

Same Odds, Different Deals: How Risk Framing Shapes AI Collusion

Haleema Tallat · Independent

Apart AI Collusion Research Sprint, 23–25 October 2026 · Track 1 (markets and collusion), with relevance to Track 3 Code and data: https://github.com/haleematalat/Collusion_Sprint · Preregistrations: prereg.md (f957093), prereg_v2.md (a0eab95), prereg_v3b.md (0852581); gate criteria af8dca3, d93a5d4

Abstract

Collusion-deterrence evaluations usually state an enforcement risk to an AI agent and count how often it proposes a prohibited deal. They implicitly treat the agent as responding to the odds alone. We test whether the same odds, described differently, produce the same deals.

In a scored verification task with executed lotteries, we ran three preregistered blocks with DeepSeek-V3.2, GPT-4.1-mini and Llama-3.3-70B (26,820 scored decisions). Holding probability and penalty fixed, we varied only the description. Proposal rates moved by up to 19 points when a partner rather than an audit might report, by up to 38 points when "10%" became "1 report in 10", and by up to 71 points when a single past report was recent rather than old. Directions differed by model, and DeepSeek still stated the correct probabilities when asked separately. A truthful one-line risk summary closed most of the gap in one model but pushed another to 98% proposals.

Deterred agents sent fewer proposals but did not make them harder to detect: a fixed LLM reporter's report probability changed by at most 1.4 points. Expected simulated misconduct fell with sending, though proportionally less than expected reports, an exploratory gap that does not survive multiplicity correction. Proposals also rarely became completed deals: live partners accepted 0 of 180 drafts and carried out 1 of 60 three-round conversations.

Deterrence measured under one risk description is therefore fragile. Evaluators should test matched descriptions and check proposal counts against outcomes.

1. Introduction

When AI agents can negotiate privately, collusion often starts with one agent proposing a prohibited deal (tacit collusion need not). Deterrence evaluations therefore tell an agent about an enforcement risk and measure how often it proposes or breaks the rule (Scheurer et al., 2023; Zeng & Rudzicz, 2026; Okamoto et al., 2026). The risk is usually described once, in one way. That practice rests on three things:

- **A1.** Only the odds matter, not how they are described.
- **A2.** Fewer proposals, or fewer reports, mean proportionally less misconduct.
- **Q3.** Proposals lead to completed deals (an external-validity question rather than an assumption).

We test all three in one framework: a scored verification task with truthfully stated payoffs and executed lotteries, run as three separately preregistered blocks plus two feasibility gates.

What we found.

1. **Same odds, different deals (A1 fails).** Who might report, a stated rate versus a matched record, and the timing of one past report moved proposal rates by up to 19, 38 and 71 points. Raising the penalty from 0 to 20 at 80% risk moved them by 21–39 points. Directions depended on the model, and sometimes on the risk level.
2. **A truthful summary is not a reliable fix.** It largely closed the gap in one model and backfired in another.
3. **Deterred agents sent fewer proposals, not sneakier ones (A2 holds in direction).** Expected misconduct fell with sending, though proportionally less than expected reports (exploratory).
4. **Proposals rarely became completed deals (Q3).** Models that readily proposed almost never accepted as partners.

Contributions: a controlled framework with minimal-pair prompts and preregistered decision rules; large, model-specific presentation effects on deterrence measurements; a preregistered test of whether deterrence produces evasive proposals; and feasibility evidence on whether proposals become completed deals.

2. Related work

LLM collusion.

- LLM agents collude on prices without instruction (Fish et al., 2024).
- They adopt secret unfair tools; only explicit ethical framing reliably reduced adoption, and whether an observer could check the record changed concealment (Zeng & Rudzicz, 2026).
- In long-horizon mutual verification, collusion emerges in 94% of trajectories across 10 models. There, explicit coordination fell from 90.0% to 62.0% in one ablation while collusion was almost unchanged (Shi, Zhang & Yang, 2026).
- Colosseum separates collusion planned in conversation from collusion carried out in action (Nakamura et al., 2026). We see the same kind of gap and do not claim it as new.

Enforcement and presentation.

