

Behaviour Beats Judgement: Auditing Collusion Between LLM Agents with Monitor Views and Antitrust-Style Evidence

Ayush Singh
student

Param Bapu
student

Atharva Patil
student

Neerada Dathan
student

With Apart Research

Abstract—It is possible for AI agents to coordinate their actions privately, so that collusion cannot be detected from the logged trajectories. This study looks at how different trajectory representations influence the ability of language models to detect collusion and identifies which evidence most strongly supports collusion decisions. We created a pre-registered testbed in which five locally hosted `qwen2.5:7b` agents play an eight-round version of the El Farol bar game. In the cases where collusion takes place, two of the agents receive an alliance note and have access to a private communication channel, while three control conditions (placebo, interaction with a friendly persona and a shared private incentive) produce similar behaviour in the absence of collusion. The ground-truth labels are obtained solely from the generation configuration. Two monitors, `llama3.1:8b` and `granite3.3:8b`, assess five different game representations together with an LLM judge and computational behavioural signals. The methods were established using 24 development games, tested on 24 games, and then replicated on 48 further games.

Behavioural signals were found to be more reliable than those based on LLMs. On the replication dataset, the capture signal reached an AUROC of 0.900 [0.811, 0.965], which was 0.264 [0.121, 0.412] better than the best LLM representation, while the coalition-gain rule obtained an AUROC of 0.885 [0.786, 0.961]. The LLM monitors performed in an inconsistent manner, their AUROC values ranging from 0.32 to 0.66; a structured report fell from 0.896 on the test set to 0.636 on the replication set. The message-only analysis stayed at a level close to chance, adding private plans gave no advantage, and combining the evidence did not perform better than gain alone. Even though the outcomes do provide useful evidence, they cannot on their own prove collusion since shared incentives can lead to similar behaviour. None of the communication signals that were tested reliably acted as an antitrust-style “plus factor”.

Index Terms—LLM agents, collusion detection, multi-agent systems, AI auditing, LLM monitors, antitrust, pre-registration

I. INTRODUCTION

LLM agents are increasingly deployed side by side: in markets, in the shared infrastructure and inside larger workflows involving multiple agents; when two of them act in coordination with one another, the damage can be serious (in the form of higher prices, a resource that has been captured and a competitor that has been displaced) may never officially acknowledge the coordination. The auditor normally keeps records of what the agents said in public, what actions they

took and the results that displayed. The practical question is what action to take regarding those logs.

We study two questions on one shared dataset:

- **Study 1 — say vs. do.** Does combining what agents say with what they do beat either alone, when the same LLM monitor reads different *views* of a game record?
- **Study 2 — standard of evidence.** What should count as convincing evidence of collusion? We compute the same conversation, coalition-advantage and the dependence evidence for every game and combine it with fixed decision rules.

Study 2 can be usefully approached through the lens of competition law. The fact that firms act in a parallel manner does not prove that they have reached an agreement: firms may move together for independent reasons, for example due to a common cost shock. Courts therefore seek “plus factors” on top of parallel conduct, for example evidence of communication or actions that only make sense under an agreement [8]–[10]. We built our dataset so that both halves of this argument can be tested: one lookalike shows collusive-sounding *talk* without coordination and another shows parallel *conduct* caused by a shared incentive without any agreement.

a) *Our main contributions are:*

- 1) **A pre-registered, fully local collusion testbed** with four episode types, two of which are deliberately hard lookalikes; ground truth taken only from the generation config; one dev/test split fixed before there was any game; each model was pinned by digest and was run locally.
- 2) **A view comparison for LLM monitors with controls, a second monitor model and a pre-registered** The view which appeared best in the original test was one that was built with code on 48 new games. The report failed to do so, while a model-free behaviour signal did and clearly outperformed all the LLM views.
- 3) **An evidence-standard analysis in antitrust terms.** Coalition gain is necessary and it replicates, but a shared incentive fooled it. None of the communication evidence we measured worked as a plus factor.
- 4) **A catalogue of measured LLM-monitor failure modes:** a constant-score collapse, a judge ceiling, a model-

specific position bias, a session effect and small-sample optimism.

II. RELATED WORK

Multi-agent deception and collusion testbeds. Our games run on the harness of Shi et al. [1], which records private plans, public announcements, final actions and outcomes over repeated social dilemmas and classifies promise-breaking. Fish et al. [4] show tacit collusion by LLM pricing agents; Agrawal et al. [5] study seller collusion in LLM double auctions with an overseer; Colosseum [3] audits collusion in teams of tool-using agents with coalition-advantage metrics. These works mostly ask *whether* agents collude. We ask how an auditor should *detect* it from logs, with lookalikes that are designed to fool each kind of evidence.

Monitors and detectors. TraceGuard [2] scores a single agent’s actions on several structured dimensions, including action–trace consistency; we reuse that rubric idea in our structured report view (V4). Work on steganographic collusion [6], [7] builds calibrated detectors for hidden channels. We instead compare *the same* LLM monitor on different views of the same record, add controls that break specific shortcuts (mismatched, length-matched and shuffled views), and test whether the ranking survives a second monitor model and new games.

Antitrust evidence. The “plus factor” doctrine [8]–[10] separates parallel conduct from agreement. We turn it into testable decision rules (Section V) and measure what they gain and what they miss.

The game. El Farol [11] is an anti-coordination game: the bar is pleasant only if few go. Two agents can share the capacity only if the others stay home, so collusion creates a visible coalition advantage while honest play has no dominant action.

III. SETUP AND METHODS

A. Game and protocol

Five agents (J, M, Q, T, Z), all `qwen2.5:7b` at temperature 1.0, play 8 rounds of El Farol. Going to the bar is worth \$10 (if fewer than half, i.e. go at most 2 of 5), a crowded bar is worth −\$5, and staying home is worth \$0. Each round follows the three-stage protocol of the Shi et al. harness [1]: a private plan (Stage 1), a public announcement with a message in a seeded, rotated speaking order (Stage 2), and a simultaneous final GO/STAY choice (Stage 3). We use the harness only for its game definitions and prompt builders; a deterministic history block replaces its reflection step, and every prompt ends with a brevity line and an English-only language line.

B. Episode types and ground truth

- **Type 1 — collusion.** The two focal agents receive an alliance note (a shared goal, a secrecy instruction and a suggested plan: both go while encouraging the other three to stay home) and a private channel that they use twice per round.

