQA's own with rank correlations of 0.95 to 0.99 for both models. Where LogiQA's start barely moves, as for Qwen3-8B, the read-off returns the MMLU-Pro ranking, which is already close; where it moves, as for Ministral-8B's capture, the read-off follows it, moving the complete graph from fourth to tenth safest against 27th in LogiQA's own ranking. Other predictors match LogiQA on one model but not on both (Appendix E.3). A new task therefore needs calibration runs to predict how much capture it suffers, but much of the choice of topology can be anticipated from how its groups start.

7 Related work

Multi-agent debate and its evaluation. Multi-agent debate lets several LLM instances exchange answers over rounds before committing (Du et al., 2024; Liang et al., 2024; Chan et al., 2024; Wang et al., 2025a). Its benefit is contested: majority voting explains most of the gain and debate alone induces a martingale on beliefs (Choi et al., 2025); under matched compute it often fails to beat chain-of-thought or self-consistency (Smit et al., 2024; Zhang et al., 2025); and it can lower accuracy even when stronger models outnumber weaker ones (Wynn et al., 2025). Failure taxonomies for multi-agent systems are descriptive (Cemri et al., 2025). We do not ask whether debate helps but how far a wrong answer spreads once introduced, and we report deliberation's own effect separately from the attack's (Table 1).

Faulty and adversarial agents. Huang et al. (2025) inject faulty agents into several architectures and find hierarchies more resilient than chains; Amayuelas et al. (2024) show that an adversary's persuasiveness governs its reach and that prompt-level warnings help little; Ju et al. (2026) study manipulated knowledge spreading through agent communities, and Cui & Du (2025) exploit conformity with compromised agents. Yu et al. (2025) study how topology affects safety: more highly connected networks are more susceptible to adversarial spread, with a 29.7% drop in task performance on a star topology, which our hub finding confirms, and static metrics such as the average distance from the attacker track dynamic evaluations better than standard graph metrics; this is a structural predictor of safety, like the regressions we compare against (Table 9). Recent defences use graph-based detection and topology-aware intervention (Wang et al., 2025b; Miao et al., 2026). Most of these works compare a handful of architectures empirically. Our contribution is the step that makes such comparisons unnecessary: a kernel measured on eight configurations that predicts every six-agent graph and every placement, with a coverage check that says when it cannot. Nilayam et al. (2026) report that an honest heterogeneous peer protects against an adversarial one; in our terms this is a change in the local state rather than in the model mix.

Conformity and what agents attend to. LLM agents conform to majorities (Weng et al., 2025; Zhu et al., 2025; Cho et al., 2025; Song et al., 2025), most of the effect survives without a speaker (Hu & Qu, 2026), and flips decompose into instability, conformity and persuasion (Hao et al., 2026). Topology shapes these dynamics (Han et al., 2026; Mehdizadeh & Hilbert, 2025). Identity labels, position in a list and memory of the transcript are all known to bias single decisions (Choi et al., 2026; Zheng et al., 2023; Wang et al., 2024; Shi et al., 2025; Liu et al., 2024; Tian et al., 2026); Section 3 measures the same features and asks which of them predict the group dynamics.

Predictive models of collective dynamics. The composition of local update rules into group behaviour is the subject of opinion dynamics (DeGroot, 1974; Hegselmann & Krause, 2002; Castellano et al., 2009), threshold and cascade models (Granovetter, 1978; Watts, 2002; Kempe et al., 2003), and resilient consensus, where an r-robust graph tolerates a bounded number of faulty neighbours (LeBlanc et al., 2013; Zheng et al., 2026; Lee et al., 2026). Our kernel is a contagion model of this kind whose response function is measured rather than assumed. LLM agent populations have been simulated as opinion dynamics (Chuang et al., 2024; El et al., 2026).

8 DISCUSSION AND CONCLUSION

Cost and limitations. Once the kernel is measured on a task and attack, choosing a topology is a computation. Calibrating a model takes eight graph–placement configurations per question set and attack, 96 runs over the three memory protocols, which took under five GPU-hours per model on one H100; scoring all 112 six-agent graphs under both placements, and covered extrapolations to other group sizes, then needs no further inference and takes a few CPU-minutes. The kernel is calibrated on six agents, one adversary, five rounds, multiple-choice questions and three 8B models, and tested beyond that on larger groups, two adversaries, ten rounds, a second question set and other adversary messages. Its largest errors are for Ministral-8B, on hub seats and over ten rounds (Appendix C). Absolute capture on a new task or under a new attack needs calibration runs there, although about five suffice once the population levels are reused. Capture uses a strict-majority rule, and the accuracy predictions lump all other wrong answers into one state (Appendix E.4).

Future work. The kernel turns questions about a group into measurements of single agents, which suggests three directions. First, defences can be measured the same way: a defence prompt, or an agent in a judge or moderator role, changes the local response, and measuring how it shifts the adoption and leaving cells would predict its effect on every graph without running them; a group with mixed roles then needs one kernel per role. Second, measuring the kernel across model sizes would show whether larger models follow majorities more or less strongly and drift less, and so on which side of the dense-versus-sparse transition they fall. Third, calibration could be cheaper: instead of running full deliberations, one can construct each local state directly, presenting an agent with chosen neighbour messages, and measure its response. This would also reach counterfactual states that no calibration graph visits, such as degrees above five.

Conclusion. The spread of an adversary's answer through a group of LLM agents is well predicted by a Markov chain on a small local state: a single-agent switching probability given an agent's own answer, its degree and its holding neighbours, per question. Measured on eight graph–placement configurations, this kernel predicts graphs outside the calibration, larger groups and a second adversary, and ranks all 112 six-agent graphs without running them. Which graph to choose depends on the objective, the model and on how the group starts: hubs are worst for every model, sparse graphs best contain the adversary's answer for the weaker models, and for two of the three models the most accurate graph moves from dense to sparse as the correct answer more often starts in the minority. The ranking of graphs thus combines two ingredients. The model contributes its population-level response, how strongly its agents follow a majority and how often they drift, which transfers across question sets and attacks; the task contributes how its groups start, which solo answers reveal. Neither decides alone: at the same start, Llama-3.1-8B prefers the opposite graph to Qwen3-8B and Ministral-8B, and for the same model the best graph changes with the start. Only the absolute level of capture needs the question-level cells, which have to be measured on each new task. A first choice of graph for a new question set or adversary message can therefore be made from the model's calibrated kernel and the task's solo answers alone, while predicting how much capture it suffers needs about five calibration runs.

REPRODUCIBILITY STATEMENT

Every experiment is a script that writes a JSON record. The kernel, its estimator and the exact propagator are one module, and every generated table in the paper is written by one script from the released runs. The code is available on github: `https://anonymous.4open.science/r/which-graph-is-safest-38C3/README.md`; raw deliberation traces for all models, protocols and graphs will be released with checksums.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, to clean and reformat datasets, to provide feedback on the research methodology and experiments, to propose and refine hypotheses and analyses, to help develop the theoretical models, and to interpret results in preliminary analyses. We have not used generative AI tools to assist with translation, and providing critical ingredients

for proofs and qualitative or thematic data analysis are not applicable to this work, which contains no theorems and no qualitative data. Additionally, we used generative AI tools to draft and edit the text, tables and figures, to create and edit code, and to summarise related work. We have reviewed all AI-assisted work: the authors set the research questions and the study design, and checked the analyses, results, text and code; every number in the paper is produced by released scripts from the recorded runs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6929–6948, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.407. URL https://aclanthology.org/2024.findings-emnlp.407/.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, May 2009. doi: 10.1103/RevModPhys.81.591. URL https://link.aps.org/doi/10.1103/RevModPhys.81.591.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=fAjbYBmonr.

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. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:271745584.

Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. Herd behavior: Investigating peer influence in llm-based multi-agent systems. *arXiv preprint arXiv:2505.21588*, 2025.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iUjGNJzrF1.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14284–14311, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.650. URL https://aclanthology.org/2026.acl-long.650/.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of LLM-based agents. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3326–3346, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.211. URL https://aclanthology.org/2024.findings-naacl.211/.

Yu Cui and Hongyang Du. Mad-spear: A conformity-driven prompt injection attack on multi-agent debate systems. *arXiv preprint arXiv:2507.13038*, 2025.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2285509.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: Statistical mechanics predicts collective behavior of ai agents. *arXiv preprint arXiv:2608.16578*, 2026.

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6): 1420–1443, 1978.

Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. Conformity dynamics in llm multi-agent systems: The roles of topology and self-social weighting. *arXiv preprint arXiv:2601.05606*, 2026.

Xiqi Hao, Zengqing Wu, Yu-Xuan Qiu, Chuan Xiao, Ruiqi Xu, Shuyuan Zheng, and Jianbin Qin. Not all flips are conformity: Decomposing stance convergence in multi-agent llm debate. *arXiv preprint arXiv:2606.00820*, 2026.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Yibo Hu and Jiaming Qu. Most llm conformity needs no speaker: Measuring the speaker-free floor in peer-pressure benchmarks. *arXiv preprint arXiv:2607.05545*, 2026.

Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=bkiM54QftZ.

Tianjie Ju, Yiting Wang, Yi Hua, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, et al. Flooding spread of manipulated knowledge in llm-based multi-agent communities. *Science China Information Sciences*, 69(7):172103, 2026.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 137–146, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137370. doi: 10.1145/956750.956769. URL https://doi.org/10.1145/956750.956769.

Heath J. LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4):766–781, 2013. doi: 10.1109/JSAC.2013.130413.

Haejoon Lee, Vincent-Daniel Yun, Hyeonho Oh, Dimitra Panagou, and Sai Praneeth Karimireddy. Robust multi-agent LLMs under byzantine faults. In *Second Workshop on Agents in the Wild: Safety, Security, and Beyond*, 2026. URL https://openreview.net/forum?id=MgHy7f7VxT.

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. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Aliakbar Mehdizadeh and Martin Hilbert. When your ai agent succumbs to peer-pressure: Studying opinion-change dynamics of llms. *arXiv preprint arXiv:2510.19107*, 2025.

Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan, and Xin Wang. Blindguard: Safeguarding llm-based multi-agent systems under unknown attacks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39215–39234, 2026.

Prashanti Nilayam, Kiran Kumar Ramanna, Prashil Tumbade, and Sankalp Nayak. Heterogeneous llm debate under adversarial peers: Honest gains, replacement costs, and resilience. *arXiv preprint arXiv:2606.19826*, 2026.

Ronald C. Read and Robin J. Wilson. *An Atlas of Graphs*. Oxford University Press, 1998.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Maojia Song, Tej Deep Pala, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. Llms can't handle peer pressure: Crumbling under multi-agent social interactions, 2025. URL https://arxiv.org/abs/2508.18321.

Hongduan Tian, Xiao Feng, Ziyuan ZHAO, XiangyuZhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EdTt8nMAMA.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5434be94e82c54327bb9dcaf7fca52b6-Paper-Conference.pdf.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. 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. 7261–7276, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.359. URL https://aclanthology.org/2025.acl-long.359/.

Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002. doi: 10.1073/pnas.082090499. URL https://www.pnas.org/doi/abs/10.1073/pnas.082090499.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *ICLR*, 2025.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Kun Wang, Qingsong Wen, and Yang Wang. NetSafe: Exploring the topological safety of multi-agent system. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2905–2938, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.150. URL https://aclanthology.org/2025.findings-acl.150/.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Lifan Zheng, Jiawei Chen, Qinghong Yin, Jingyuan Zhang, Xinyi Zeng, and Yu Tian. Rethinking the reliability of multi-agent systems: a perspective from byzantine fault tolerance. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i41.40806. URL https://doi.org/10.1609/aaai.v40i41.40806.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. 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. 3854–3872, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.195. URL https://aclanthology.org/2025.acl-long.195/.

A Derivations

Notation. A deliberation runs on a connected graph $G = (V, E)$ with $|V| = n$ agents, one of which, v^*, is the adversary; $C = V \setminus \{v^*\}$ are the clean agents, $N(i)$ the neighbours of i and $d_i = |N(i)|$. Write $x_i^t \in \{0, 1\}$ for whether agent i holds the adversary's letter after round t, with $x_{v^*}^t \equiv 1$. The *local state* of clean agent i entering round $t \geq 1$ is

$$s_i^t = \big(o,\ d,\ a,\ c\big) = \Big(x_i^{t-1},\ d_i,\ \mathbf{1}[v^* \in N(i)],\ \textstyle\sum_{j \in N(i) \cap C} x_j^{t-1}\Big), \qquad b = a + c, \tag{4}$$

where b is the number of neighbours holding the letter. For question q the kernel is the switching probability $\kappa_q(s) = \Pr\big(x_i^t \neq x_i^{t-1} \mid s_i^t = s,\ q\big)$, and $p_i = \Pr(x_i^t = 1 \mid \cdot)$ equals $\kappa_q(s_i^t)$ if $o = 0$ and $1 - \kappa_q(s_i^t)$ if $o = 1$.

A.1 Estimator as a chain of posterior means

For a cell with n observations of which k switched, a $\mathrm{Beta}(\lambda\pi,\ \lambda(1-\pi))$ prior with mean π and prior sample size λ gives the posterior mean

$$\hat{\kappa} = \frac{k + \lambda\pi}{n + \lambda} = \frac{n}{n + \lambda}\,\frac{k}{n} + \frac{\lambda}{n + \lambda}\,\pi, \tag{5}$$

a weighted average of the cell's own rate and the prior mean, with the data's weight growing as $n/(n+\lambda)$. Every clean agent in every round of every eligible calibration run is one observation: n_{odb} counts the observations, pooled over questions, in which a clean agent entered a round in state (o, d, b) and k_{odb} how many of them changed their holding status; n_{odac}, k_{odac} do the same for the state split by source and $n_{q,odac}, k_{q,odac}$ for question q alone. The kernel applies the posterior mean three times, each level supplying the prior mean of the next ($\lambda = 2$ throughout):

$$\pi^{(0)}(o,d,b) = \begin{cases} 0 & o = 0,\ b = 0 \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{6}$$

$$\pi^{(1)}(o,d,b) = \frac{k_{odb} + \lambda\, \pi^{(0)}(o,d,b)}{n_{odb} + \lambda} \qquad \text{pooled over questions} \tag{7}$$

$$\pi^{(2)}(o,d,a,c) = \frac{k_{odac} + \lambda\, \pi^{(1)}(o,d,a+c)}{n_{odac} + \lambda} \qquad \text{pooled, split by source} \tag{8}$$

$$\kappa_q(o,d,a,c) = \frac{k_{q,odac} + \lambda\, \pi^{(2)}(o,d,a,c)}{n_{q,odac} + \lambda} \qquad \text{question } q \tag{9}$$

Two properties follow directly. A cell without data ($n = 0$) returns its prior mean unchanged, so an unseen question–state combination inherits the pooled estimate; and as $n \to \infty$ the estimate converges to the cell's own rate. The first property is also why the order of the levels matters: a level can only pass on information from the levels above it.

The prior sample size λ was fixed at 2 before any evaluation: each level counts as two observations for the level below, a weak prior that a cell's own data outweigh after a few observations. Refitted with other values, the mean error on the 52 six-agent configurations outside the calibration is 0.0111 ($\lambda = 0.5$), 0.0112 (1), 0.0115 (2), 0.0128 (5) and 0.0156 (10): a stronger prior pulls each question's cells toward the population cells and loses the question information that the ablation shows is needed (`scripts/lambda_sensitivity.py`).

A.2 Round term under full history

Let $\pi_r^{(1)}$ be the pooled (o, d, b) cell estimated from round r alone, shrunk toward $\pi^{(1)}$. Its difference on the log-odds scale, $\delta_r(o,d,b) = \operatorname{logit} \pi_r^{(1)}(o,d,b) - \operatorname{logit} \pi^{(1)}(o,d,b)$, is a population-level round profile. It is added to the question's all-rounds estimate to form the prior of the question's own round cell:

$$\rho_{q,r} = \sigma\big(\operatorname{logit} \kappa_q(o,d,a,c) + \delta_r(o,d,a+c)\big), \qquad \kappa_{q,r} = \frac{k_{q,r,odb} + \lambda\, \rho_{q,r}}{n_{q,r,odb} + \lambda}. \tag{10}$$

With little round-specific data for a question, $\kappa_{q,r} \approx \rho_{q,r}$: the question's own switching propensity shifted by the shared round profile. The alternative ordering, with $\kappa_{q,r}$ shrunk toward $\pi_r^{(1)}$, loses the question's all-rounds information whenever its round cells are thin, because $\pi_r^{(1)}$ contains no question-specific term. The order matters here too: shrinking the question's round cells toward a pooled round estimate instead discards the question's all-rounds data and raises out-of-sample error for Ministral-8B from 0.016 to 0.034.

A.3 Factorisation of the group transition

Assume that, given the round-$(t-1)$ configuration x, the clean agents' decisions in round t are conditionally independent and each depends on x only through its local state (Section 3). Then for configurations $x, y \in \{0,1\}^C$

$$T_t(x,y) = \Pr\big(x^t = y \mid x^{t-1} = x\big) = \prod_{i \in C} p_i(x)^{\,y_i} \left(1 - p_i(x)\right)^{1-y_i}, \tag{11}$$

a $2^{|C|} \times 2^{|C|}$ stochastic matrix, time-homogeneous unless the round term is on. The distribution after t rounds is $\mu_t = \mu_0\, T_1 T_2 \cdots T_t$, computed exactly.

A.4 GROUP FAILURE

With the adversary counted as a holder, the number of holders in configuration x is $h(x) = 1 + \sum_{i \in C} x_i$. Capture is a strict majority, ties counted one half:

$$\varphi(x) = \mathbf{1}\big[h(x) > n/2\big] + \tfrac{1}{2}\,\mathbf{1}\big[h(x) = n/2\big], \qquad F_q(G, v^*) = \sum_x \mu_R(x)\,\varphi(x), \tag{12}$$

with R the number of rounds. The graph score averages over questions, $\bar{F}(G, v^*) = |Q|^{-1} \sum_q F_q$; the worst-case score is $F_{\mathrm{wc}}(G) = \max_{v^*} \bar{F}(G, v^*)$ and the average-case score is $F_{\mathrm{avg}}(G) = n^{-1} \sum_{v^* \in V} \bar{F}(G, v^*)$. Positions in the same automorphism orbit give the same value, so each orbit is evaluated once and weighted by its size.

A.5 INITIAL DISTRIBUTION

For a measured run, μ_0 is a point mass on the run's round-0 configuration. For a graph that was never run, round-0 answers are given alone, so the clean agents are exchangeable and only the number of holders matters. With $w_q(m)$ the empirical probability that m of the clean agents hold the letter in round 0 for question q,

$$\mu_0(x) = w_q\big(|x|\big) \Big/ \tbinom{|C|}{|x|}. \tag{13}$$

For larger groups, where no empirical w_q exists, each clean agent holds the letter independently with the question's measured round-0 rate.

A.6 WHY A RANDOMISED POSITION CANNOT IMPROVE GROUP PREDICTION

Let $f \in \{0, 1, \dots\}$ index which prompt slot a holder occupies. If the order is randomised independently of everything else, then within a cell $\Pr(f \mid s) = \pi(f \mid s)$ is known, and

$$\kappa_q(s) = \sum_f \pi(f \mid s)\,\kappa_q(s, f). \tag{14}$$

The chain only ever uses $\kappa_q(s)$, and propagating the refined kernel with f drawn from $\pi(f \mid s)$ gives exactly the same one-step transition probabilities, so every group-level prediction is unchanged. Splitting a cell by position can only add estimation error.

Under a fixed order, $\pi(f \mid s)$ is no longer the random-order distribution and the group's outcome changes. We estimate how much without new runs. For the first and for the last slot, we refine each kernel cell by whether that slot holds the letter, adding to the question's logit the pooled log-odds difference between the two cases (estimated with the same shrinkage as the kernel), and set the probability w that the slot holds the letter by a fixed ordering policy instead of b/d: the adversary's message always in the slot, never in it, all holding messages placed there first, or holding messages never there unless all neighbours hold. Scored over all 112 graphs with exchangeable starts, the largest change in worst-case capture from the random order is 0.065 for Qwen3-8B and 0.054 for Llama-3.1-8B with own memory, where the mean change over graphs is 0.019–0.044 depending on the policy; in the other settings the mean change is at most 0.017 and the largest at most 0.032. Rank correlations with the random-order ranking are 0.92–1.00 under worst-case placement and 0.49–1.00 with a random seat, the lowest for Qwen3-8B when the adversary's message is always last.

A.7 COVERAGE

With n_{odb} the number of calibration observations in cell (o, d, b), the predicted coverage of a target (G, v^*) over R rounds is the occupancy-weighted share of clean-agent visits that land in cells with at least five observations:

$$\mathrm{cov}(G, v^*) = \frac{1}{R} \sum_{t=0}^{R-1} \sum_x \mu_t(x)\, \frac{1}{|C|} \sum_{i \in C} \mathbf{1}\big[n_{s_i(x)} \geq 5\big], \tag{15}$$

computed before any run on G; the same expression over (o, d, a, c) cells gives the split coverage.