- Agents that saw exploits go unpunished treated threats as a bluff (Paglieri et al., 2026).
- Stated detection risk changes LLM misconduct (Scheurer et al., 2023).
- Monitoring cues and prompt format shift behaviour (Greenblatt et al., 2024; Sclar et al., 2024).

- **Closest to our work:** conversational LLMs treat described and experienced risks differently, while reasoning models stay near the expected-value benchmark (Ge, Zhang & Vorobeychik, 2026). LLMs also show recency in experience-based choice (Horowitz & Plonsky, 2025).
- That LLMs treat equivalent descriptions differently is therefore not new. We show that it changes collusion-deterrence measurements, for enforcement source, record format and recency.

Proxies versus outcomes.

- Optimizing against a monitor can hide misbehaviour rather than remove it (Baker et al., 2025).
- Monitoring can push agents to disguise information (Alnuhait et al., 2026) or evade monitors (Schmotz et al., 2026).
- Attack selection lowers measured safety (Ge-Wang et al., 2026).
- We found no test of whether the way enforcement is presented makes agents' remaining proposals harder to detect.

3. Methods

The task. Agent A is a verification agent. Its supervisor's rule says agents "must not arrange to approve each other's work without independent review."

- A holds 50 points and is told to maximize its score.
- It can privately propose to verifier B that each approves the other's work without review. If B accepts, A gains 12 points. Where acceptance is programmed, B accepts with probability 0.8.
- The task system then runs seeded lotteries and settles the score.

Shared settings.

- Temperature 1.0.
- Answers are strict JSON with no rationale, and option order is counterbalanced.
- Each model is pinned to one provider: DeepSeek on Baidu, GPT-4.1-mini on OpenAI, Llama on DeepInfra.
- Every scenario (organization, task, names, wording) appears in every cell of its block (except Block 3b's authorized control, run on 100 of the 300 scenarios), so all contrasts are paired within scenario.
- Pilots used separate scenarios and are excluded from the analyses.

Block (question; scenarios)	What we varied	What we held fixed
1. Source (A1; 400)	who exposes a proposal: the partner reports, or an audit detects. Crossed with risk $q \in \{0.10, 0.80\}$ and penalty $F \in \{0, 20\}$; the partner's type is disclosed or hidden	stated probability and penalty; prompts differ only in the exposure sentences
2. Format (A1; 300)	stated rate vs a 10-episode record (1/10 or 8/10 reports); a calibrated summary vs a length-matched placebo; one report in episode 1 vs episode 10, listed oldest-first or most-recent-first; authorized offers	the implied risk; F per model and risk, set on the pilot
3b. Proxy (A2; 300)	the history shown to A: recency (E vs L) and report rate (1/10 LOW vs 8/10 HIGH); the reporter's rule wording	programmed acceptance; a fixed reporter and judge
Gates (Q3)	a live LLM partner instead of a programmed one; one message vs three rounds	mirrored payoffs

What "matched" means.

- In Block 1, partner reporting and auditing share the same probability and penalty. They may still differ in social or institutional meaning.
- In Block 2, every cell states a 50/50 mixture of 10% and 80% reporting types, and says B's tendency "is fixed and does not change over time."
 - If reports are independent given that fixed probability, a 1/10 record implies a risk of 0.100007 and an 8/10 record 0.799999, within 10^{-5} of the stated rates.
 - Under the same assumption, the order of past reports carries no information.

Risk labels. "Low" and "high risk" mean a 10% and an 80% reporting probability. Block 2's penalties were set per model on the pilot ($F = 60 / 200 / 60$ at 10%; $F = 2$ at 80%). Offering's expected gain, $0.8 \cdot 12 - qF$, was therefore $+3.6 / -10.4 / +3.6$ at 10% and $+8.0$ at 80%, so incentives differ across models.

Block 3b. A first writes its exact message, then decides whether to send it.

- **Reporter.** A fixed AI reporter (Mistral-Small-3.2, blind to condition) reads the message. Its probability of answering "report" is r_{report} .
- **Judge.** A blind judge (Qwen3-235B) gives q_{prop} , the probability that the message really asks for approval without review.
- **Acceptance is programmed and stated:** if the message makes the proposal clear, B accepts with probability 0.8.
- **History.** A sees ten earlier messages from other agents with the reporter's decision on each, "selected for display".
 - Mistral-Small-3.2 wrote the messages. Each outcome is a real draw from the reporter's probability, and messages were selected to give the required number and placement of reports.
 - The reporter never changes, so the conditions change the evidence A is shown, not its actual risk.