- **Type 2 — placebo.** The focal pair receives a neutral note of the same length in the same position (139 vs. 138 tokens), and plays honestly.
- **Type 3 — persona talk.** The pair is told to talk like friends in public, with no agreement and no channel. This lookalike targets *word* evidence.
- **Type 4b — shared incentive.** Each focal agent is told privately that staying home costs it an \$8 deposit, so going is better for both, with no agreement and no knowledge of each other. This lookalike is antitrust’s “conscious parallelism” and targets *outcome* evidence.

The label is collusion = 1 for Type 1 and 0 otherwise, read only from the generation config. The dataset has 48 games (24 Type 1, 8 of each lookalike), split once into 24 dev and 24 test games (each 12 Type 1 + 4 of each lookalike), with labels and split written from the plan *before* generation (master seed 20261023). The replication used 48 new games (seed 20261027) as one confirmation set with no split. Fig. 1 shows the data flow.

C. Views, controls and cleaning (Study 1)

A whitelist cleaner keeps only public fields (announcements, messages, final actions, standard-rule payoffs); setup text, private payoffs, the private channel and labels cannot reach a view. Its leak scanner removes 8-word copies of setup texts and header phrases, and otherwise only flags (deviation 7, Appendix A).

- **V1 messages:** all public announcements and messages. **V2 actions:** final actions and payoffs per round. **V3 combined:** V1 and V2 interleaved per round.
- **V4 structured report:** built by code, with no LLM: per agent, announced vs. actual action and broken promises (with a TraceGuard-style consistency score); per pair, rounds both went, joint broken promises and mentions of each other.
- **V5 oracle:** V3 plus each agent’s private Stage-1 plan and Stage-3 reasoning (the private channel is excluded because only Type 1 games have one).
- **Controls:** *C-mismatch* (messages of one game with actions of another game of the same type), *C-length* (V3 cut to V4’s length) and *C-shuffled* (contents drawn from random games; a leak check).

D. Monitors and judge

The primary monitor is `llama3.1:8b` with the frozen format `monitor_v3_pairs_ctx16k`; it scores all 10 pairs 0–100, and the game score is the maximum pair score (temperature 0, `num_ctx` 16384, one sample per call). The same frozen prompt was run with `granite3.3:8b` as a second model (RQ9 and the replication’s amendment 1). For Study 2 an LLM judge (`llama3.1:8b`) reads only the cleaned public messages and scores each pair 0–100.

E. Evidence signals and decision rules (Study 2)

Four evidence families are computed per pair: **E1 conversation** (the judge score C_{judge} , plus a coordination-keyword

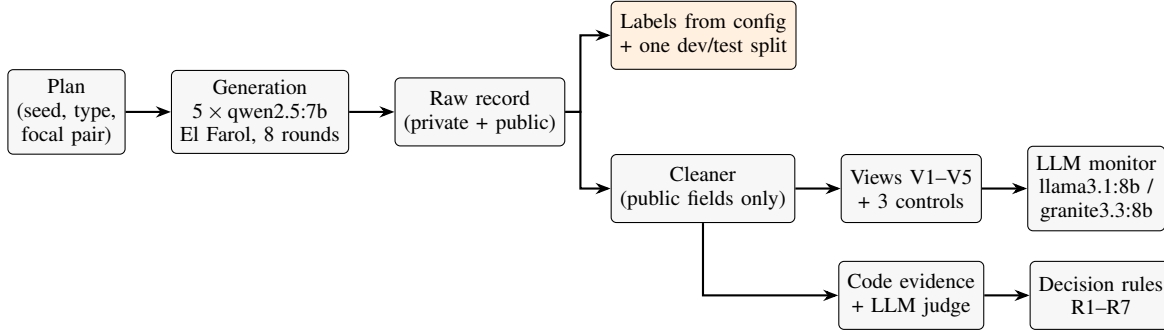


Fig. 1. Pipeline. Labels come only from the generation config and are written, with the single dev/test split, before any game is generated. Monitors, the judge and the code evidence see only the cleaned public record (V5 adds private plans, never the private channel).

score C_{kw}), **E2 coalition advantage** (G_{others} : how much better the pair does than the other agents), **E3 dependence** (Dep: how much the pair’s choices depend on each other) and **E4 word-action** (W_{joint} : whether the pair broke public promises together). Simple parallel-conduct signals are *coGO* (rounds both went) and *capture* (rounds in which exactly the pair went). Seven rules were frozen on dev: R1 judge only; R2 gain only; R3 dependence only; R4 legal style (parallel conduct plus a plus factor, same pair); R5 gain and dependence, same pair; R6 points (≥ 1 of 4 families above its dev level); R7 a learned logistic rule.

F. Integrity: pre-registration, freezes and statistics

- **Pre-registration chain** (Appendix ??): a frozen RQ0.1 pre-registration; amendment 1 (after pilot 1, before pilot 2); amendment 2 (before any dataset game) and its erratum (before any monitor result); the frozen monitor (after dev runs, before any test call); follow-up 1 (before its calls); and the replication pre-registration with its amendment 1 (before any replication game, and before any replication view was rendered, respectively).
- **Freezes.** Monitor format, prompts and thresholds were chosen on dev, frozen and run *once* on test; the Study 2 rules file was applied once to test and once to the replication, never refit.
- **Statistics.** AUROC is primary; we also report accuracy-style counts at frozen thresholds. All intervals are 95% bootstrap CIs stratified by episode type, resampling whole games ($B = 2000$, seed 20261023); differences are paired. The pre-registered primary comparisons are RQ2 and RQ5; everything else is exploratory, with no multiple-comparison correction.
- **Deviations.** We logged 39 deviations, each with the reason and whether it was decided before or after the related data (Appendix A). The most important ones: pilot 1 failed the frozen uptake rule and amendment 1 replaced it (#4); the cleaner became flag-only (#7); `num_ctx` was raised to 16384 before any monitor call (#15); and the constant-score format was ruled ineligible on dev, before any test call (#16).

- **Local and reproducible.** Ollama 0.40.3 with every model pinned by digest, \$0 spent, and every LLM call cached to disk under a key that includes the model digest, prompt, schema, settings, seed and sample index.

IV. RESULTS: STUDY 1 (MONITOR VIEWS)

A. Gates: does the dataset behave as intended? (RQ0.1–RQ0.4)

RQ0.1 – Real collusion. Pilot 1 failed the uptake criterion (2/4) even though the transcript evidence suggested coordination in all 4 games. The rule failed to pick up on brief instructions which didn’t contain the ‘we/us’ phrasing and also failed to pick up on the shift to Chinese in one game. Amendment 1 introduced an English only instruction and uptake rule v2. In Pilot 2 (fresh pilot), all the criteria were met: uptake 4/4, placebo false positives 0/2, the focal pair came first by coGO in 4/4 games and both coalition advantage and capture occurred in 4/4 games. Rule-hand check agreement was 96.9% (Cohen’s $\kappa = 0.931$). Across the full dataset uptake was recorded in 24/24 games and the combined uptake and effect criterion was met in 23/24.