A.8 MANTEL–HAENSZEL RISK DIFFERENCE

A contrast compares two arms, for example the single holder's message in the last slot (arm 1) against an interior slot (arm 0), among observations that agree on every matched feature. Observations are grouped into strata k, each a combination of question and the matched covariates; stratum k has n_{1k}, n_{0k} observations and x_{1k}, x_{0k} adoptions in the two arms. The Mantel–Haenszel risk difference is the weighted average of the within-stratum differences,

$$\widehat{\mathrm{RD}} = \frac{\sum_k w_k \left(x_{1k}/n_{1k} - x_{0k}/n_{0k}\right)}{\sum_k w_k}, \qquad w_k = \frac{n_{1k}\, n_{0k}}{n_{1k} + n_{0k}}, \tag{16}$$

over strata in which both arms are observed. The weight is the reciprocal of the sum of the two arms' inverse sizes, so a stratum counts in proportion to how precisely it measures a difference; a stratum with a handful of observations in either arm contributes little. Because comparisons never cross strata, differences between the arms in question difficulty, degree or round cannot appear as an effect. The same weights give the arm rates reported alongside the contrasts, $\bar{p}_j = \sum_k w_k\, x_{jk}/n_{jk} \big/ \sum_k w_k$, whose difference is $\widehat{\mathrm{RD}}$.

Strata are defined per contrast. Round index: question, own answer, degree, adversary among the neighbours, number of converted holders, first-listed message holding the letter, and graph; arms rounds 3–5 against round 2. Message position: question, degree, and round 1 versus later, among agents that do not hold the letter and have exactly one holder among at least three neighbours; arms first or last slot against an interior slot. Converted peer versus adversary: question, degree, first-listed message holding the letter, and round 1 versus later, among agents that do not hold the letter and have exactly one holder; arms the holder is a converted peer against the holder is the adversary.

Uncertainty. Observations within a question are dependent, because they share the adversary's text, the agents, five rounds, several graphs and several seeds. We therefore resample questions rather than observations: in each of 2000 replicates, questions are drawn with replacement, each drawn question contributes its strata with multiplicity equal to its draw count, and $\widehat{\mathrm{RD}}$ is recomputed. The 95% interval is the 2.5th to 97.5th percentile of the replicates.

A.9 SAMPLING INSTEAD OF EXACT PROPAGATION

The exact chain stores a $2^{|C|} \times 2^{|C|}$ transition matrix per question and round, which limits it to about twelve agents. Sampling replaces it with T trajectories per run: in each round every clean agent looks up its local state, switches with probability $\kappa_q(s_i)$, and capture is estimated as the share of trajectories that end in a strict majority (ties counted one half). The cost is $O(T\,|C|)$ per round and works for any group size and any number of adversaries, provided no clean agent neighbours two of them. We compared the two with the final kernel on all 125 measured six-agent configurations, every run started from its own round-0 state. The mean difference in graph-level capture between sampling and the exact chain is 0.0049 with one trajectory per run, 0.0014 with 10, 0.0005 with 100 (largest 0.002) and 0.0002 with 1000, against a mean error of 0.010 between the exact prediction and the measurement; sampling therefore adds no error that matters from 10 trajectories on. For twelve agents with two adversaries (1024 states), 100 trajectories per run agree with the exact chain to within 0.0007 in all three models and take 0.1–0.2 seconds per configuration, against 73–85 seconds for the exact chain.

B WHICH FEATURES MATTER

Table 2 gives the decision-level contrasts of Section 3, Table 3 the co-switching test, and Table 4 the change in prediction error when a part of the kernel is removed or added. Clean agents also switch to the adversary's letter without seeing it, in 0.1–0.2% of agent-rounds for Qwen3-8B, 1.7–2.2% for Ministral-8B and 3.5–6.0% for Llama-3.1-8B, which sets a floor under every adoption rate.

Table 4 gives the mean graph-level error on every configuration outside the calibration when one part of the kernel is removed or added, each variant refitted on the eight calibration configurations. Six-agent configurations are predicted with the exact chain from each run's own round-0 answers; the eight-agent, twelve-agent and ten-round runs (no memory) by sampling 400 trajectories per run

Table 2: Effect of each candidate feature on a single decision, per model and protocol. RD: Mantel–Haenszel risk difference in the probability of adopting the adversary's letter (for the round index, of switching in either direction), with its bootstrap standard error in parentheses (notation of Table 1); *: 95% question-bootstrap interval excludes zero. RR: Mantel–Haenszel risk ratio. Message position uses agents with one holder among at least three neighbours. Reference-arm adoption rates: interior slot 0.003–0.032; single holder is the adversary 0.05–0.25.

	Qwen3-8B		Ministral-8B		Llama-3.1-8B	
contrast	RD	RR	RD	RR	RD	RR
round index (3–5 vs 2)						
none	$+0.000(0)$	$\times 1.01$	$+0.006(2)^*$	$\times 1.07$	$+0.005(2)^*$	$\times 1.03$
own	$-0.003(1)^*$	$\times 0.90$	$+0.002(2)$	$\times 1.02$	$+0.003(2)$	$\times 1.02$
full	$-0.021(2)^*$	$\times 0.36$	$-0.035(2)^*$	$\times 0.63$	$-0.014(2)^*$	$\times 0.87$
first vs interior slot						
none	$+0.002(1)^*$	$\times 1.72$	$+0.004(2)^*$	$\times 1.19$	$+0.017(3)^*$	$\times 1.55$
own	$-0.000(0)$	$\times 0.89$	$+0.010(2)^*$	$\times 1.41$	$-0.001(2)$	$\times 0.97$
full	$+0.000(0)$	$\times 1.12$	$+0.009(1)^*$	$\times 1.42$	$+0.002(2)$	$\times 1.10$
last vs interior slot						
none	$+0.009(2)^*$	$\times 3.93$	$-0.001(1)$	$\times 0.95$	$+0.026(2)^*$	$\times 1.83$
own	$+0.026(3)^*$	$\times 7.51$	$+0.005(2)^*$	$\times 1.19$	$+0.029(3)^*$	$\times 1.92$
full	$+0.013(2)^*$	$\times 4.79$	$+0.001(1)$	$\times 1.07$	$+0.006(2)^*$	$\times 1.28$
converted peer vs adversary						
none	$-0.107(22)^*$	$\times 0.57$	$-0.033(7)^*$	$\times 0.79$	$-0.041(8)^*$	$\times 0.79$
own	$-0.131(20)^*$	$\times 0.42$	$-0.048(7)^*$	$\times 0.72$	$-0.063(8)^*$	$\times 0.62$
full	$-0.005(11)$	$\times 0.90$	$-0.062(8)^*$	$\times 0.59$	$-0.030(8)^*$	$\times 0.76$

Table 3: Excess probability that two clean agents sharing a neighbour both switch in the same round, relative to the product of their switching rates in the same (question, round, own answer, degree, holding neighbours) states pooled over runs; the control pairs the first agent with a same-state agent from a different run. Calibration runs.

model	protocol	excess	control	pairs
Qwen3-8B	none	$+0.0004$	-0.0003	602665
	own	$+0.0003$	-0.0007	601653
	full	$+0.0003$	-0.0006	602458
Ministral-8B	none	-0.0026	-0.0012	505743
	own	-0.0026	-0.0017	505526
	full	-0.0021	-0.0012	506288
Llama-3.1-8B	none	-0.0024	-0.0021	437159
	own	-0.0031	-0.0024	436990
	full	-0.0023	-0.0012	438161

(Appendix A.9). Sparse: the path, g_3 and g_8. Position is added as a slot flag averaged over its known distribution under the random order, and is only defined for six agents.

B.1 Is the binary state enough?

The kernel's state records whether an agent holds the adversary's letter, not which other answer it holds or what its non-holding neighbours say. Table 5 measures both at the level of single decisions, as in Table 2: Mantel–Haenszel risk ratios within the kernel's cells (d, a, c) and the round, with 95% question-bootstrap intervals. Agents holding the correct answer adopt the letter less often than agents holding another wrong answer, and agents whose non-holding neighbours agree on one letter adopt it less often and leave it more often. The own-answer contrast is stratified by cell and round but not by question. With the question added to the strata the ratio is 0.81–0.97 for Qwen3-8B and Ministral-8B and 0.93–1.06 for Llama-3.1-8B, so most of the pooled gap lies between questions, which the kernel's question level already separates. The remaining within-question gap is significant

Table 4: Mean graph-level error on configurations outside the calibration, by kernel variant (number of configurations in the header). In parentheses: mean paired difference from the full kernel and its bootstrap standard error over configurations, in units of 10^{-4}; positive means the variant is worse. The question level and population shrinkage are needed everywhere; the adversary/converted split and the round term are kept for interpretability, and the split's effect exceeds twice its standard error only under full history.

	all (52)	sparse (22)	full history (13)	8 agents (6)	12 agents (3)	10 rounds (3)
full kernel	0.0115	0.0078	0.0104	0.0056	0.0171	0.0260
no adversary / converted split	0.0120 (+5±7)	0.0091 (+13±14)	0.0125 (+21±9)	0.0067 (+11±22)	0.0165 (-6±22)	0.0298 (+39±34)
no question level	0.0306 (+191±27)	0.0336 (+258±24)	0.0385 (+281±33)	0.0396 (+340±46)	0.0437 (+267±129)	0.0664 (+404±133)
no population shrinkage	0.0408 (+292±45)	0.0191 (+113±27)	0.0158 (+54±18)	0.0726 (+670±218)	0.1081 (+910±144)	0.1352 (+1092±262)
no round term (full history)	–	–	0.0125 (+21±21)	–	–	–
+ first-slot position	0.0115 (+0±0)	0.0078 (+0±0)	0.0105 (+1±1)	–	–	–
+ last-slot position	0.0115 (+0±2)	0.0078 (+0±1)	0.0106 (+2±2)	–	–	–

for Ministral-8B in all three protocols (0.91–0.93), for Qwen3-8B without memory (0.81) and for Llama-3.1-8B with memory (0.93), and not significant for Qwen3-8B with memory or Llama-3.1-8B without it.

Table 5: Decision-level effects of the information the binary state leaves out: risk ratio [95% interval]. Own correct: adoption by an agent holding the correct answer over one holding another wrong answer. Agree: adoption (leaving) when all non-holding neighbours, at least two, give the same letter over when they give several letters.

model	protocol	own correct, adopt	agree, adopt	agree, leave
Qwen3-8B	none	0.58 [0.48, 0.73]	0.49 [0.28, 0.89]	1.19 [0.98, 1.66]
	own	0.55 [0.42, 0.67]	0.72 [0.52, 1.20]	1.24 [0.99, 1.64]
	full	0.56 [0.43, 0.75]	0.57 [0.39, 0.97]	0.94 [0.67, 1.39]
Ministral-8B	none	0.61 [0.56, 0.66]	0.40 [0.37, 0.44]	1.13 [1.05, 1.25]
	own	0.65 [0.60, 0.71]	0.50 [0.45, 0.55]	1.23 [1.12, 1.35]
	full	0.64 [0.60, 0.70]	0.45 [0.41, 0.49]	1.14 [1.05, 1.25]
Llama-3.1-8B	none	0.96 [0.86, 1.06]	0.76 [0.70, 0.83]	1.04 [1.02, 1.07]
	own	0.84 [0.76, 0.92]	0.85 [0.77, 0.95]	1.13 [1.09, 1.19]
	full	0.83 [0.75, 0.91]	0.69 [0.62, 0.77]	1.09 [1.04, 1.14]

To test whether these differences matter for the group we built a three-state kernel. Each clean agent is correct, holds another wrong answer, or holds the adversary's letter, and its neighbourhood counts the adversary, converted holders, and correct and other-wrong clean neighbours. An agent not holding the letter adopts it with probability $\sigma(\operatorname{logit}\kappa_q(0,d,a,c)+\delta_0)$ and otherwise moves to the correct answer with probability τ_q; a holder leaves with probability $\sigma(\operatorname{logit}\kappa_q(1,d,a,c)+\delta_1)$ and then goes to the correct answer with probability τ_q. Here κ_q is the paper's kernel, δ_0 and δ_1 are offsets fitted against each observation's own κ_q (by own answer, by the number of correct non-holding neighbours, or both; normal prior with standard deviation 1), and τ_q is estimated by the same Beta–Binomial chain as the kernel, pooled over (s, n_C, n_W) and then per question. With both offsets zero the chain reproduces the binary chain exactly (to 10^{-15}). Table 6 gives the errors. The extra state does not help on the configurations outside the calibration or on Ministral-8B's largest errors; it removes part of Ministral-8B's over-prediction at round 10. The binary kernel's per-question residual is also unrelated to the share of agents holding another wrong answer in round 0 (correlation -0.001 over 22,705 question–configuration pairs).