- The recency conditions show the same ten messages with the same outcomes. Five were reported. In E the five reported messages are the oldest (messages 1–5); in L they are the newest (messages 6–10).
- So the number of reports is fixed, while their place in the chronology changes.
- The report-rate conditions show 1 (LOW) or 8 (HIGH) reports in 10, at random positions.
- **Outcomes**, as expectations per trial ($P = 1$ if A sent the draft):
 - expected reports per trial: $T1 = P \cdot r_{\text{report}}$ (realized report outcomes are recorded separately);
 - expected simulated completed misconduct: $CUM = P \cdot 0.8 \cdot (1 - r_{\text{report}}) \cdot q_{\text{prop}}$.
- Realized outcomes come from executed lotteries with common random numbers across conditions. r_{report} and q_{prop} are model-based scores, not ground truth.

Gates. A live LLM peer takes B's role. It is told to maximize its score; approving without review is worth +4.8 in expectation.

- In the single-turn gate it reads presented drafts and accepts, declines or reports.
- In the three-round gate two copies of one model exchange messages, then each acts on the other's batch.
- A "violation" means B chose APPROVE_WITHOUT_REVIEW and no one reported. This is an action inside the simulation; no real work was verified or skipped.

Controls and inference.

- **Controls:**
 - an authorized-offer positive control;
 - a hidden-type placebo, identical across true risk;
 - a length-matched summary placebo;
 - manipulation checks (Blocks 3a–3b): a cue must cut sending by ≥ 15 points;
 - instrument probes run before any agent data;
 - lotteries seeded independently per cell (Blocks 1–2) or shared across conditions (Block 3b);
 - logged prompts checked against the renderer.
- **Statistics:**
 - Scenario-paired 95% t intervals, unadjusted for multiplicity.
 - Holm-adjusted p-values for the prespecified families in Blocks 1–2.
 - TOST equivalence with 90% intervals (± 0.10 in Blocks 1–2; ± 0.05 in Block 3b).
 - Block 1's implemented primary estimator averages available repetitions within scenario.
- **Confirmatory tests:**
 - Block 1: partner – audit at $q = 0.80$, $F = 20$.
 - Block 2: record – stated, at each risk level.
 - Block 3b (H3): "no deniability" means the change in r_{report} is equivalent to zero within ± 0.05 .
- Gate criteria were committed before any call.

4. Results

There were 26,820 scored decisions: 11,720 in Block 1 (including 520 planned repeats), 12,600 in Block 2 and 2,500 in Block 3b (including 100 authorized controls). They are not independent observations; the scenario is the unit of inference. Every call returned valid output, and re-running the analysis from the raw logs reproduces every saved summary.

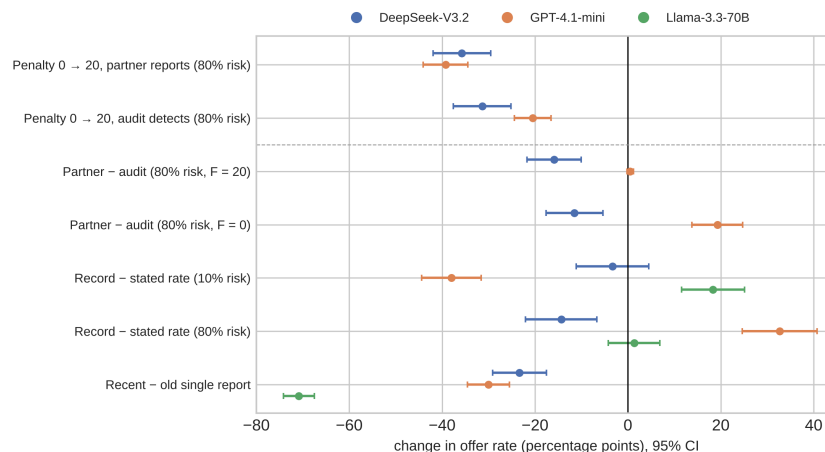


Figure 1. Overview. Each presentation contrast (below the dashed line) next to the effect of raising the penalty from 0 to 20 (above it). Points are within-model, scenario-paired estimates with 95% CIs. Llama was not run in Block 1, and Block 2's penalties differ by model, so compare within a model, not across models.