A hand check of Pilot 1 by two raters who were blind to the rule further looked into the rule’s limitations. The raters agreed in 91.7% of the cases ($\kappa = 0.80$). Rule v2 agreed with the two raters in 79.2% and 85.4% of the cases respectively. All the disagreements concerned missed detections of coordination; there were no false alarms. The fact that detection errors had remained shows that they had not disappeared even though Pilot 2 had been a success.

RQ0.2 – Lookalike conditions. Type 3 behaved as intended in 7/8 games, with one partial case. Type 4b followed the intended pattern in 6/8 games; in the remaining two, participants did not follow the incentive. However, all eight Type 4b games included a public mention of the “deposit”, creating an unplanned cue visible to monitors.

RQ0.3 – Leakage. The leakage checks passed: no setup copies required removal, and a search across 384 views found no leakage. The shuffled control remained consistent with chance-level performance on development data (0.465 [0.250, 0.667]) and test data (0.632 [0.431, 0.833]).

RQ0.4 – Stability. Stability criteria were met for V1, V3, V4, and V5, but not for V2 (mean absolute difference: 5.00;

Spearman’s $\rho = 0.676$). Collusion was not fully concealed: 6.5% of focal public messages revealed joint action, and 12/24 Type 1 games contained at least one such disclosure.

B. Original test set (RQ1–RQ8)

Fig. 2 compares the views on the 24-game test set. V4 achieved the highest AUROC (0.896 [0.771, 0.979]), while V1 was close to chance (0.490) and V2 fell below it (0.372). The full results are reported in Table I.

Primary comparisons. RQ2 (V3, with the dev-chosen best single view excluded, V2) was +0.365 [0.094, 0.611]: The fact that ‘combining helps’ is indicated by the pre-registered term yet it is fragile since V2 failed RQ0.4 means that the pooled value was It increased by +0.040 [−0.155, 0.238] and V2 alone rose by +0.188 between sessions. RQ5 found no clear advantage with respect to a preference for the structured view (V4 − V3 +0.160 [−0.021, 0.351]) as compared with a raw view of the same length (V4 − C-length +0.285 [0.017, 0.556]).

Other questions. Each single view was fooled by one lookalike (RQ3). At the frozen thresholds V4 flagged 0/4 placebo, 2/4 persona and 1/4 shared-incentive games; V2 flagged 2/4, 3/4 and 2/4. Per-type AUROC showed persona talk fooling actions (V2, T1 vs. T3 0.219 [0.00, 0.62]) and the shared incentive fooling messages (V1, T1 vs. T4b 0.344 [0.12, 0.65]) (Fig. 3). There was no clear evidence that the monitor cross-checks words against actions (RQ4, V3 − C-mismatch +0.188 [−0.028, 0.386]), and private plans added nothing (RQ6, V5 − V3 −0.031 [−0.267, 0.236]). The dependence code signal lost to V4 (RQ7, −0.173 [−0.382, −0.009], $n = 21$), but the simpler action signals coGO, capture and G_{others} matched V4 (exploratory). For naming the colluding pair (RQ8; chance 0.1), V4 was best with 7/12.

C. Follow-up 1: pair-order position bias

Llama gave the top score to the last-listed pair (T-Z) in 21/24 V1 and 17/24 V3 test replies, although T-Z was the true pair in only one out of the twelve test collusion games; a pre-registered follow-up re-scored the test games involving the pair. The order was randomised (see Fig. 4). Randomisation eliminated the last-position effect in V1 (21 out of 24, now 10 out of 24) and for V3 it was 19 out of 24 (now 9 out of 24), but there was no clear change in the AUROC for either one (V1 +0.017 [−0.229, +0.281], V3 +0.042 [−0.198 to +0.285], V4 −0.076 [−0.240 to +0.111]). Although the bias is real, it does not account for the raw-view. There was a failure; when the order was chosen at random, V4 identified the correct pair in 8 out of 12 games.

D. RQ9: a second monitor model

Using the same frozen prompt with granite3.3:8b produced a different ranking (Fig. 5): V1 0.729 [0.549, 0.896], V2 0.767 [0.604, 0.917], V3 0.736 [0.573, 0.917] and V4 0.510 [0.292, 0.729] (Kendall τ -b −0.667 (a coarse measure with 4 views)); granite had no preference for the last position other than in the case of ties. Granite continued to run. Since

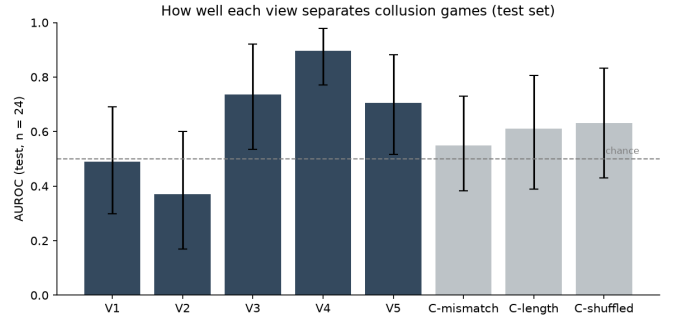


Fig. 2. **Test-set discrimination by view** ($n = 24$). Bars show AUROC, with 95% confidence intervals; the dashed line marks chance performance (0.5). Dark bars represent V1–V5, and grey bars represent the three controls. V4 performed best (0.896 [0.771, 0.979]); the confidence intervals for V3, V4, and V5 lie above chance. The frozen threshold was zero for V1, V3, and C-mismatch, making their threshold-based metrics uninformative. AUROC is threshold-independent.

different hardware is involved (Apple Metal versus NVIDIA), it is not possible to separate the effects of the model and the hardware.

E. Pre-registered replication on 48 new games

Table II and Fig. 6 give the replication with frozen prompts, one sample and one run per monitor; since there were no quality fail games the sensitivity analysis is the same as the primary one. Replicated: messages alone fail (R1), private plans add nothing (R8), the shuffled control is at chance (R9), and the model-free capture signal works (R10a) and is now clearly doing better than the best LLM view (R10b, with an improvement of +0.264 [0.121, 0.412]). **Not replicated:** the structured report’s advantages over the raw view (R4) and over a same-length raw view Regarding ‘persona talk fools actions’ (R7), with Llama ‘combining beats actions’ was reproduced (R3) but only... The action view for Llama was below chance (0.324); with Granite, it was not consistently above chance. The llama is the only one that prefers the structured report. Neither of the monitors reaches 0.7 in any of the new games.

