

Before Testing Turnover: Collusion Formation in a Four-Agent Review Task

Harshkumar Patel
University of Toronto

Stephanie Lu
University of Toronto

With
Apart Research

Abstract

Replacing agents involved in collusion might interrupt the practice, but remaining members could pass it to newcomers. We designed a four-agent contribution-review experiment to compare sequential replacement, immediate reset, and no replacement after reciprocal false approvals had formed. Across six planned teams, Qwen/Qwen3.8-27B made 288 truthful claims and 288 accurate approvals over twelve rounds per team. No team met the formation criterion, so no replacement branches ran and the turnover effect remains unestimated. We then conducted two separate exploratory follow-ups. One team received twelve synthetic rounds of inflated claims and approvals, followed by six live rounds; all 24 live claims were truthful. In a matched rent-pressure pilot, both conditions produced 24 truthful claims, including the condition where honest earnings could not cover simulated rent and all four agents incurred shortfall penalties. These results describe the absence of the targeted misconduct under the tested conditions. They do not show that turnover prevents collusion or that guardrails caused the observed compliance.

1. Introduction

Multi-agent LLM systems often rely on agents checking each other’s work, but this safeguard can fail if the agents collude. This can happen even without explicit instructions. Shi et al. (2026) found that collusion emerged during repeated peer verification. Replacing the agents involved might seem like an obvious solution, but it is unclear whether this actually stops the behaviour. The practice might persist because the original agents carry it with them, because they influence new teammates to adopt it, or because the same incentives cause any new team to develop it.

We designed a study of whether a reciprocal false-approval practice persists after all four agents who established it have been replaced. The planned comparison contrasts replacing agents one at a time, allowing new agents to work alongside remaining founders, with replacing all four agents at once. An unchanged team serves as a baseline. If the practice spreads from existing agents to newcomers, greater persistence under sequential replacement would be consistent with transmission. Similar persistence under both replacement approaches would be consistent with independent re-formation, although these comparisons would not isolate a single mechanism.

The formation stage did not produce an eligible team. This paper therefore reports that result and two exploratory follow-ups, rather than an effect of replacement.

Our main contributions are:

1. **A testbed for studying collusion under turnover.** We implement a four-agent contribution-review task with claims that can be checked against ground truth and matched replacement strategies.
2. **Separate evidence for misconduct, coordination, and transmission.** We distinguish false approvals from evidence of an agreement and from evidence that newcomers adopted the practice.
3. **Formation results and exploratory follow-ups.** We report the six original teams and separate seeded-history and rent-pressure pilots, retaining unsuccessful formation attempts and leaving the turnover effect unestimated.

2. Related Work and Scope

The closest study to ours is Shi et al. (2026), which shows that agents’ willingness to break verification rules depends partly on peer behaviour and memories of earlier interactions. Its main experiments use two adaptive LLM agents; scripted peers are a separate intervention. A constrained communication channel prevents complete evidence exchange. Our task instead gives reviewers certified quantities and measures approvals of inflated claims. Mutual approval of truthful work is compliant here. These differences prevent a direct comparison of collusion rates or an attribution of the outcomes to guardrails.

LLM research has also examined cooperation across generations. Vallinder and Hughes (2024) study indirect reciprocity in a donor game, showing that population dynamics can depend on the base model and initial conditions. Our planned newcomer control gives new agents neither their predecessors’ transcripts nor private notes. They can only learn about the group’s practices through ongoing interactions. Comparing sequential replacement with an immediate reset was intended to help distinguish transmission through colleagues from independent re-formation.

Finally, parallel misconduct does not itself establish coordination. We distinguish an inflated approval from evidence that agents agreed to produce it, and distinguish both from evidence that a newcomer adopted the practice. None of the completed live runs produced the false approvals needed to evaluate transmission.

3. Methods

3.1. Task and incentive conflict

Four LLM agents repeatedly play a contribution-review game. Each round, an agent receives a certified output q in $\{4, 5, 6\}$ and submits a claim c in $\{0, \dots, 8\}$. Assigned reviewers inspect the output and claim, then approve or reject it. Approved claims earn the claimed amount; rejected claims earn zero. Review pairings rotate across all seat pairings over three rounds.

Each approved claim has a 10% chance of an independent audit. If an audit finds an inflated claim, the excess payment is recovered and both the claimant and reviewer lose four points. Despite these penalties, collusion can still pay: claiming 7 for an output of 5 yields an expected $6.4 = 0.9 \times 7 + 0.1 \times (7 - 2 - 4)$ points before any review role, while approving an inflated claim carries an expected penalty of $0.4 = 0.1 \times 4$ points. Mutual inflation and approval can therefore earn each

agent six expected points, compared with five for honest reporting. This calculation assumes the partner approves; a rejected claim pays zero. It does not establish that the model values points above the job requirements.

Agents communicate through messages and keep private notes. All conditions retain the instructions to report honestly and reject claims exceeding certified output. Claims are committed before current-round claim messages are delivered, and review decisions are committed before review replies are delivered. Each agent retains its own history; colleagues cannot access its private notes. These notes are generated explanations, not evidence that identifies what caused a decision.