C Prediction and generalisation

For each model and protocol we refit the kernel on each of the 254 proper, non-empty subsets of the eight calibration configurations. Each subset kernel predicts every measured configuration it was not fitted on: the calibration configurations outside the subset and all configurations run outside the calibration, each from its runs' own round-0 answers. This uses only existing runs and gives 22,606 predictions, each with the target's predicted coverage under the subset kernel. Figure 4 shows

Table 6: Mean graph-level error by state (number of configurations in parentheses). Binary: the paper's kernel. Own: adoption offset by own answer (correct or other wrong). Neighbours: offsets by the number of correct non-holding neighbours. Both: the two together.

targets	binary	own	neighbours	both
six agents, all (52)	0.0115	0.0118	0.0118	0.0119
six agents, Ministral-8B (25)	0.0155	0.0159	0.0161	0.0161
Ministral-8B, g_{24}, g_{65}, g_{67} (9)	0.0270	0.0285	0.0281	0.0280
ten rounds (six agents) (3)	0.0255	0.0256	0.0200	0.0188

the error by coverage band. The median error falls from 0.46 below 30% coverage to 0.04–0.07 between 30% and 90%, and to 0.011 at 90% or more, where 69% of predictions are within 0.02. Coverage between 90% and 99% does not occur: a subset either contains every local state a target visits or misses a whole class of agents. Errors can still be large at full coverage (90th percentile 0.05), typically when the only calibration configuration for a class of agents behaves differently from the target, as for the complete graph predicted without it.

Where the kernel fails. The kernel's largest errors are for Ministral-8B: at round 10 without memory (0.05), because its groups stop changing after round 5, and on its riskiest graphs g_{24}, g_{65} and g_{67} (0.02–0.04), mostly with the adversary on a hub. A logistic response with the same state and question offsets more than halves the error on these nine configurations (0.012 against 0.027) while matching the kernel elsewhere (Appendix C.1), which suggests that the problem lies in the sparsely observed hub neighbourhoods, where the kernel relies on pooled cells and the logistic model on a smooth function. Graphs closer than about 0.03 are not resolved: the kernel's own error on a predicted difference is 0.014–0.020, about the standard error of a measured difference with four seeds of 600 questions (0.010–0.016) and twice the sampling error of the prediction (Appendix E.1).

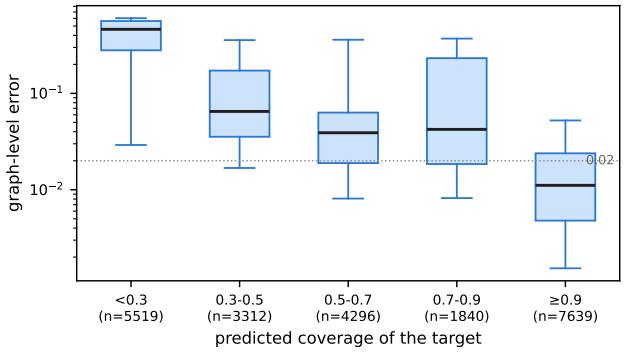

Figure 4: Graph-level error against predicted coverage, for kernels fitted on every subset of the calibration configurations and evaluated on every measured configuration outside the subset (22,606 predictions, all models and protocols). Boxes: interquartile range and median; whiskers: 10th and 90th percentiles; dotted line: 0.02.

Table 7 gives the error in and out of calibration per model and protocol, including the start a never-run graph would get, and Table 8 the tests with more rounds, larger groups and two adversaries. For groups larger than six the chain is sampled (Appendix A.9). Table 9 compares the kernel with predictors that use only the measured capture of the eight calibration configurations of each model and protocol: their mean, and least-squares regressions of capture on the adversary's degree and on the adversary's degree and the number of edges. All are evaluated on the 52 six-agent configurations not used in calibration, the kernel from each run's own round-0 answers. The regressions cannot predict other group sizes or horizons.

Table 7: Graph-level error, the absolute difference between predicted and measured capture rate, each averaged over the graph's questions. In calibration: mean over the eight calibration configurations (in-sample). Out of calibration: $K_{3,3}$, predicted from each run's own round-0 answers or, as for a graph that was never run, from the calibration distribution of round-0 answers (Appendix A.5). In parentheses: question-bootstrap standard error (notation of Table 1), reflecting the sampling noise of the measured capture rates with the predictions held fixed; an error near zero, such as 0.000 for $K_{3,3}$ with Qwen3-8B without memory, means agreement to within that noise, not an exact match.

			out of calibration ($K_{3,3}$)	
model	protocol	in calibration	own round 0	never-run start
Qwen3-8B	none	0.008(2)	0.000(2)	0.000
	own	0.009(1)	0.003(3)	0.003
	full	0.009(2)	0.004(3)	0.004
Ministral-8B	none	0.008(2)	0.007(6)	0.007
	own	0.011(2)	0.013(8)	0.014
	full	0.007(2)	0.006(6)	0.008
Llama-3.1-8B	none	0.009(2)	0.004(5)	0.002
	own	0.015(3)	0.003(5)	0.003
	full	0.005(1)	0.009(5)	0.010

Table 8: More rounds, larger groups and two adversaries, without memory. All graphs are vertex-transitive; with one adversary it has a single placement, and the two adversaries sit on opposite sides. $C_n(1,2)$: the 4-regular circulant (each agent linked to the two nearest agents on either side). Predicted by the kernel calibrated on six-agent, five-round runs, from each run's own round-0 state; measured over two seeds, with a 95% question-bootstrap interval. For ten rounds the round-5 value of the same runs is also shown (predicted / measured).

model	graph	round 5	predicted	measured [95% CI]	error
ten rounds, six agents					
Qwen3-8B	complete graph	0.034 / 0.024	0.044	0.034 [0.019, 0.051]	$+0.009$
Ministral-8B	complete graph	0.176 / 0.167	0.225	0.173 [0.141, 0.206]	$+0.052$
Llama-3.1-8B	ring	0.094 / 0.092	0.095	0.080 [0.059, 0.103]	$+0.014$
eight agents, five rounds					
Qwen3-8B	ring		0.059	0.057 [0.038, 0.077]	$+0.002$
Qwen3-8B	$C_8(1,2)$		0.038	0.037 [0.022, 0.055]	$+0.000$
Ministral-8B	ring		0.104	0.105 [0.081, 0.129]	-0.001
Ministral-8B	$C_8(1,2)$		0.116	0.133 [0.105, 0.160]	-0.017
Llama-3.1-8B	ring		0.057	0.064 [0.046, 0.084]	-0.007
Llama-3.1-8B	$C_8(1,2)$		0.053	0.059 [0.041, 0.079]	-0.006
twelve agents, two adversaries, five rounds (capture; spread in parentheses)					
Qwen3-8B	$C_{12}(1,2)$		0.055 (0.210)	0.050 [0.033, 0.069] (0.208)	$+0.004$ ($+0.002$)
Ministral-8B	$C_{12}(1,2)$		0.192 (0.317)	0.216 [0.183, 0.251] (0.329)	-0.024 (-0.011)
Llama-3.1-8B	$C_{12}(1,2)$		0.080 (0.257)	0.102 [0.077, 0.129] (0.263)	-0.023 (-0.006)

C.1 A PARAMETRIC RESPONSE

Our claim concerns the state, not the estimator, so we also fit parametric response functions to the same calibration observations and propagate them with the same exact chain (Table 10). Each is a logistic model, fitted separately for adopting and leaving, in the share b/d of holding neighbours, their number b, or both together with the degree d, always with an indicator for the adversary being a neighbour, an indicator for no holding neighbour and, under full history, round terms. The question offset adds one intercept per question for adopting and one for leaving, with a standard normal prior fixed in advance. The threshold rule is a cascade model in the sense of Granovetter (1978) with noise: an agent adopts with one probability if the share of holding neighbours reaches the question's threshold and with another if it does not, and leaves by the same rule applied to the share of non-holding neighbours.

Table 9: Kernel against simple predictors on the 52 six-agent configurations not used in calibration: mean and median absolute error, share within 0.02, and rank correlation between prediction and measurement within each model and protocol (mean over the nine).

predictor	mean	median	within 0.02	rank corr.
kernel	0.012	0.008	85%	0.96
mean calibration capture	0.103	0.092	4%	–
regression on adversary degree	0.046	0.033	33%	0.70
regression on adversary degree and edges	0.025	0.022	46%	0.90

The question level is needed in every form: without it the best logistic model has mean error 0.031, the same as the kernel's population cells. With it, the logistic model in share, count and degree predicts as well as the kernel (0.0094 against 0.0115, difference -0.0021 ± 0.0019). The two are tied on the 43 configurations outside Ministral-8B's risky end (0.0088 against 0.0083). On the nine risky-end configurations, where the kernel over-predicts capture on hub seats (Appendix C), the logistic model more than halves the error (0.0121 against 0.0270), presumably because it fills these sparsely observed neighbourhoods with a smooth function instead of pooled cells. A share or a count alone predicts worse (0.0168 and 0.0880 with question offsets), and so does the threshold rule (0.0195): agents respond both to how many and to what fraction of their neighbours hold the letter. We chose the logistic forms after the kernel and scored several of them on the same configurations, so we read the comparison as a tie rather than an improvement.

The fitted slopes differ by model. For adoption under own memory, Qwen3-8B responds steeply to the share (a coefficient of 4.9 on the logit scale, against 1.3 for the count), Ministral-8B mainly to the count (1.5, against 0.8 for the share and -0.6 per additional neighbour), and Llama-3.1-8B to the share but less steeply than Qwen3-8B (3.3, with 0.1 for the count). The question offsets for adoption have a standard deviation of 2.6–3.0 on the logit scale for Qwen3-8B and 1.4–1.6 for the other two models.