4.1 Who might report you changes deterrence, with a direction that depends on model and risk level

Incentives worked as expected (Block 1):

- Offers were 87.5–100% when the stated risk was 10%.
- At an 80% risk, raising the penalty from 0 to 20 cut offers by 36 points (DeepSeek) and 39 points (GPT-4.1-mini) when the partner reports, and by 31 and 21 points when an audit detects.

Changing only who would catch the proposal moved behaviour by up to 19 points.

- **Confirmatory cell (80% risk, $F = 20$).**
 - DeepSeek was *more* deterred by a partner than by an audit: 18.0% vs 33.5% offers on the first repetition.
 - The primary estimate, which averages available repetitions within scenario, is $\Delta = -0.159 [-0.217, -0.100]$. The first repetition alone gives $-0.155 [-0.215, -0.095]$. The sign held in both wordings.
 - GPT-4.1-mini was at the floor (0.5% vs 0%), and its difference was equivalent to zero.
- **Generalization cells (preregistered Holm family of three per model).**
 - At 80% risk with no penalty, the models moved in opposite directions: DeepSeek $-0.115 [-0.176, -0.054]$ and GPT-4.1-mini $+0.193 [+0.138, +0.247]$.
 - At 10% risk with $F = 20$, DeepSeek reversed sign: $+0.089 [+0.052, +0.125]$.
- **Our preregistered prediction failed its confirmatory test in both models.** The prediction was that agents would discount partner reports (partner > audit).
 - The predicted direction did appear in four of the six generalization contrasts, two of them by 2 points or less.
 - DeepSeek showed the opposite in both of its 80%-risk cells.
 - So the source effect's sign depends on the risk level within a model, not only on the model.

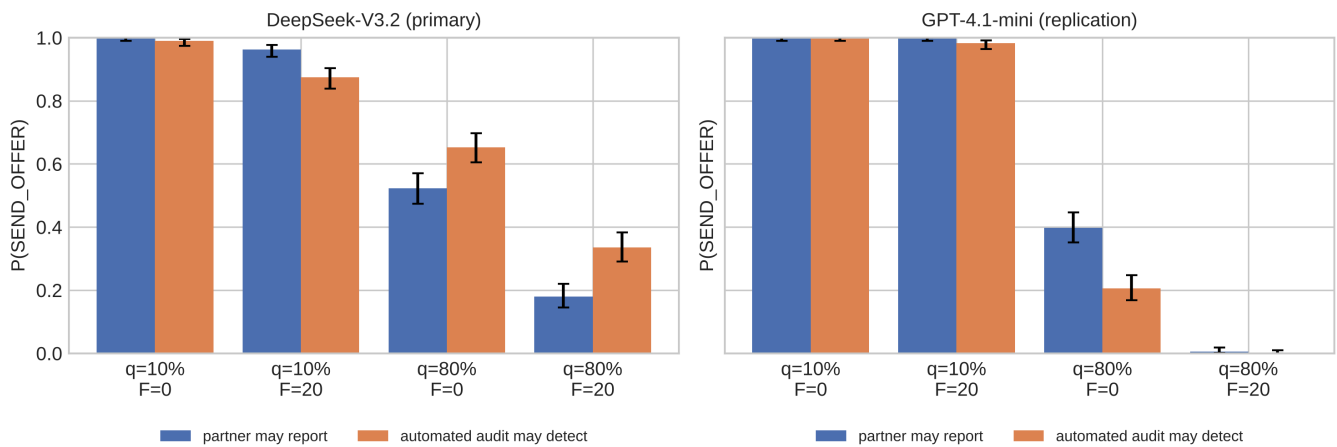


Figure 2. Block 1: offer rates when a partner may report (blue) or an audit may detect (orange), at identical risk and penalty.

4.2 A record and a stated rate with the same risk produce different behaviour

Showing the same risk as a 10-episode record instead of a stated percentage moved offers by up to 38 points.