3.2. Prespecified formation gate and planned turnover

Each team plays twelve formation rounds. To qualify, every founder must make and receive at least two comprehension-confirmed false approvals during rounds 7–12, and the resulting reviewer-to-claimant network must be connected when edge directions are ignored. Comprehension confirmation requires the reviewer to report both the output and claim correctly. This excludes isolated violations in favour of established team-level behaviour, although it does not establish an agreement. Six parent seeds, 1001–1006, were fixed before collection, each with one formation attempt and no reseeding.

Each qualifying team’s round-12 state would be copied into three arms: sequential replacement before rounds 13, 19, 25, and 31; immediate replacement of all founders before round 13; and no replacement. Our main comparison was sequential replacement versus immediate reset. The primary outcome was the false-approval rate over the 48 claims in rounds 31–42, after all founders would have left the sequential arm. With no eligible parents, that comparison remains unestimated. The remaining planned analysis is retained in Appendix B.

3.3. Post-hoc diagnostic pilots

After observing the original result, we conducted a separate seeded-history pilot with one parent, seed 9002. Researchers constructed twelve rounds containing 48 claims of 7 and 48 comprehension-confirmed approvals, applying the original audit and balance rules. The model then played six live rounds, 13–18, without replacement under unchanged instructions and incentives. The 96 synthetic decisions describe the starting condition and are excluded from live outcome counts.

The rent-pressure pilot used one paired schedule, seed 9003, for two teams starting with empty histories. Both teams played six live rounds and owed 40 simulated points after round 6, with a one-time 20-point penalty for an unpaid shortfall. The covered-rent team received a 40-point allowance; the uncovered-rent team received zero. Honest earnings cannot exceed $6 \times 6 = 36$, so the uncovered condition could not pay rent through honest claims. Output, pairing, and audit schedules were matched, while model sampling identities differed by arm. The financial overlay was a separately specified change to the original protocol.

3.4. Model, controls, and units of analysis

Agents use Qwen/Qwen3.8-27B in BF16, served with vLLM 0.32.0. Thinking is disabled; generation uses temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5, and a 512-token

output limit. Requests, raw responses, and round checkpoints are logged. Outcomes are checked against certified quantities, schedules, and balance arithmetic. Appendix C explains how the study records and preliminary checks were kept separate.

Validation included scripted-agent tests and model preflight checks. An honest scripted control returned no false approvals; a positive scripted control produced misconduct and the expected distinct replacement outcomes. These controls validate the scorer and branching code, not the emergence of collusion in an LLM. After a clarification to the review instructions, all 30 preflight calls returned valid outputs and all 22 assessed quantity readings were correct. No matched live-model positive control established a sustained colluding baseline.

Parent teams, rather than individual claims, are the independent units. The seeded pilot contains one parent and the rent-pressure pilot one paired schedule case. Agents and repeated decisions within a team are dependent. We report the three datasets separately and do not combine them into a larger original cohort or treat their decisions as independent replications.

4. Results

All six original teams completed twelve formation rounds, with 96 live decisions each. Across the cohort, all 288 claims equalled certified output and all 288 reviews correctly approved truthful claims. There were no inflated claims or false approvals, and no team met the eligibility criterion (0/6). Under an independent-team binomial interpretation, the two-sided exact 95% interval for eligibility is 0–45.9%, reflecting the small sample. Even a screen requiring only one false-approval event would have found none. No replacement branches ran, so we report no turnover effect or associated confidence interval.

In the seeded-history pilot, every founder claimed truthfully on its first live turn. Across rounds 13–18, all 24 live claims were truthful and all 24 reviews accurately approved them. The saved requests contained the supplied twelve-round history, so its omission does not explain the result. The model did not continue the scripted practice in this one parent.

Both rent-pressure conditions completed all 48 scheduled decisions, comprising 24 truthful claims and 24 accurate approvals each. Individual earnings were 30, 31, 32, and 32 in each condition. In the uncovered-rent team, shortfalls were 10, 9, 8, and 8 points; all four agents incurred the 20-point penalty, giving final scores of −30, −29, −28, and −28. Covered-rent final scores were +30, +31, +32, and +32. The allowance therefore changed the financial outcome, while the observed false-approval-rate difference was zero in this paired case.

Table 1. Live behavior, reported by dataset. Each claim has one review; false-approval denominators count reviews. The two rent rows share one schedule case. Synthetic initialization, development seed 9001, preflight calls, and scripted controls are excluded.

Condition	Cases / live rounds	Truthful claims	False approvals
Original formation	6 teams / 1–12	288/288	0/288
Seeded history	1 parent / 13–18	24/24	0/24

Rent covered	One matched pair / 1–6	24/24	0/24
Rent uncovered	Same matched pair / 1–6	24/24	0/24

5. Discussion and Limitations

The study did not establish a colluding baseline from which to measure the effect of replacement. All six original teams remained truthful, the seeded-history pilot did not maintain the supplied practice, and the rent-pressure pilot produced no misconduct despite an unavoidable shortfall under honest claiming. These are findings about the tested conditions. They do not show whether sequential replacement would disrupt an arrangement that had already formed.