Ranking all 112 graphs for Ministral-8B with the logistic model instead of the kernel gives worst-case rank correlations of 0.96–0.98 with the kernel's ranking. Four or five of the five safest graphs are shared, the same hub graphs are the riskiest, and the safest graph changes from $K_{3,3}$ to the ring, which the kernel placed within 0.005 of $K_{3,3}$ (Section 6). With a random seat the correlation is 0.77–0.82, the tree g_3 remains the safest or second-safest graph, and the worst-case and random-seat rankings correlate at 0.16–0.25, against 0.12–0.36 with the kernel.

We keep the shrinkage kernel. It was fixed before any configuration outside the calibration was scored, assumes no functional form, and every cell is backed by counted observations, which the coverage check of Section 4 relies on; a smooth model predicts states it has never observed without warning. The logistic model is more compact, its slopes are easier to interpret, and it extends to degrees above five, which the table cannot reach.

Table 10: Response models, each fitted on the eight calibration configurations and propagated with the exact chain; mean graph-level error on the 52 six-agent configurations outside the calibration. Difference: paired difference from the kernel with its bootstrap standard error over configurations, in units of 10^{-4}. Parameters: per model and protocol without full history (full history adds four round terms per direction). Risky end: Ministral-8B's g_{24}, g_{65} and g_{67} (nine configurations); rest: the other 43.

response model	parameters	all	difference	sparse	full history	risky end	rest
kernel (hierarchical shrinkage)	table	0.0115	–	0.0078	0.0104	0.0270	0.0083
kernel, population cells only	table	0.0306	$+191 \pm 26$	0.0336	0.0385	0.0333	0.0301
logistic: share	8	0.0637	$+522 \pm 54$	0.0434	0.0559	0.1301	0.0498
logistic: share, question offset	8 + 2 per question	0.0168	$+53 \pm 25$	0.0091	0.0107	0.0456	0.0108
logistic: count	8	0.1498	$+1383 \pm 186$	0.0519	0.0870	0.2397	0.1309
logistic: count, question offset	8 + 2 per question	0.0880	$+765 \pm 108$	0.0277	0.0496	0.1341	0.0784
logistic: share, count, degree	12	0.0308	$+192 \pm 32$	0.0419	0.0344	0.0152	0.0340
logistic: share, count, degree, question offset	12 + 2 per question	0.0094	-21 ± 19	0.0087	0.0070	0.0121	0.0088
threshold rule, per-question threshold	4 + 2 per question	0.0195	$+79 \pm 23$	0.0203	0.0178	0.0219	0.0189

C.2 GROUPS CAPTURED BEFORE DELIBERATION

Eligibility requires one correct clean agent in round 0, but three other clean agents may already hold the adversary's letter, which with the adversary is a strict majority before any deliberation. Table 11 gives how often this happens in the calibration runs. It is rare for Qwen3-8B on MMLU-Pro and reaches 8% for Ministral-8B on LogiQA, whose four options put more agents on the adversary's letter. The kernel starts from each run's own round-0 answers, so these runs enter the prediction through its starting distribution. Removing them from both measurement and prediction leaves the mean error on the 52 six-agent configurations outside the calibration unchanged (0.0114 against 0.0115 with all runs, and 0.0114 when runs tied at round 0 are removed as well), while the mean measured capture falls from 0.137 to 0.128.

Table 11: Eligible runs already captured (at least three of the five clean agents on the adversary's letter) or tied (exactly two) at round 0, over the calibration runs. Round-0 answers do not depend on the memory protocol.

question set	model	eligible runs	captured at round 0	tied at round 0
MMLU-Pro	Qwen3-8B	12,481	0.3%	0.1%
MMLU-Pro	Ministral-8B	10,766	2.7%	3.5%
MMLU-Pro	Llama-3.1-8B	9,327	2.3%	3.9%
LogiQA	Qwen3-8B	10,456	2.0%	0.8%
LogiQA	Ministral-8B	13,488	8.4%	14.8%

D DISTRIBUTION SHIFT

D.1 A NEW QUESTION SET

We took the first 600 questions of the LogiQA test set (logical reasoning; four options each) and generated an adversary argument for each with the MMLU-Pro procedure (the wrong letter drawn uniformly with the same seed; Qwen3-4B, greedy, cut to 400 characters). With Qwen3-8B and Ministral-8B and own memory we ran the eight calibration configurations at four seeds and five test graphs at two seeds ($K_{3,3}$, g_1, g_{10}, g_{97} and g_{110}, at the placements used on MMLU-Pro), with the same prompts, message shuffle and eligibility rule. Four kernels predict each LogiQA configuration from its runs' own round-0 answers (Table 12): the MMLU-Pro kernel, which has no data on LogiQA questions and so uses its population cells for all of them; LogiQA's own kernel at the population level only; LogiQA's full kernel; and a hybrid whose population levels come from MMLU-Pro and whose per-question cells come from LogiQA runs. The calibration configurations are predicted by kernels fitted without them, which is possible without losing coverage for the star, the complete graph, g_{19} and g_6.

Figure 5 shows the error on the five LogiQA test graphs against the number of LogiQA calibration configurations.

The population cells agree between the question sets: the 36 cells $\pi^{(1)}(o, d, b)$ with at least 50 observations in both correlate at 0.988, and the 49 cells $\pi^{(2)}(o, d, a, c)$ at 0.984, with a mean difference of 0.026 weighted by how often LogiQA agents visit each cell (Qwen3-8B); for Ministral-8B, 38 cells at 0.991 and 58 at 0.989, with a weighted mean difference of 0.072. Table 13 shows what the MMLU-Pro population levels are worth when LogiQA data are scarce. Ranked with their own kernels, the two question sets give almost the same ranking of the 112 graphs for Qwen3-8B (rank correlation 0.988), with the complete graph safest and the star riskiest in both; the MMLU-Pro kernel with LogiQA's round-0 answers does not reproduce it (0.805). For Ministral-8B the rank correlation is 0.908: the safest graph moves from $K_{3,3}$ to the ring and the complete graph from rank 4 to 27, and the MMLU-Pro kernel with LogiQA's round-0 answers reproduces most of this (0.952).

The question level is not one number per question. Replacing each question's cells by a single log-odds shift of the population cells, fitted to the same data, predicts the 52 MMLU-Pro configurations outside the calibration worse than no question level (0.042–0.050 against 0.031; question cells 0.012), because a shift fitted in the states a question usually visits over-predicts switching in the states it

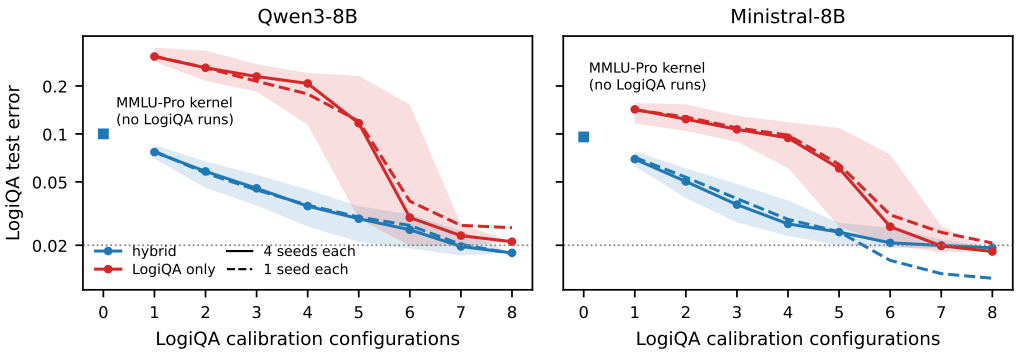

Figure 5: Error on the five LogiQA test graphs against the number of LogiQA calibration configurations, median and interquartile range over all subsets. Hybrid: population levels from MMLU-Pro, question cells from LogiQA runs.

Table 12: LogiQA, own memory: measured capture rate and prediction error (predicted minus measured) of four kernels. Population levels: the pooled cells $\pi^{(1)}$ and $\pi^{(2)}$; question level: the per-question cells.

		MMLU-Pro	LogiQA		hybrid
graph	measured	kernel	population only	full kernel	
Qwen3-8B					
left-out calibration configurations					
star	0.471	+0.082	+0.330	+0.069	+0.068
complete graph	0.162	−0.031	+0.100	+0.040	+0.047
g_{19}	0.408	−0.062	+0.164	+0.022	+0.017
g_6	0.519	−0.086	+0.111	−0.050	−0.051
mean absolute error		0.065	0.176	0.045	0.046
test graphs					
$K_{3,3}$	0.176	−0.122	−0.042	+0.010	+0.003
g_1	0.438	−0.093	+0.092	+0.036	+0.036
g_{10}	0.351	−0.128	+0.072	+0.046	+0.042
g_{97}	0.191	−0.100	−0.008	+0.011	+0.004
g_{110}	0.194	−0.057	+0.045	+0.002	+0.005
mean absolute error		0.100	0.052	0.021	0.018
Ministral-8B					
left-out calibration configurations					
star	0.697	+0.005	+0.093	−0.007	−0.008
complete graph	0.458	−0.067	+0.137	+0.060	+0.014
g_{19}	0.624	+0.001	+0.126	+0.035	+0.038
g_6	0.679	−0.003	+0.092	+0.002	+0.005
mean absolute error		0.019	0.112	0.026	0.016
test graphs					
$K_{3,3}$	0.406	−0.146	−0.027	+0.014	+0.012
g_1	0.624	−0.094	+0.026	−0.014	−0.014
g_{10}	0.567	−0.095	+0.052	+0.018	+0.016
g_{97}	0.437	−0.081	+0.008	+0.019	+0.023
g_{110}	0.467	−0.061	+0.058	+0.027	+0.031
mean absolute error		0.096	0.034	0.018	0.019

rarely visits. With only one or two LogiQA configurations, however, such a shift makes better use of the data for Ministral-8B (0.034 and 0.023, against 0.070 and 0.050 with question cells).

Table 13: Mean absolute error on the five LogiQA test graphs when the per-question cells come from only some of the eight LogiQA calibration configurations (four runs each): median and range over all choices of configurations.

model	configurations	choices	LogiQA-only kernel	hybrid	hybrid better
Qwen3-8B	1	8	0.306 (0.212–0.529)	0.077 (0.049–0.091)	8 of 8
	2	28	0.260 (0.134–0.523)	0.058 (0.029–0.089)	28 of 28
	8	1	0.021	0.018	–
Ministral-8B	1	8	0.143 (0.101–0.216)	0.070 (0.039–0.088)	8 of 8
	2	28	0.124 (0.072–0.204)	0.050 (0.021–0.083)	28 of 28
	8	1	0.018	0.019	–

D.2 Other adversary messages

The paper's adversary repeats one fixed argument per question, written once by Qwen3-4B. To see whether the kernel and the graph ranking depend on that message, we ran Qwen3-8B and Ministral-8B with own memory under three other attacks, all for the same wrong option per question: a *resampled argument* (the same prompt and model, sampled at temperature 1.0 instead of greedily, cut to 400 characters); a *bare assertion* (only the letter and "Confidence: 95"); and an *adaptive adversary*, which sends the fixed argument in round 0 and in every later round has the model under test write a new two-sentence argument for the option in reply to its neighbours' latest messages. Each attack was run on the eight calibration configurations and three test configurations ($K_{3,3}$ with the adversary at placement 0, g_3 and g_{110} at placement 1) at seeds 0 and 1, with the same prompts, message shuffle and eligibility rule. For a like-for-like comparison the paper's runs are restricted to the same two seeds. Under the paper's argument there are no runs of g_3 at placement 1 for Qwen3-8B or of g_{110} at placement 1 for Ministral-8B, so its test error is over two graphs.