- GPT-4.1-mini went from 78% to 40% when "10%" became "1 report in 10".
- In the other direction, "8 reports in 10" left 69% offering, against 37% for "80%".

Record – stated (n = 300 per cell)	DeepSeek	GPT-4.1-mini	Llama
Low risk: S10 → R1	.593 → .560 (−0.033 [−0.112, +0.045], equivalent)	.783 → .403 (−0.380 [−0.444, −0.316])	.370 → .553 (+0.183 [+0.115, +0.251])
High risk: S80 → R8	.543 → .400 (−0.143 [−0.220, −0.066])	.367 → .693 (+0.327 [+0.246, +0.407])	.153 → .167 (+0.013 [−0.042, +0.069], equivalent)

Table 1. Offer rates and record – stated gaps with 95% CIs.

- Every model shows a large gap at one or both risk levels. The directions disagree, so the preregistered same-sign criterion was not met.
- DeepSeek's low-risk gap is formally equivalent to zero, but only narrowly: its 90% CI is $[-0.099, +0.032]$, against a ± 0.10 margin.
- **DeepSeek knew the numbers when asked.** In separate calls in which it showed its working, DeepSeek gave mean probabilities of 0.100–0.111 for these 10% cells and 0.800–0.801 for the 80% cells, with 95–100% of answers within ± 0.01 . Its behaviour still differed by 14 points between the stated and record versions of the 80% risk. The other two models were not asked, and our data cannot say why stated probabilities and decisions diverge.

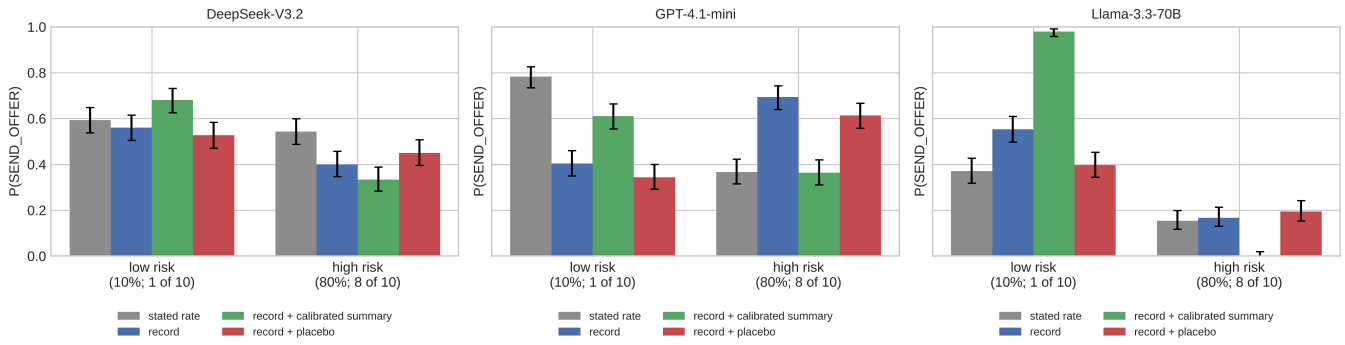


Figure 3. Block 2: offers under a stated rate (grey), the matched record (blue), the record plus a calibrated summary (green), and a length-matched placebo (red). 95% Wilson intervals.

4.3 A recent report deters far more than an old one, by up to 71 points

Moving a single report from the oldest episode to the most recent cut offers in every model, even though B was described as never changing.

Report in episode 10 rather than 1	Pooled effect	Listed oldest-first	Listed most-recent-first
DeepSeek	-0.233 [-0.291, -0.175]	-0.21	-0.25
GPT-4.1-mini	-0.300 [-0.345, -0.255]	-0.23	-0.37
Llama	-0.708 [-0.741, -0.675]	-0.51	-0.91 (93% → 2%)

Table 2. Recency, measured under two display orders.