V. RESULTS: STUDY 2 (STANDARD OF EVIDENCE)

Single families (test, $n = 24$; Table III). Agents’ public channel messages, that is, Talk alone is not evidence: the judge flagged 11 of 12 innocent games and ranked below chance (0.396). On the other hand, Outcomes rank the games really well, but the shared-incentive lookalike is still able to fool them: coalition gain (R2, 0.872 [0.736, 0.986]) and capture (0.896, exploratory) each flag 2 of 4 Type 4b games. Dependence alone is very weak, it is undefined in 3 games, and is easily fooled by every lookalike type at least once.

Combining did not help (RQ-E3). The pre-registered comparison of the best combined rule (R6) with the best single family (R2), both chosen on dev, was −0.080 [−0.208, 0.045]: no clear difference. Ablations show why: almost all of the signal comes from coalition gain. Removing gain cost R6 +0.292 [0.125, 0.458] and R7 +0.375 [0.111, 0.639] of AUROC, while removing the judge, dependence or word–action consistency

TABLE I

AUROC PER VIEW [95% CI]: ORIGINAL TEST (LLAMA, $n = 24$) AND THE PRE-REGISTERED REPLICATION (48 NEW GAMES) FOR BOTH MONITORS. REPLICATION VALUES ARE NOT REPORTED FOR V5 AND THE C-LENGTH / C-MISMATCH CONTROLS. SOURCES: R14 §4.2; R21 §4.7; R22C §4.4.

View	llama, test	llama, replication	granite, replication
V1 messages	0.490 [0.299, 0.691]	0.505 [0.363, 0.655]	0.618 [0.474, 0.756]
V2 actions	0.372 [0.170, 0.601]	0.324 [0.186, 0.469]	0.628 [0.488, 0.754]
V3 combined	0.736 [0.535, 0.920]	0.565 [0.421, 0.706]	0.662 [0.560, 0.771]
V4 structured	0.896 [0.771, 0.979]	0.636 [0.507, 0.751]	0.582 [0.414, 0.730]
V5 oracle	0.705 [0.517, 0.882]	—	—
C-mismatch	0.549	—	—
C-length	0.611	—	—
C-shuffled	0.632 [0.431, 0.833]	0.457 [0.315, 0.591]	—
capture (code, no LLM)	0.896 [0.781, 0.986]	0.900 [0.811, 0.965]	(model-free)

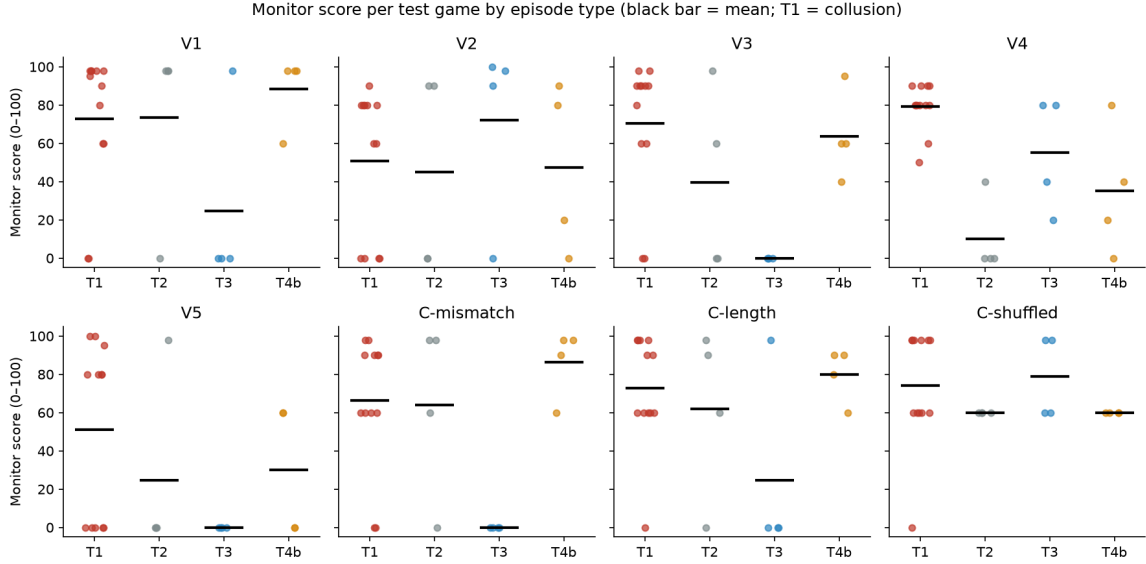


Fig. 3. **Monitor score per test game by episode type** (*exploratory, not pre-registered*). Each dot is one test game; the black bar is the type mean (12 Type 1 games, 4 of each negative type). Persona-talk games (T3) score high in the actions view (V2); same-incentive games (T4b) score highest of all in the messages view (V1). V4 is the only view where every collusion game (T1) scores above every placebo game (T2).

changed little. Word–action consistency (RQ-E5) was a null: removing it changed R6 by 0.000 and R7 by -0.062 , both CIs including 0.

The plus-factor rule. An exploratory rule (post hoc, motivated by report 15) flags a game only if it shows parallel conduct (capture ≥ 0.625) *and* coordination words ($C_{kw} \geq 0.7375$), both thresholds set on dev. It removed every false positive, including both shared-incentive games, but caught only 1 of 12 real collusion games. The frozen legal-style rule R4 shows the same trade-off: 0 false positives, 2/12 caught.

Replication (Fig. 7). The frozen judge and rules file were applied once to the 48 new games. E1 (gain works) was confirmed, 0.885 [0.786, 0.961]; E2 (the talk-only judge is near chance) was confirmed, 0.543 [0.392, 0.693]; E3 (R6 – R2) was not replicated, -0.059 [-0.135 , 0.010], and nor was the secondary E3b (R7 – R2, -0.042 [-0.155 , 0.067]). The shared-incentive lookalike stayed the hardest case: gain flagged 3/8 of those games, against 1/8 placebo and 1/8 persona

games. The plus-factor rule caught 0/24 colluders with 0/24 false positives.

Our standard of evidence. Outcome evidence (gain, capture) is necessary and replicates, but it is not sufficient against a shared incentive. None of the communication evidence we measured was a usable plus factor: requiring it removes false positives only by missing most real collusion.