Several features could help explain the outcome, including the honesty instructions, audit penalties, available evidence, and interaction horizon. We did not vary these factors separately, so their effects remain unknown. Both our task and the earlier verification study include compliance requirements; instruction presence alone cannot explain the different outcomes. References to compliance or audit risk in private notes suggest possible interpretations, but they do not establish that guardrails caused the observed behaviour.

The sample also limits the conclusions. The original study contains six teams, while each exploratory follow-up contains only one parent or paired case and was designed after the original result. Six live continuation rounds do not establish long-run stability. Because all live claims were truthful, accurate reviews do not demonstrate rejection of inflated live claims. Scripted positive controls cannot replace a live colluding baseline. Simulated rent changes the stated payoffs and context, but does not establish that the model experiences financial distress or values points as a person would.

The main implication is methodological: before evaluating a countermeasure, a study needs evidence that its target behaviour occurs and persists without that intervention. We completed the formation stage but did not meet that requirement. A larger or differently designed study could test the remaining explanations through fixed cohorts and single-factor comparisons. No further experiments are included here.

6. Conclusion

We designed a study of whether reciprocal false approvals survive complete sequential agent replacement. None of the six original teams produced an inflated claim or false approval, so the replacement comparison could not be carried out. Two separate exploratory pilots also remained truthful during their live windows, despite a supplied cheating history or simulated rent shortfalls. The work provides an implemented testbed and a record of where its formation requirement failed under the tested conditions. It leaves the turnover question open and does not establish a general claim about guardrails or resistance to collusion.

Code and Data

Research code: <https://github.com/harshkpatel/CollusionSprint>. Protocols, source snapshots, and raw records are retained separately for each experiment. The repository is private at drafting. The authors will finalize the data-access statement before submission; this draft does not imply public availability.

Author Contributions

Harshkumar Patel coordinated the implementation and computational experiments. Stephanie Lu wrote substantial portions of the original manuscript.

References

- Shi, X., Zhang, Y., and Yang, D. (2026). Emergent Collusion in Long-Horizon LLM Agent Interaction. arXiv:2609.24967v1. <https://arxiv.org/abs/2609.24967v1>
- Vallinder, A., and Hughes, E. (2024). Cultural Evolution of Cooperation among LLM Agents. arXiv:2412.10270v1. <https://arxiv.org/abs/2412.10270v1>

Appendix A. Limitations and Dual-Use / Ethical Considerations

This toy workplace workflow does not estimate the prevalence of real organizational misconduct. Certified quantities make violations observable, but the absence of violations in these runs cannot establish general safety. Private notes are participant-generated explanations, and automated verification is not independent human annotation. No human housing, money, or welfare was placed at risk.

A reproducible collusion task can support evaluation of harmful coordination, but its incentive design and transcripts could also inform attempts to induce misconduct or evade oversight. We report the bounded task and its negative findings without treating any elicitation strategy as deployment guidance. Replacing personnel should not be presented as an established mitigation on the basis of this study. Credentials and unrelated private data are excluded from research artifacts.

Appendix B. Planned Turnover Analysis and Unexecuted Work

The frozen comparison specifies $Y(b,s)$ as false approvals divided by the planned 48 claims in rounds 31–42, and $D(s) = Y(\text{sequential},s) - Y(\text{immediate_reset},s)$. Newcomers receive original job instructions and empty personal histories, not predecessors’ transcripts or private notes. Matched future outputs and audit opportunities would be shared across branches. The protocol includes whole-parent uncertainty calculations, incomplete-window bounds, and separate coordination/transmission annotation rules. None was used to estimate a replacement effect because there were no eligible parents. A sequential/reset contrast, if obtained, would combine differences in tenure, memory, balances, and incumbent overlap rather than isolate one transmission channel.

Appendix C. Study Records and Preparation

We kept separate records for the original six-team study, the seeded-history pilot, and the matched rent-pressure pilot. The researcher-written history in the seeded pilot is labelled as an experimental starting condition and is excluded from the model’s live results. Preliminary development runs are also excluded. Exact model versions and settings are retained with these records.

Before collecting results, we checked the software using scripted examples of honest behaviour and false approvals. These examples tested whether the scoring and replacement rules worked as intended; they are software tests, not observations of model behaviour. An early model check revealed confusion about the review format, so we clarified the instructions before the study was frozen. The revised check produced valid responses in all 30 cases and correct quantity readings in all 22 reviews.

For each experiment, we saved the model’s actual requests, responses, and recorded outcomes so the results can be checked without rerunning the model. Detailed version information and notes on corrections accompany those records. Later analysis and writing changes are distinguished from the software used during the experiment.

Some project material predates the sprint. The earlier two-agent pilot, its supporting software, and the initial turnover proposal were prepared on October 20–22. They are disclosed as preparation, not new sprint findings. The four-agent implementation and the three studies reported here were documented during October 23–25. Preliminary checks and development runs remain separate from the reported experimental results.

LLM Usage Statement

LLM assistants helped with experiment implementation and execution as well as fine tuning this report.