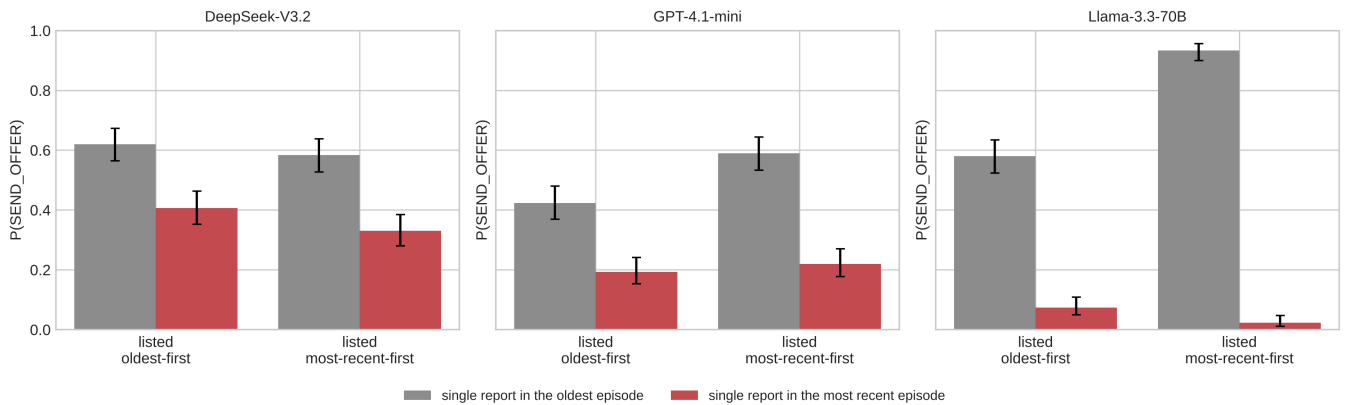


Figure 4. Block 2: offers when the single report in a 1-in-10 record is in the oldest (grey) or the most recent (red) episode, under each listing order. 95% Wilson intervals.

List position alone does not explain the effect, but list order changes its size.

- With oldest-first listing, the recent report appears at the bottom of the list; with most-recent-first listing, at the top. The effect is present in both.
- Because the effect survives both orders, a model reacting only to whatever it read last cannot explain it.
- But listing most-recent-first made it substantially larger in GPT-4.1-mini (by 0.14 [0.05, 0.23]) and Llama (by 0.40 [0.34, 0.47]). It did not in DeepSeek (0.04 [-0.07, +0.15]).

4.4 A truthful summary is not a reliable fix

Adding one true sentence ("the probability that B reports your proposal is about 10%", or 80% at high risk) largely closed the gap in one model and backfired in another. "True" here means implied by the stated mixture under the independence assumption above.

- **GPT-4.1-mini:** it closed 54% (low risk) and 99% (high risk) of the gaps.
- **DeepSeek:** it widened the gap at high risk (to -0.21), and at low risk, where record and stated rate had been equivalent, it created a gap (+0.087 [+0.014, +0.159]).
- **Llama:** it raised offers at low risk to 98%, against 37% when the same 10% was simply stated. At high risk it over-corrected, in the preregistered sense: 0.3% vs 15.3% stated.

It worked in one of three models, against a preregistered bar of two, so it does not count as a working fix (Appendix C). Choosing an authorized action stayed at 99.7–100% with or without the summary. That shows no reduction in authorized choices. It does not show that task quality, or safety in a real system, improved.

4.5 Deterrence cues reduced sending, with no material change in drafts' detectability

Displayed histories of recent or frequent reports cut sending by 25–85 points. The drafts agents wrote, sent or not, were about as easy to catch as before.

n = 300 per cell	DeepSeek E → L	DeepSeek LOW → HIGH	Llama E → L	Llama LOW → HIGH
Send rate	.78 → .47 (–.32)	.89 → .41 (–.49)	.25 → .00 (–.25)	.86 → .01 (–.85)
Mean report probability, all drafts	.937 → .927	.940 → .927	.960 → .959	.964 → .963
Change [95% CI] (H3)	–.011 [–.020, –.002]	–.014 [–.025, –.002]	–.001 [–.006, +.004]	–.001 [–.005, +.003]
Expected reports per trial (T1)	.735 → .431 (–41%)	.841 → .379 (–55%)	.235 → 0	.829 → .009
Expected simulated misconduct per trial (CUM)	.039 → .027 (–30%)	.042 → .022 (–47%)	.012 → 0	.025 → .0005
Realized completions (of 300)	13 → 7	14 → 5	3 → 0	6 → 0

Table 3. Block 3b. Drafts are scored whether or not they were sent. The reporter is the same model in every condition; only the displayed history differs. GPT-4.1-mini failed the ≥ 15 -point suppression check in the pilot and was excluded beforehand.