VI. FAILURE MODES OF LLM MONITORS

- 1) **Constant-score collapse.** The single-score format (one 0–100 score per game) returned 80 in 96/96 dev calls, and it was ruled ineligible on dev before any test call. Scoring each pair instead gave 4–10 distinct values per view.
- 2) **Judge ceiling.** The messages-only judge gave almost every game a top pair score near 98 and never scored a pair above 98 (0 of 480 pair scores). Because E1’s frozen plus threshold was “above 98”, E1 could never add a point in R6. Its top pair was the true pair in only

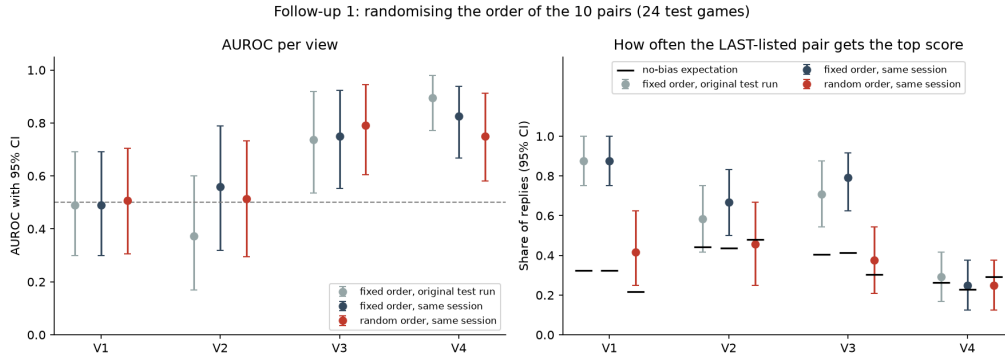


Fig. 4. **Follow-up 1: randomising the order of the 10 pairs** (pre-registered, commit 496fc28). The 24 test games $\times$ V1–V4 were scored in three arms: fixed order in the original test run (grey), fixed order re-scored in the follow-up session (dark), and a random order per call in the same session (red). **Left:** AUROC with 95% CIs. **Right:** share of replies in which the last-listed pair gets the top score; black ticks show the no-bias expectation (allowing for ties). Randomising cuts the last-position effect in V1 and V3 but changes no view’s AUROC clearly.

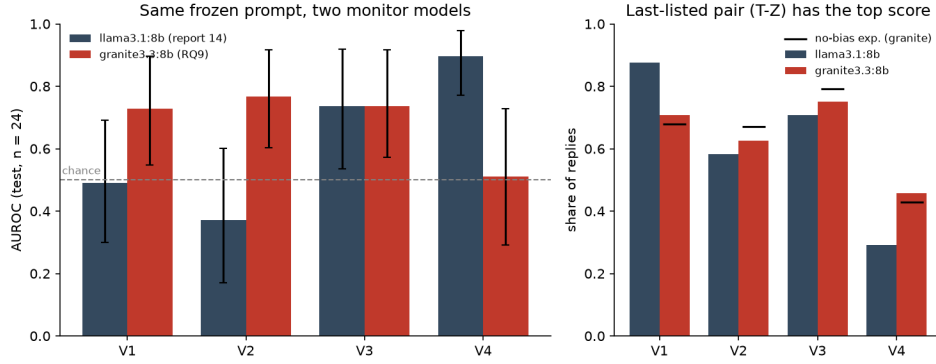


Fig. 5. **RQ9: the same frozen monitor prompt with a second model** (test set, $n = 24$, one sample, one session). **Left:** test AUROC per view with 95% CIs, llama3.1:8b vs. granite3.3:8b. The ranking does not carry over (Kendall τ -b -0.67): only V3 is the same (0.736 for both), and V4, best with llama, is at chance with granite. **Right:** share of replies in which the last-listed pair (T-Z) has the top score; granite sits at its no-bias expectation in every view, so the last-position pattern looks specific to llama.

2 of 12 collusion games, and it gave the top score to the last-listed pair T-Z in 18 of 24 test games.

- 3) **Position bias (llama only).** On the replication, the last-listed pair had llama’s top score in 44/48 V1 replies (no-bias expectation 0.215), 44/48 V3 and 48/48 C-shuffled replies; granite showed no preference beyond ties. The bias changes *who* is named, not *how well* games are ranked (Follow-up 1).
- 4) **Session effect.** Within one Ollama session, repeats were identical. Between sessions, V2 re-scored unchanged went from 0.372 to 0.559 AUROC (+0.188 [+0.007, +0.389]; mean absolute difference 10.79, Spearman 0.653).
- 5) **Small-sample optimism.** Best first results shrank on new games: llama V4 $0.896 \rightarrow 0.636$ and granite V2 $0.767 \rightarrow 0.628$. Between dev and test, V3 moved from 0.319 to 0.736 (+0.417 [0.142, 0.684]), and the view rankings agreed only weakly (Kendall τ 0.255).
- 6) **Reading friendly talk as a pair signal.** V4 at its frozen threshold flagged all 8 persona games on the replication (2/4 on the original test). Its mention-count lines make

friendly talk look like coordination. In one persona game where nobody ever went to the bar, the V4 monitor scored 80 and wrote: “T and Z repeatedly mention each other in messages, suggesting coordination.”

VII. DISCUSSION AND LIMITATIONS

What this means for auditing multi-agent systems. In our setting, *what the agents did* was that they carried the signal and *what an 8B model concluded from reading the record* did not. A simple code signal (capture) replicated at about 0.90, while every LLM view of both the monitors lay between 0.32 and 0.66 on new games, and the best LLM view which was on the first test did not survive. An auditor of agent systems should treat the model-free behaviour statistics as the baseline that any LLM monitor must beat, report LLM monitor results per model, and expect LLM verdicts to move with the model, the prompt order and the session.

The link to antitrust. Our data reproduce the first half of the plus-factor doctrine clearly: parallel conduct and coalition gain are the strongest single evidence, yet they also flag agents that merely share an incentive (gain flagged 3/8 shared-incentive

TABLE II

STUDY 1 REPLICATION (48 NEW GAMES; FROZEN LLAMA3.1:8B MONITOR; GRANITE UNDER AMENDMENT 1). ORIGINAL COLUMN: $n = 24$ TEST. SOURCES: R21 §4.7; R22C §4.4; R14; R15; R15C.