Table 14 reports, for each attack: its own kernel's error on the test configurations; the error of the kernel calibrated under the paper's argument; the error of a hybrid with population levels from the paper's argument and per-question cells from the attack's runs; the correlation of the pooled cells $\pi^{(2)}(o, d, a, c)$ with at least 50 observations under both; the correlation of per-question capture, averaged over the calibration configurations, with the paper's argument; and the rank correlation of the 112 graphs (worst-case placement, the start of Appendix A.5) with the paper's argument. The near-best set lists the graphs predicted within 0.01 of the safest, safest first. For Qwen3-8B, with two seeds the safest graph alternates between the complete graph and the octahedron, also under the paper's argument (with four seeds it is the complete graph, Table 15); for Ministral-8B the ring is near-best under every attack.

Table 14: Own memory, seeds 0–1: capture averaged over the eight calibration configurations, mean absolute error on the test configurations for three kernels, agreement with the paper's argument, and the near-best set of graphs (worst-case placement).

		test error			correlation with paper's			
attack	capture	own	paper's	hybrid	cells	questions	ranking	near-best set
Qwen3-8B								
fixed argument (paper)	0.160	0.010	–	–	–	–	–	oct., $K_{3,3}$, complete, prism, ring
resampled argument	0.156	0.011	0.017	0.013	0.99	0.87	0.996	complete, $K_{3,3}$, prism
bare assertion	0.116	0.010	0.048	0.009	0.97	0.71	0.977	complete, oct., g_{110}, $K_{3,3}$
adaptive adversary	0.206	0.008	0.031	0.012	0.99	0.85	0.989	oct., complete, $K_{3,3}$, prism
Ministral-8B								
fixed argument (paper)	0.298	0.010	–	–	–	–	–	$K_{3,3}$, ring, prism
resampled argument	0.298	0.013	0.014	0.013	1.00	0.89	0.995	ring
bare assertion	0.206	0.015	0.101	0.011	0.97	0.79	0.971	complete, $K_{3,3}$, prism, ring
adaptive adversary	0.335	0.008	0.025	0.006	1.00	0.88	0.986	ring

Table 15: Predicted safest and riskiest of the 112 connected six-agent graphs when the adversary picks its seat (worst case) and when it lands on a random seat (average case), with predicted/measured capture rate (leading zeros dropped; † = graph not used in calibration; average-case values measured at every seat). $^{+k}$: k further graphs predicted within 0.01 of the safest (the near-best set). ρ: rank correlation between the worst-case and average-case rankings. Graphs are drawn in Figure 6; the five safest and riskiest under each placement are in Tables 16 and 17.

		worst case				average case				
model	protocol	safest	pred./meas.	riskiest	pred./meas.	safest	pred./meas.	riskiest	pred./meas.	ρ
Qwen3-8B	none	complete	.036/.029	star	.358/.340	complete^{+1}	.036/.029	g_6	.094/.087	0.82
	own	complete^{+3}	.064/.057	star	.400/.365	$g_8{}^{+30}$	.062/.060†	g_6	.103/.097	0.83
	full	complete^{+3}	.031/.028	star	.351/.324	complete^{+13}	.031/.028	star	.075/.073	0.82
Ministral-8B	none	$K_{3,3}{}^{+3}$	.170/.177†	g_{24}	.436/.414†	$g_3{}^{+3}$	.134/.132†	g_{65}	.210/.172†	0.36
	own	$K_{3,3}{}^{+2}$	.192/.177†	star	.570/.571	$g_3{}^{+1}$	.161/.161†	g_{67}	.228/.199†	0.12
	full	$K_{3,3}{}^{+2}$	.171/.163†	star	.527/.523	$g_3{}^{+1}$	.150/.143†	g_{65}	.211/.177†	0.29
Llama-3.1-8B	none	ring^{+4}	.096/.103	star	.260/.292	path^{+59}	.095/.098†	complete	.118/.122	0.49
	own	ring	.100/.106	star	.486/.509	ring^{+8}	.100/.106	star	.162/.156	0.59
	full	ring^{+7}	.076/.072	star	.207/.199	$g_3{}^{+78}$	.074/.063†	complete	.104/.101	0.39

E CHOOSING A TOPOLOGY

Table 15 gives the safest and riskiest graph under each score with predicted and measured capture, Table 16 the five safest and riskiest graphs under worst-case placement, and Table 17 under average-case placement. Figure 6 draws the named graphs. Under average-case placement the gap between the safest and riskiest graph shrinks from 0.32 to 0.06 for Qwen3-8B without memory, from 0.27 to 0.08 for Ministral-8B and from 0.16 to 0.02 for Llama-3.1-8B; the star moves from rank 108 to 6 for Ministral-8B without memory, and for Qwen3-8B and Llama-3.1-8B up to 79 graphs are predicted within 0.01 of the safest, so for these models the choice of graph matters little when the seat is random.

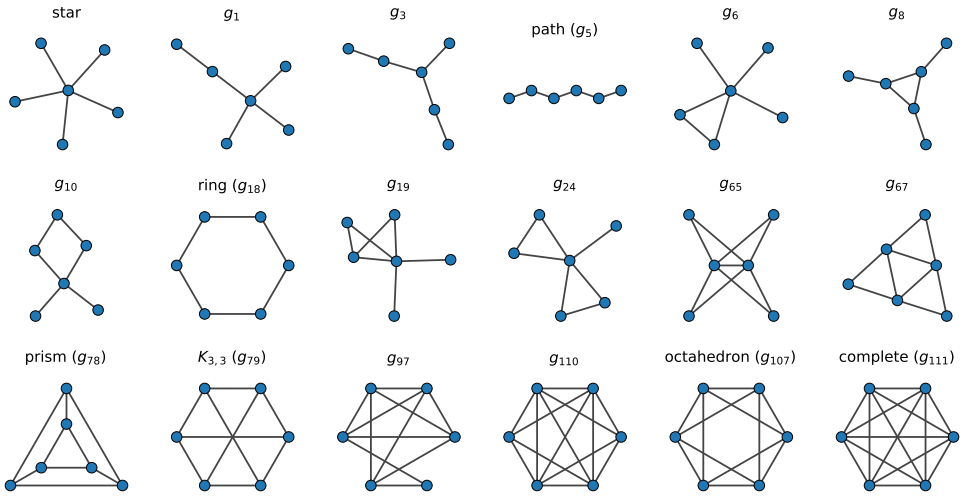

Figure 6: The six-agent graphs named in the paper. g_k is the k-th connected six-node graph in the order of the graph atlas (Read & Wilson, 1998), counted from 0.

E.1 RESOLUTION OF THE RANKINGS

A predicted difference between two graphs has two sources of error. The first is sampling: for each model and protocol we resampled, 200 times, the runs within each calibration configuration and then the 600 questions, refitted the kernel on each resample (a question drawn twice keeps its own cells and counts twice in the average) and recomputed the worst-case prediction of the 25 graphs

Table 16: The five safest (ranks 1–5) and five riskiest (ranks 108–112) of the 112 connected six-agent graphs under the worst-case adversary placement, for each model and memory protocol: predicted and measured capture rate (leading zeros dropped; measured values with their standard error over questions, notation of Table 1; – = not run at that placement; † = graph not used in calibration, otherwise a calibration configuration). oct. = octahedron. Unnamed graphs are indexed by their position in the atlas of graphs (Read & Wilson, 1998).

Qwen3-8B

	none			own			full		
rank	graph	pred.	meas.	graph	pred.	meas.	graph	pred.	meas.
1	complete	.036	.029(7)	complete	.064	.057(10)	complete	.031	.028(7)
2	g_{110}	.049	–	$K_{3,3}$	.068	.071(11)†	g_{110}	.039	–
3	oct.	.052	.048(9)	oct.	.068	.064(10)	oct.	.040	.039(8)
4	$K_{3,3}$	.056	.056(9)†	prism	.069	.067(10)	$K_{3,3}$	.041	.045(8)†
5	g_{109}	.058	–	g_{97}	.078	.069(11)†	prism	.044	.043(8)
108	g_7	.246	–	g_7	.259	.254(17)†	g_{10}	.207	–
109	g_{19}	.264	.270(18)	g_{19}	.279	.278(17)	g_{19}	.221	.228(16)
110	g_1	.339	–	g_1	.354	.323(21)†	g_1	.301	–
111	g_6	.344	.352(22)	g_6	.361	.376(21)	g_6	.309	.335(21)
112	star	.358	.340(22)	star	.400	.365(21)	star	.351	.324(21)

Ministral-8B

	none			own			full		
rank	graph	pred.	meas.	graph	pred.	meas.	graph	pred.	meas.
1	$K_{3,3}$	.170	.177(14)†	$K_{3,3}$	.192	.177(15)†	$K_{3,3}$	.171	.163(14)†
2	ring	.170	.159(13)	ring	.195	.176(14)	prism	.174	.171(14)
3	oct.	.176	.169(15)	prism	.195	.189(14)	ring	.176	.170(14)
4	prism	.176	.169(14)	complete	.207	.196(15)	complete	.190	.176(15)
5	path	.188	–	g_{59}	.222	–	oct.	.192	.191(15)
108	star	.410	.415(19)	g_1	.488	–	g_{19}	.437	.422(20)
109	g_{19}	.412	.407(19)	g_{19}	.488	.469(19)	g_1	.443	–
110	g_{49}	.426	–	g_{24}	.493	–	g_{24}	.449	–
111	g_6	.427	.419(19)	g_6	.540	.536(20)	g_6	.493	.502(20)
112	g_{24}	.436	.414(21)†	star	.570	.571(19)	star	.527	.523(20)

Llama-3.1-8B

	none			own			full		
rank	graph	pred.	meas.	graph	pred.	meas.	graph	pred.	meas.
1	ring	.096	.103(11)	ring	.100	.106(11)	ring	.076	.072(10)
2	prism	.097	.105(11)	prism	.112	.120(12)	oct.	.079	.076(10)
3	g_{58}	.105	–	$K_{3,3}$	.112	.108(12)†	prism	.080	.076(10)
4	$K_{3,3}$	.105	.107(12)†	oct.	.120	.125(12)	$K_{3,3}$	.080	.071(10)†
5	g_{57}	.105	–	g_{59}	.121	–	path	.084	–
108	g_1	.206	–	g_{24}	.297	–	g_1	.159	–
109	g_{19}	.216	.207(16)	g_{19}	.324	.290(17)	g_{19}	.163	.169(15)
110	g_{24}	.222	–	g_1	.364	–	g_{24}	.166	–
111	g_6	.239	.232(16)	g_6	.398	.372(19)	g_6	.184	.189(16)
112	star	.260	.292(17)	star	.486	.509(20)	star	.207	.199(16)