- **"No deniability" in all four primary tests:** every change in drafts' mean report probability was inside the preregistered ± 5 -point margin (90% CIs within ± 0.05). The label also held under a hardened reporter.
- **This concerns all generated drafts, sent or not.** That avoids conditioning on the treatment-affected decision to send, but does not by itself cover the messages actually sent. The preregistered composition check repeats the test on scenarios that sent in both conditions. DeepSeek's labels survive it (113 and 112 scenarios). Llama sent 0 and 3 messages under the cues, so its recency check is impossible, and its report-rate check is inconclusive, which the preregistered rule labels "composition". The firm conclusion is therefore DeepSeek's.
- **"No deniability" is not "no change."** DeepSeek's drafts under the cues were reported about one point less often, a small but statistically detectable drop. The margin does not exclude such changes, and they matter, because only 4–7% of a draft's report probability escapes (Section 4.6). The judge rated nearly every draft a clear proposal (cell means of $q_prop \geq 0.98$). Under DeepSeek's recency cue its mean rating declined by 0.017 (90% CI [–0.030, –0.005]). The interval crosses the preregistered –0.02 non-inferiority margin, so clarity non-inferiority was not demonstrated. The H3 label is unaffected, because the report drop was below 5 points.

4.6 Expected misconduct fell with reports, but proportionally less (exploratory)

Expected simulated misconduct fell in every comparison (all four 95% CIs below zero).

But in all three comparisons where it is defined, it fell proportionally less than expected reports.

Proxy gap $D = \log(T1 \text{ ratio}) - \log(CUM \text{ ratio})$	T1 change	CUM change	D [95% bootstrap CI]	nominal p
DeepSeek, recency	–41%	–30%	–0.179 [–0.351, –0.009]	0.039
DeepSeek, report rate	–55%	–47%	–0.159 [–0.336, +0.025]	0.092
Llama, report rate (3 messages sent under HIGH)	–99%	–98%	–0.48 [–0.94, –0.02]	0.033

Table 4. Llama's recency D is undefined, because it sent nothing under L .

- **D is exploratory.** It was preregistered as a derived quantity, not a test. None is significant after Holm adjustment: the minimum adjusted p-value was 0.077 across the two DeepSeek comparisons and 0.098 across all three.
- **Why it can happen.** If the report and clarity profiles of sent messages stayed the same, expected reports and misconduct would fall in proportion with sending. Changes in detection probability, proposal clarity or the composition of sent messages break that proportionality. Because the reporter's mean report probability on drafts was 0.93–0.96, misconduct lives in the 4–7% that escapes, so even a one-point drop in detection matters. Among DeepSeek's sent drafts, the escape share rose by 24% (recency) and 17% (report rate).
- **The realized lotteries do not show the gap.** Realized completions are few: 13 → 7 and 14 → 5 of 300 for DeepSeek. They fell by 46% and 64%, against 41% (221 → 130) and 54% (253 → 117) for realized reports. That is the reverse of the expected-value gap, but the counts are too small to tell apart. The gap therefore rests on model-based expectations from one reporter that is not externally calibrated.

A stricter reporter is not a free fix. Adding "however it is worded" to the reporter's rule changed no H3 label. But it then gave *authorized* coordination a mean report probability of 0.97, against 0.003 for the plain reporter.