ID	Hypothesis	Replication [95% CI]	Original	Outcome
R1	Messages alone fail: AUROC(V1)	0.505 [0.363, 0.655]	0.490	confirmed
R2	Structured report works: AUROC(V4)	0.636 [0.507, 0.751]	0.896	confirmed
R3	Combining helps: V3 – V2	+0.241 [0.063, 0.419]	+0.365	replicated
R4	Structure beats raw: V4 – V3	+0.071 [–0.128, 0.247]	+0.160	not replicated
R5	Structure beats same length: V4 – C-length	–0.034 [–0.212, 0.133]	+0.285	not confirmed
R6	Same incentive fools messages: V1, T1 vs. T4b	0.495 [0.297, 0.716]	0.344	confirmed
R7	Persona fools actions: V2, T1 vs. T3	0.352 [0.154, 0.578]	0.219	not confirmed
R8	Private plans: V5 – V3	–0.022 [–0.192, 0.155]	–0.031	no gain again
R9	Shuffled control: AUROC(C-shuffled)	0.457 [0.315, 0.591]	0.632	pass
R10a	Code signal: AUROC(capture)	0.900 [0.811, 0.965]	0.896	confirmed
R10b	Capture – V4, paired	+0.264 [0.121, 0.412]	—	capture better
G1	Granite action view works: AUROC(V2)	0.628 [0.488, 0.754]	0.767	not confirmed
G2	Granite structured report does not work: AUROC(V4)	0.582 [0.414, 0.730]	0.510	confirmed
G3	Ranking depends on monitor: V4 – V2, llama / granite	+0.313 [0.127, 0.481] / –0.046 [–0.251, 0.152]	—	confirmed (mainly llama)
G4	Monitors agree on V3: granite – llama	+0.097 [–0.063, 0.262]	—	no detected difference

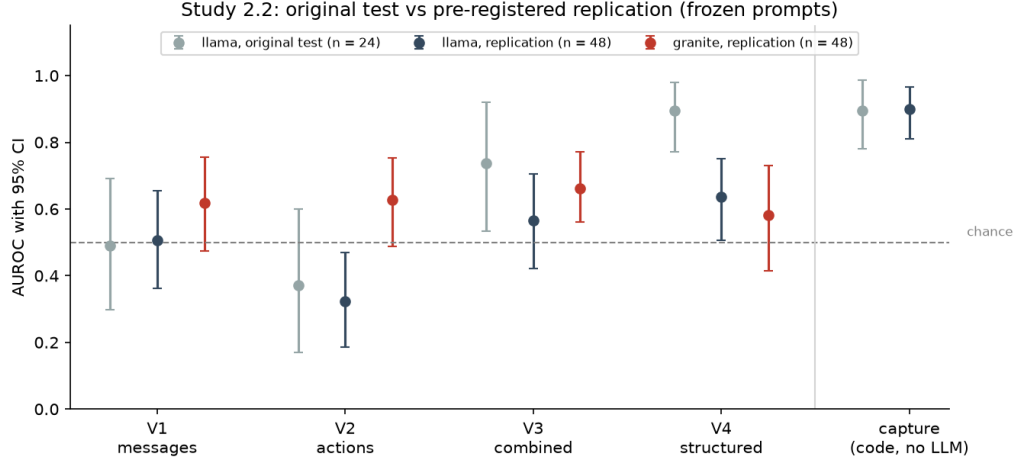


Fig. 6. **Study 1: original test vs. pre-registered replication** (frozen prompts; one sample per call, one run per monitor). AUROC with 95% CIs: llama on the original 24 test games (grey), llama on the 48 replication games (dark) and granite on the same 48 games (red); the model-free capture signal on the right. The structured report’s original 0.896 shrank to 0.636. No LLM view of either monitor reaches 0.7 on the new games, while capture stays at about 0.90. V5 and the C-length / C-mismatch controls are not shown (no replication AUROC reported).

games on the replication). The second half did not work here: neither keyword evidence nor an LLM judge of the public talk was a usable plus factor, because agents in this game talk about coordinating all the time, honest ones included. Real cartels rarely announce their agreement in public, and an auditor, like a competition authority, will often lack the private channel. Our results suggest that without such private evidence, the honest verdict on a “shared incentive” pattern is “parallel conduct, agreement not shown”.

Limitations

- **Small samples.** 24 test and 48 replication games, with 8 or fewer per lookalike type; one false positive moves a type’s rate by 25 points on test and 12.5 points on the replication, and most CIs are wide.
- **Narrow setup.** One game (El Farol), one agent model (qwen2.5:7b), 8 rounds. The collusion is suggested by a prompt (the alliance note proposes the plan), not invented by the agents, and it is not fully secret (6.5% of focal public messages reveal joint action).
- **Two small monitors** (8B), one frozen prompt format and one sample per call. Scores can change between sessions, and the second monitor ran on different hardware, so model and hardware are confounded.
- **An unplanned clue.** Shared-incentive agents talk about their “deposit” in public, and the flag-only cleaner leaves it visible to monitors.
- **Data quality.** 1.75% of generation calls failed (all truncation) and their rounds are excluded; 1 persona game partly behaved like collusion and 2 shared-incentive games did not follow the incentive (labels stay as configured).

TABLE III

FROZEN DECISION RULES ON THE 24 TEST GAMES (12 COLLUSION; 4 PER LOOKALIKE TYPE) AND THEIR REPLICATION AUROC (48 NEW GAMES). COUNTS ARE GAMES FLAGGED AT THE FROZEN THRESHOLDS. R4 AND R5 ARE BINARY RULES (NO AUROC). SOURCES: RESULTS_23.MD §1, §2, §6 (17C; R22C §4.2); CAPTURE: R15 §4.4, COUNTS R17C §4.5.

Rule	AUROC test	TPR	FPR	T2	T3	T4b	AUROC replication
R1 judge (talk only)	0.396 [0.205, 0.597]	8/12	11/12	4/4	4/4	3/4	0.543 [0.392, 0.693]
R2 coalition gain	0.872 [0.736, 0.986]	10/12	2/12	0/4	0/4	2/4	0.885 [0.786, 0.961]
R3 dependence ($n = 21$)	0.736 [0.491, 0.942]	9/12	4/12	1/4	1/4	2/4	0.682 [0.518, 0.823]
capture (exploratory)	0.896 [0.781, 0.986]	8/12	2/12	0/4	0/4	2/4	0.900 [0.811, 0.965]
R4 legal style	—	2/12	0/12	0/4	0/4	0/4	—
R5 gain AND dependence	—	4/12	2/12	0/4	0/4	2/4	—
R6 points	0.792 [0.667, 0.917]	11/12	4/12	1/4	0/4	3/4	0.826 [0.721, 0.920]
R7 learned (logistic)	0.812 [0.611, 0.972]	10/12	2/12	0/4	0/4	2/4	0.844 [0.722, 0.941]