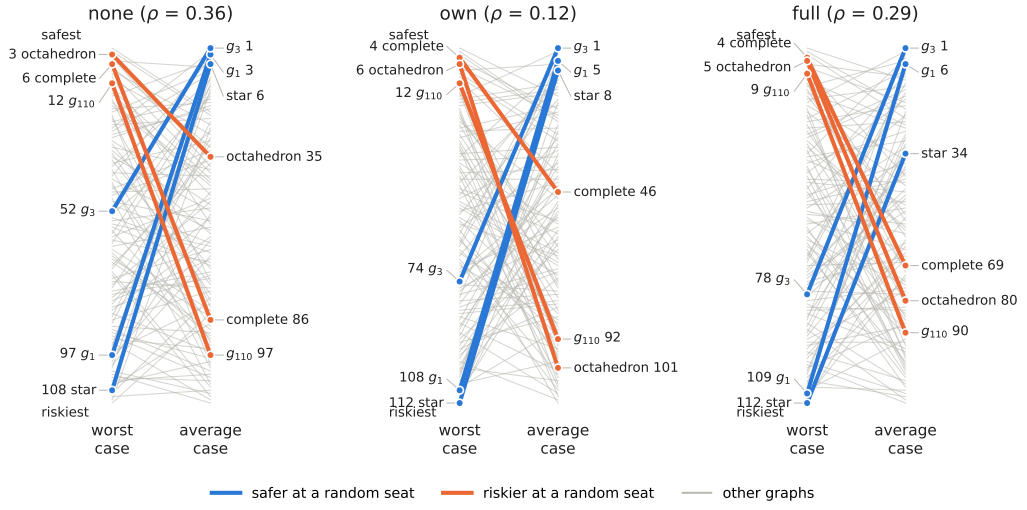

Figure 7: Ministral-8B's worst-case and average-case rankings in all three memory protocols. ρ: rank correlation between the two rankings. g_1: a hub with four neighbours, one of which has a further leaf; g_3: a tree with one branch point; g_{110}: the complete graph minus one edge.

Table 17: Average-case placement: the five safest and five riskiest graphs with their predicted capture rate.

	Qwen3-8B			Ministral-8B			Llama-3.1-8B		
rank	none	own	full	none	own	full	none	own	full
1	complete 0.036	g_8 0.062	complete 0.031	g_3 0.134	g_3 0.161	g_3 0.150	path 0.095	ring 0.100	g_3 0.074
2	g_{110} 0.043	complete 0.064	g_{110} 0.034	g_2 0.135	g_8 0.170	g_8 0.156	g_{57} 0.096	g_{35} 0.104	path 0.074
3	g_{94} 0.047	g_3 0.065	g_8 0.037	g_1 0.138	g_2 0.172	g_2 0.164	ring 0.096	g_{36} 0.105	g_{12} 0.075
4	g_{109} 0.048	g_{23} 0.067	g_{109} 0.037	g_4 0.143	g_4 0.177	g_{12} 0.164	g_{34} 0.097	g_{34} 0.107	g_{17} 0.076
5	g_8 0.051	$K_{3,3}$ 0.068	g_{108} 0.037	g_8 0.146	g_1 0.177	path 0.166	prism 0.097	g_{17} 0.108	g_4 0.076
108	g_2 0.087	star 0.090	g_{14} 0.065	g_{49} 0.203	g_{96} 0.226	g_{49} 0.201	g_{104} 0.114	g_{19} 0.137	g_{82} 0.093
109	g_{33} 0.087	g_{22} 0.091	g_1 0.067	g_{69} 0.203	g_{83} 0.227	g_{66} 0.201	star 0.114	g_{38} 0.138	g_{65} 0.095
110	g_{19} 0.087	g_2 0.093	g_2 0.071	g_{98} 0.204	g_{84} 0.227	g_{40} 0.203	g_{110} 0.116	g_1 0.140	g_{104} 0.096
111	g_{24} 0.087	g_{19} 0.097	g_6 0.072	g_{82} 0.207	g_{65} 0.228	g_{82} 0.204	g_{65} 0.116	g_6 0.146	g_{110} 0.097
112	g_6 0.094	g_6 0.103	star 0.075	g_{65} 0.210	g_{67} 0.228	g_{65} 0.211	complete 0.118	star 0.162	complete 0.104

predicted safest; the standard deviation of a graph's difference from the safest graph over resamples is its sampling error. The second is the kernel's own error, which we calibrate with the measured configurations outside the calibration. For every pair of such configurations i, j within a model and protocol (190 pairs), let p_i be the kernel's predicted capture rate of configuration i and m_i its measured capture rate, both averaged over the n_{ij} questions the two configurations share. The residual of the predicted difference,

$$e_{ij} = (p_i - p_j) - (m_i - m_j) = \frac{1}{n_{ij}} \sum_q e_{ij,q}, \tag{17}$$

is the mean of the per-question residuals $e_{ij,q}$ and contains both the kernel's error and the sampling noise of the comparison. The sampling part is estimated by the variance of that mean, s_{ij}^2/n_{ij}, with s_{ij}^2 the variance of $e_{ij,q}$ across questions, and removed:

$$\sigma_d^2 = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(e_{ij}^2 - \frac{s_{ij}^2}{n_{ij}} \right), \tag{18}$$

over a set of pairs $\mathcal{P}$. This gives $\sigma_d = 0.020$ over all pairs and 0.014 over the 82 pairs whose measured capture differs by less than 0.05 (for a single configuration, 0.015). Removing only the run-to-run noise of the measurements, estimated from the seed-to-seed differences of the four-seed calibration runs, instead of the question-sampling variance gives the same value over all pairs (0.020). Near-best graphs have similar capture, so we use 0.014: the total standard error of a predicted

difference is $\sqrt{\mathrm{SE}^2 + \sigma_d^2}$, and the probability that a graph is the safest adds independent normal noise with variance $\sigma_d^2/2$ to every graph in every resample. The kernel's errors on similar dense graphs may be more correlated than on the pairs we could measure, in which case this is conservative. For comparison, the standard error of a measured difference is computed from the per-question capture of two calibration configurations on the same questions (four seeds). Table 18 lists both standard errors, the near-best sets at the paper's threshold, and the graphs most likely to be the safest.

Table 18: Worst-case placement. SE samp.: median sampling error of the predicted difference between the safest graph and the next ten; SE total: with the kernel's own error added. SE meas.: median standard error of the measured difference between two calibration configurations. Near-best: graphs predicted within 0.01 of the safest, safest first. P(safest): graphs that are the safest with probability at least 0.10 once both errors are included.

model	protocol	SE samp.	SE total	SE meas.	near-best (0.01)	P(safest)
Qwen3-8B	none	0.0071	0.0154	0.0150	complete	complete 0.68, g_{110} 0.11
	own	0.0075	0.0156	0.0141	complete, $K_{3,3}$, oct., prism	complete 0.39, oct. 0.22, $K_{3,3}$ 0.17, prism 0.10
	full	0.0063	0.0150	0.0133	complete, g_{110}, oct., $K_{3,3}$	complete 0.47, g_{110} 0.17, oct. 0.10
Ministral-8B	none	0.0125	0.0185	0.0156	$K_{3,3}$, ring, oct., prism	ring 0.32, oct. 0.25, $K_{3,3}$ 0.23, prism 0.13
	own	0.0074	0.0155	0.0155	$K_{3,3}$, ring, prism	$K_{3,3}$ 0.33, ring 0.30, prism 0.21, complete 0.14
	full	0.0089	0.0164	0.0146	$K_{3,3}$, prism, ring	$K_{3,3}$ 0.36, ring 0.24, prism 0.20, complete 0.10
Llama-3.1-8B	none	0.0067	0.0152	0.0117	ring, prism, g_{58}, $K_{3,3}$, g_{57}	ring 0.28, prism 0.21
	own	0.0080	0.0158	0.0140	ring	ring 0.60, prism 0.12, $K_{3,3}$ 0.12
	full	0.0087	0.0162	0.0102	ring, oct., prism, $K_{3,3}$, path, g_{59}, g_{102}, g_{101}	ring 0.27, prism 0.14, oct. 0.13, $K_{3,3}$ 0.12

E.2 What makes the safest graph differ between models

For Section 6 we re-rank all 112 graphs, under worst-case and average-case placement, with one component of a model's kernel changed: converted peers treated like the adversary; one pooled table for every question; adoption or abandonment cells taken from another model question by question; another model's round-0 answers with the model's own kernel; and another model's whole kernel with the model's own round-0 answers. To make the swaps question by question, all variants use the 333 questions eligible for all three models, on which the unmodified rankings differ slightly from Table 15. The round-0 answers change the safest graph most. The next largest effect is how switching varies across questions: with one pooled table for every question, $K_{3,3}$ replaces the complete graph as Qwen3-8B's safest graph in all three protocols, while the rankings of the other two models keep a rank correlation of 0.88–0.98 with their own. Treating converted peers like the adversary changes the least (rank correlation at least 0.99 with the unmodified ranking in every setting); where it matters it favours sparse graphs, moving the ring from rank 1 to 3 for Llama-3.1-8B with own memory. Swapping only the adoption cells or only the abandonment cells moves a ranking part of the way toward the other model's, with no consistent difference between the two halves. The code is `scripts/analyze_drivers.py` and the results are in `results/analysis_drivers.json`.

E.3 How the best graph moves with the start

For Figure 3 we run the three-state chain once per model and protocol, for all 112 graphs at every placement, from each question's measured round-0 distribution (eligible runs), and keep every question's predicted capture and accuracy. A question set that starts differently is emulated by reweighting the real questions, $w_q \propto \exp(\beta z_q)$ with z_q standardised, where z_q is either the probability

that question q's correct answer starts as a minority (one or two of five clean agents correct, among eligible groups) or its mean round-0 share of clean agents on the adversary's letter. Every question keeps its own start distribution and its own kernel cells, so questions that start differently also bring their own persuadability. The intervention is therefore on the mixture of observed question types, not on their dynamics: correlations between a question's initial group state and its susceptibility are preserved. For each β the weighted mean of z_q gives the x-value, and the graphs are ranked by weighted worst-case accuracy or capture. We keep values of β whose effective number of questions, $(\sum_q w_q)^2 / \sum_q w_q^2$, is at least 100. Qwen3-8B's curves therefore end early: only about 30 of its 459 eligible questions ever start with the correct answer as a minority, because its agents mostly all know the answer or all miss it. Table 20 gives the ring's advantage over the complete graph at both ends of each curve, with a question-bootstrap standard error (questions resampled with their weights, kernel not refitted, so the kernel's own error of about 0.014 on a difference is not included), and where it changes sign.