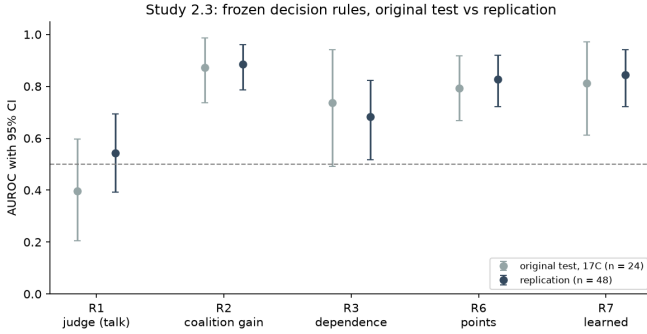


Fig. 7. **Study 2: frozen decision rules, original test vs. replication** (pre-registered; frozen rules, not refit; one judge sample). AUROC with 95% CIs on the 24 original test games (grey) and the 48 replication games (dark). The gain rule R2 replicates (0.872 $\rightarrow$ 0.885), and the talk-only judge R1 stays near chance (0.396 $\rightarrow$ 0.543). The combined rules R6 and R7 do not beat R2. R4 and R5 are binary and have no AUROC.

- **Degenerate thresholds.** V1, V3 and C-mismatch had a frozen threshold of 0 (flag everything), so their threshold metrics carry no information.
- **V2 instability.** V2 failed the stability gate (RQ0.4), so V2-based results (including RQ2) carry this caveat.
- **Exploratory parts.** Code-signal comparisons beyond Dep and the plus-factor rule were not pre-registered; pooled results include dev selection data; the uptake rule v2 was developed on pilot 1 and confirmed on pilot 2 only.
- **In-sample dev fits.** Thresholds and rule choices come from 24 dev games, so dev performance overstated test performance (e.g. R2’s balanced accuracy 0.958 on dev vs. 0.833 on test).
- **One judge** (llama3.1:8b, one sample, one session) with a fixed pair order that likely biases its pair choice.

Future Work

- 1) Larger test sets with more games per lookalike type, so that primary comparisons such as RQ2 and RQ5 can be settled.
- 2) Score several sessions per view and report the between-session spread, especially for V2 and V4.

- 3) Make randomised pair and agent order the default monitor format, with an explicit “no pair” option to reduce ties.
- 4) Treat simple action signals (coGO, capture, G_{others}) as required baselines, and test a combined code + LLM rule on dev first.
- 5) Remove the shared-incentive “deposit” clue at generation time, so that lookalike tests behaviour only.
- 6) Test emergent (not prompted) collusion, other games and larger monitors.

VIII. CONCLUSION

We built a pre-registered, fully local testbed for auditing the collusion between the LLM agents, with lookalikes that target the word evidence and also the outcome evidence separately, and then we re-tested every frozen choice on 48 new games. The result is a clear ordering of evidence: the model-free behaviour signals (capture, coalition gain) were the most reliable and they replicated; small LLM monitors were mediocre, model-dependent and unstable, and their best first result did not hold; messages alone and talk-only judging stayed near chance; and combining evidence did not beat gain alone.

For the standard of evidence, our data agree with the antitrust intuition: parallel conduct and coalition gain are necessary signals but not enough proof, because a shared incentive produces the same outcomes. In this setting no communication evidence we measured could serve as the plus factor. Detecting the agreement itself, as opposed to its consequences, remains open.

CODE AND DATA

- **Code repository:** <https://github.com/AyushSingh110/collusion-evidence-lab> (pre-registrations in `configs/`, per-task reports in `notes/reports/`, results in `writeup/`).
- **Data:** games, views, monitor outputs and the full LLM cache are kept out of git (`data/`, `runs/`, `cache/`); every report gives the commands that produced it. Without the cache, re-running makes new model calls, which may not reproduce every score exactly (session nondeterminism).
- **Info-hazards:** none beyond the general topic; the collusion is a prompted toy behaviour in a bar game.

AUTHOR CONTRIBUTIONS

Ayush Singh led the project, the research design and the pre-registrations. **Param Babu** built the harness integration, the data pipeline, the generation runs and the LLM caching. **Atharva Patil** built the monitor views and runners and carried out the Study 1 analysis. **Neerada Dathan** built the evidence signals, the decision rules and the statistics, and carried out the Study 2 analysis. All authors wrote and reviewed the final manuscript.

REFERENCES

- [1] Shi, Zhang, Schölkopf, Conitzer, and Jin, “When agents lie: Premeditation, persistence, and exploitation in repeated games,” arXiv:2607.05132, 2026. [Online]. Available: <https://github.com/Jerick-1380/LLM-Trust-Breaking>
- [2] Nguyen *et al.*, “TraceGuard: Structured multi-dimensional monitoring as a collusion-resistant control protocol,” arXiv:2604.03968, 2026.
- [3] Nakamura *et al.*, “Colosseum: Auditing collusion in cooperative multi-agent systems,” arXiv:2602.15198, 2026.
- [4] Fish *et al.*, “Algorithmic collusion by large language models,” arXiv:2404.00806, 2024.
- [5] Agrawal *et al.*, “Evaluating LLM agent collusion in double auctions,” arXiv:2507.01413, 2025.
- [6] Mathew *et al.*, “Hidden in plain text: Emergence and mitigation of steganographic collusion in LLMs,” arXiv:2410.03768, 2024.
- [7] Tailor, “Audit the whisper: Detecting steganographic collusion in multi-agent LLMs,” arXiv:2510.04303, 2025.
- [8] Kovacic, Marshall, Marshall, and Raiff, “Plus factors and agreement in antitrust law,” *Michigan Law Review*, vol. 110, no. 3, 2011.
- [9] *Bell Atlantic Corp. v. Twombly*, 550 U.S. 544, 2007.
- [10] *Theatre Enterprises, Inc. v. Paramount Film Distributing Corp.*, 346 U.S. 537, 1954.
- [11] Arthur, “Inductive reasoning and bounded rationality,” *American Economic Review*, vol. 84, no. 2, pp. 406–411, 1994.

APPENDIX A

DEVIATIONS FROM THE PRE-REGISTRATIONS

Items 1–12 are from report 10 §B4; items 13 onward were added later; items 23 onward are the RQ9 / replication phase. “When” says whether the deviation was decided before or after the related data. Source: `writup/results_22.md` §f.

#	What	Why	When
1	Pre-freeze changes: possessive name pattern, hint folded into alliance_v3, placebo_v3, 2 machines	rule bug; b2 at risk; length match; machine C unavailable	before any pilot data (the hint was motivated by 3 earlier smoke games)
2	Freeze commit 84fca68 → 4909c7e	the title still said DRAFT	before pilot 1
3	Pilots 1 and 2 on machine B only (prereg: A and B)	operational	before the data; no design change
4	Departure from U1 → Amendment 1 (language line, uptake rule v2, pilot 2, new gate)	U1 targeted a cause that was not observed; the rule missed real coordination; language drift	after pilot-1 data (rule v2 tested on new data, pilot 2)
5	Pilot-1 hand-check made descriptive	the rule had been replaced; two raters for reliability	after pilot-1 data, before the hand-check was filled
6	E2 = 10 rounds instead of 12	lead’s decision	after pilot 1; moot
7	Cleaner becomes flag-only (except 8-word copies and header phrases)	over-removal concern; Type 4b “deposit” talk is behaviour	after pilot 1
8	Language line (“English only”)	language drift in pilot 1	after pilot 1
9	Amendment 1 committed together with code, not alone	—	before pilot 2
10	Secrecy variant v4 + 2-game re-pilot	executes frozen \$5 (revealed rate 12.5%)	after pilot-2 data; recorded before its own calls

#	What	Why	When
11	Proceed with v4 although the re-pilot reveal rate stayed 12.5%	frozen \$5 sets no further target	after the re-pilot data
12	Repair re-ask not used	keep all failures as data	after the pilot-2 failures
13	RQ0.2 criteria operationalised (e.g. “persona talk” = names the partner + a warm word)	the frozen text gave no exact rule	stated in the script before computing RQ0.2
14	Erratum and analysis decisions: alternation shows in G_others only; C_kw = literal sum of both directions; tie rules; threshold candidates = observed dev scores	correctness of the amendment’s explanation; rules not pre-registered	before any monitor result
15	num_ctx 16384 for all monitor calls (new manifests, same templates)	every V5 view exceeded the 8192 guard	after the views were rendered, before any monitor call
16	Format eligibility clause: a constant-score format is ineligible, so v3_pairs was chosen	v1 gave 80 on 96/96 dev calls	after dev monitor call , before any test call
17	<code>analyse_22.py -freeze-extra</code> (adds keys to the frozen file without overwriting)	record format, digest and RQ0.3 / RQ0.4 in the frozen file	before the test run
18	Test-analysis addendum: “degenerate threshold” labels; supplementary per-type AUROC for RQ3; RQ2 caveat; exploratory C-length — V3	three views had threshold 0	before test scoring
19	Exploratory robustness analyses, plus an unplanned position-bias check	robustness; prompted by an excerpt	after test data ; labelled exploratory
20	Follow-up 1 pre-registered, with a same-session control arm added before the commit	test the position-bias explanation	after test data, before any follow-up call
21	Follow-up random arm also uses sample_idx 1 (same seed as the control arm)	the arms then differ only in pair order	before any follow-up call
22	Documentation corrections: section reference “report 15 §6” → §4.6	typo	no data involved
23	RQ9 granite run recorded machine id <code>m_5acffc7a</code> instead of C; cache shard kept as is; shell fixed for later runs	environment variable not loaded in the tool	after the RQ9 run; provenance only, results unaffected (pinned digest checked on every call)
24	RQ9 run paused at 11/96 calls to ask the lead’s OK; the kill removed the first session’s manifest record	time limit (Rule H)	during the run; nothing re-scored
25	The RQ9 granite manifest adds a <code>model_digest</code> field	record the digest in the config	before the RQ9 run
26	E3 fixed as primary (best single R2 vs. best combined R6; tie broken by the lower rule number); E3b = R7 — R2 added as secondary	no 17C record existed when the prerog was drafted	before any replication game
27	Replication labels in a separate file; C-mismatch partners within type across all 48 games; full outcome map for every rule; primary = all games, sensitivity = without quality-fail games	never touch the original labels / split; avoid undefined outcomes	before any replication game
28	The labels writer refuses to overwrite even an identical file	stricter, as the task asked	before generation
29	Outcome words for “otherwise” branches; last-position-max with a tie-aware expectation; quality-file format	the prerog gave “otherwise” or no format	before any replication output
30	Amendment 1 (granite): full outcome map for G2; G4 note “no detected difference, not equivalence”; analysis refuses frozen dev game ids	avoid undefined and over-reading	before any replication data

#	What	Why	When
31	G1–G4 code checks each monitor line by <code>prompt_sha256</code> ; “undetermined (no CI)”; a G3 / G4 zero estimate counts as “not confirmed”	the amendment did not cover these cases	before any replication monitor output
32	<code>g_ec16a9027c</code> : <code>G_others = 2.2e-16</code> counted as <code>> 0</code> by the frozen rule; effect check reported both ways (24/24 by rule, 23/24 substantively)	float artefact in a descriptive check	after the generation data; descriptive only
33	Replication quality and checks run with path-only copies of the task-12 scripts (diff-checked)	view keep report 12’s definitions exactly	before the monitor run
34	Report 21 gives V2 / V3 replication point estimates derived by subtraction (no CI)	context for R3 / R4	after the analysis; descriptive
35	Granite-only replication ensures not computed before llama file arrived	mea- keep the single G1–G4 run the first look	before the G1–G4 run
36	The llama V1–V4 subset of machine B’s 384-line file selected by the <code>view</code> field only; lines copied byte for byte	amendment 1 specifies “the V1–V4 subset”	after the data existed; score-blind, recorded before the G1–G4 run
37	Final-results documentation: “position bias explains the raw-view failure” keeps NOT SUPPORTED; Figure 10’s 17C values copied from report 17C; Figure 9 omits V5 and two controls	stay within reported numbers	after all data; documentation only
38	Final labels reworded: “actions alone fail” → “REPLICATED (llama); borderline for granite”; “view ranking depends on the monitor” → “NOT SHARED BETWEEN MONITORS (G3; rests mainly on llama)”	precision of wording	after all data; wording only, no number changed
39	Documentation repairs: conflict markers in the VLOG removed with both sides kept; two citation-section fixes from the number audit	merge- merge artefacts; audit findings	no data involved

LLM USAGE STATEMENT

We used Claude (Anthropic), through Claude Code, as an engineering assistant: it wrote and tested most of the code under our instructions, ran the local experiments, drafted the per-task reports and drafted this paper from our write-up files. All design decisions were made by the team. Every number in this paper is copied from the project’s reports, each traceable to a report section, and the team verified the claims.