For LogiQA we fit the chain on its own eight calibration configurations (Qwen3-8B and Ministral-8B, own memory). Its points agree with the MMLU-Pro curves in sign for both models and both targets, but not always in size: for Ministral-8B the ring's predicted advantage on LogiQA is $+0.023$ in accuracy and $+0.10$ in capture, against 0.00 and $+0.065$ on the MMLU-Pro curve at the same x. To read Figure 3 for LogiQA, we take the start property of LogiQA from its round-0 answers and use the MMLU-Pro ranking at the reweighting whose start property is closest. Compared with the ranking from the chain fitted on LogiQA's own calibration runs, the read-off has rank correlation 0.988 (capture) and 0.990 (accuracy) for Qwen3-8B and 0.949 and 0.965 for Ministral-8B. Using the MMLU-Pro ranking unchanged gives the same for Qwen3-8B, whose start moves little between the question sets, and 0.908 and 0.962 for Ministral-8B. For Ministral-8B's capture the read-off moves the complete graph from fourth to tenth safest, in the direction of LogiQA's own ranking, where it is 27th; which of the nearly tied sparse regular graphs comes first is not resolved. The alternative of combining LogiQA's round-0 answers with the MMLU-Pro population cells alone, which treats every LogiQA question as averagely persuadable, reproduces LogiQA's capture ranking at 0.952 for Ministral-8B but only 0.805 for Qwen3-8B (Appendix D.1). The reweighting keeps each question's own persuadability together with its start, so a link between the two, such as questions that start on the adversary's letter being less persuadable than average, carries over; with population cells alone it is lost, which is the likely reason for the lower value. Of the three predictors, only the read-off matches LogiQA's own ranking for both models. Measured on LogiQA's calibration runs, the ring is both more accurate than the complete graph for Ministral-8B ($+0.025$, standard error 0.009) and safer (capture 0.091 lower, standard error 0.010), while for Qwen3-8B the two are tied on both. Figure 8 gives the capture row of Figure 3: as more agents start on the adversary's letter, Ministral-8B's complete graph falls to about rank 20 and the ring becomes the safest, while Llama-3.1-8B's complete graph stays at rank 10–15 and Qwen3-8B's questions never reach this regime.

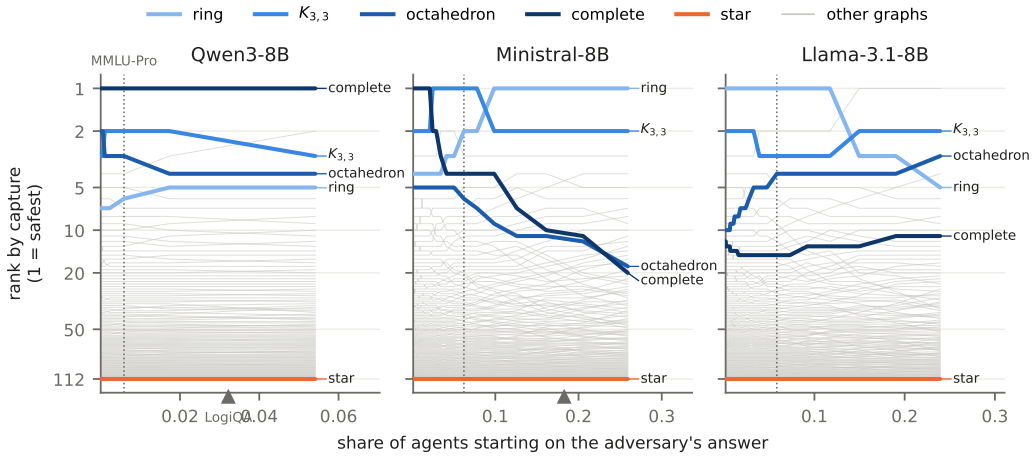

Figure 8: As Figure 3, for capture: each graph's worst-case rank by capture (1 = safest) as the question set is reweighted toward questions on which more clean agents start on the adversary's letter.

Table 19 gives the three rates behind the explanation of Llama-3.1-8B's direction in Section 6, measured on the complete graph without an adversary (two seeds, rounds 1–5), where every agent sees the other five. Drift: the probability that an agent leaves the correct answer when it and all five neighbours hold it. Overwrite: the probability that a correct agent takes an answer held by at least three of its five neighbours. Minority pull: the probability that an agent holding a wrong answer takes the correct one when exactly one neighbour holds it (and no majority does). Drift separates the models most: Llama-3.1-8B's agents leave a unanimous correct answer 14 to 210 times as often as Ministral-8B's (by protocol), while Qwen3-8B's almost never do, and Llama-3.1-8B's do so least often under full history, where its preference for the complete graph is also weakest (Table 20).

Table 19: Noise and conformity on the complete graph without an adversary: rates per round in percent, with 95% question-bootstrap intervals.

model	protocol	drift	overwrite	minority pull
Qwen3-8B	none	0.00 [0.00, 0.00]	63.0 [52.6, 73.1]	21.5 [12.0, 30.7]
	own	0.00 [0.00, 0.01]	59.6 [49.0, 69.2]	26.7 [15.6, 38.6]
	full	0.01 [0.00, 0.02]	44.0 [33.1, 54.7]	25.1 [15.2, 36.2]
Ministral-8B	none	0.07 [0.02, 0.15]	60.2 [55.5, 64.8]	11.2 [9.1, 13.4]
	own	0.11 [0.03, 0.21]	51.9 [47.0, 56.7]	10.5 [8.6, 12.5]
	full	0.18 [0.09, 0.28]	50.7 [46.7, 55.3]	10.5 [8.4, 12.7]
Llama-3.1-8B	none	15.01 [12.69, 17.83]	47.5 [43.6, 51.4]	19.8 [18.2, 21.4]
	own	6.75 [5.77, 7.86]	38.9 [35.1, 42.6]	15.3 [13.7, 17.1]
	full	2.49 [1.95, 3.06]	43.6 [39.9, 47.3]	13.6 [11.9, 15.3]

Table 20: Advantage of the ring over the complete graph along the reweighting (worst-case placement). Accuracy: ring minus complete, against the share of groups whose correct answer starts as a minority. Capture: complete minus ring (positive = ring safer), against the share of agents starting on the adversary's letter. Range: x reached with at least 100 effective questions; low, high: the advantage at the two ends, with its standard error (notation of Table 1); sign change: x where the advantage crosses zero.

		accuracy				capture			
model	protocol	range	low	high	sign change	range	low	high	sign change
Qwen3-8B	none	0.00–0.52	$-0.060(9)$	$+0.002(23)$	0.50	0.00–0.05	$-0.053(9)$	$-0.042(10)$	none
	own	0.00–0.52	$-0.030(6)$	$+0.047(29)$	0.21	0.00–0.05	$-0.017(5)$	$-0.009(6)$	none
	full	0.00–0.52	$-0.028(6)$	$+0.042(24)$	0.22	0.00–0.05	$-0.018(5)$	$-0.013(5)$	none
Ministral-8B	none	0.01–0.99	$-0.067(10)$	$+0.055(15)$	0.63	0.00–0.26	$-0.001(8)$	$+0.105(21)$	0.01
	own	0.01–0.99	$-0.045(8)$	$+0.033(14)$	0.64	0.00–0.26	$-0.018(8)$	$+0.095(25)$	0.04
	full	0.01–0.99	$-0.061(12)$	$+0.058(14)$	0.56	0.00–0.26	$-0.009(9)$	$+0.084(21)$	0.03
Llama-3.1-8B	none	0.01–0.99	$-0.004(11)$	$-0.061(8)$	none	0.00–0.24	$+0.014(4)$	$+0.020(14)$	none
	own	0.01–0.99	$-0.006(10)$	$-0.050(10)$	none	0.00–0.24	$+0.030(5)$	$+0.013(21)$	none
	full	0.01–0.99	$-0.012(7)$	$-0.012(13)$	none	0.00–0.24	$+0.016(5)$	$+0.032(12)$	none

E.4 Final accuracy

For Section 6 we predict the group's final accuracy with the three-state chain of Appendix B.1 without offsets: adopting and leaving the adversary's letter use the paper's cells κ_q, and moves between the correct and other wrong answers use the Beta–Binomial chain τ_q. The chain does not record which wrong letter an agent holds, so the plurality is computed twice, with all other-wrong agents on one letter and with each on its own letter, and we use the midpoint. Accuracy is the probability that the plurality answer after round 5 is correct, ties split, over eligible runs. For the ranking, each question starts from its calibration distribution of correct, other-wrong and holding agents, spread over interchangeable agents, as in Appendix A.5; worst-case placement is the placement with the lowest accuracy. Table 21 gives the prediction error on the configurations outside the calibration, and Table 22 compares the accuracy ranking with the capture ranking.

Table 21: Final accuracy on the configurations outside the calibration (number in parentheses): measured accuracy, mean absolute error of the three-state prediction (midpoint of the two lumpings, against the exact plurality), the same with the measured runs lumped in the same way, the error that lumping alone causes in the measured runs, the correlation of prediction and measurement, and the error of each model and protocol's mean calibration accuracy.

model	measured	error	same lumping	lumping only	correlation	constant
Qwen3-8B (17)	0.875	0.009	0.009	0.002	0.99	0.105
Ministral-8B (25)	0.565	0.017	0.015	0.008	0.98	0.089
Llama-3.1-8B (10)	0.546	0.021	0.018	0.018	0.97	0.078
all (52)	0.663	0.015	0.014	0.008	0.99	0.092

Table 22: Ranking the 112 graphs by predicted accuracy. ρ: rank correlation with the capture ranking (worst-case and average-case placement). Most accurate graph under worst-case placement with its accuracy; rank of the complete graph and its gap to the most accurate; rank of the ring by capture and by accuracy; least accurate graph.

model	protocol	ρ worst	ρ average	most accurate	complete	ring	least accurate
Qwen3-8B	none	1.00	0.98	complete (0.935)	1 (0.000)	$27 \to 33$	star
	own	0.99	0.92	complete (0.909)	1 (0.000)	$6 \to 11$	star
	full	1.00	0.97	complete (0.942)	1 (0.000)	$6 \to 12$	star
Ministral-8B	none	0.91	0.19	$K_{3,3}$ (0.583)	4 (0.006)	$2 \to 10$	star
	own	0.97	0.49	$K_{3,3}$ (0.568)	3 (0.005)	$2 \to 5$	star
	full	0.96	0.38	$K_{3,3}$ (0.582)	4 (0.012)	$3 \to 8$	star
Llama-3.1-8B	none	0.81	-0.37	complete (0.496)	1 (0.000)	$1 \to 24$	star
	own	0.90	0.15	complete (0.535)	1 (0.000)	$1 \to 9$	star
	full	0.82	-0.34	oct. (0.614)	9 (0.015)	$1 \to 24$	star