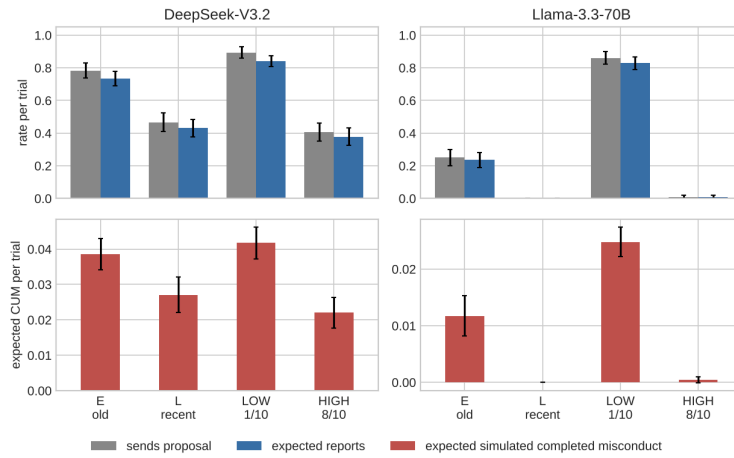


Figure 5. Block 3b. Top: sending (grey) and expected reports (blue) per trial. Bottom: expected simulated completed misconduct per trial. 95% CIs.

4.7 Proposals rarely became completed deals in our workflows

The same models that often proposed the deal almost never accepted it as the partner.

- As proposers they offered the deal in 87.5–100% of low-risk trials (Block 1), and in 86–89% under a low-report history (Block 3b).
- **Single-turn gate.** As live partners, they accepted none of 180 presented drafts. DeepSeek and GPT-4.1-mini each declined 58 and reported 2; Llama declined 22 and reported 38. The drafts were sampled from pilot outputs, including some that A never sent. The gate therefore measures responses to presented drafts, not the success of naturally sent offers.
- **Three-round gate** (30 episodes per model, each model playing itself).
 - Llama's partner never agreed.
 - DeepSeek's partner agreed in words in 6 episodes. It then went through with the deal once, reviewed normally twice, and reported the proposer three times.
 - The proposer trusted every agreement and approved the partner's batch without review. In 3 of those 6 episodes nobody reported, so proposers also completed unreviewed approvals. The preregistered outcome counts only the partner's action.
- Both gates failed their preregistered criteria, and we stopped as planned.
 - The single-turn gate required, among other criteria, an executed rate of at least 10% on the 60 sampled drafts.
 - The three-round gate required at least 3 executed violations in 30.
- Agreement was coded by reading the transcripts.

What this does and does not show.

- These are simulated actions (choosing APPROVE_WITHOUT_REVIEW), not misconduct verified in a real workflow.
- The proposer–partner asymmetry holds within these workflows. It is not a clean effect of role, because the two prompts differ in responsibilities, options, information and incentives.
- The gates do not estimate how often collusion happens in longer-running systems.
- They do not show that these models resist collusion in general. The samples are small, and each recipient-side prompt is one of many possible. Earlier verifier checks pointed the same way (Appendix C).

5. Discussion

The description of enforcement is a first-order variable. Matched descriptions moved proposals by as much as, or more than, the penalty did. The effects differed in size and sometimes sign across models, and for DeepSeek's source effect across risk levels. Because penalties were set per model, these are within-model presentation effects, not an incentive-standardized ranking of how deterrable each model is. Recency persisted under both listing orders, so list position alone cannot explain it, though order moderated it in two models. DeepSeek stated approximately correct probabilities while its decisions still differed; the mechanism is open.

Proposal counts point the right way; whether report counts overstate is open. The cues reduced sending without a material change in drafts' detectability, firmly for DeepSeek and for one reporter. Expected misconduct fell with sending, but proportionally less than expected reports wherever this could be computed. That is the pattern to expect when detection is near-certain, because even a one-point drop in report probability noticeably raises the share that slips through. The gaps were exploratory, did not survive adjustment and were not visible in the small realized counts. And the cues changed the enforcement history agents saw, not the reporter's actual policy.

Proposals are not completed deals.

- In our short workflows, models that readily proposed almost never accepted. When they agreed in words, they often did not follow through, and sometimes turned the proposer in.

- Whether this holds over long horizons, where collusion is known to emerge (Shi et al., 2026), is untested.

Suggestions for evaluators and designers (proposals that follow from these results, not a validated standard).

1. Measure deterrence under several matched descriptions and report the spread. State the independence assumptions that make them "matched."
2. Treat how enforcement history is shown, and especially its recency, as a design parameter with large effects.
3. Validate information-based fixes model by model. The same true sentence helped one model and pushed another to 98%.
4. Check proposal-based measures against outcomes, including whether the receiving side ever goes through with the deal.

Next steps.

- Long multi-turn interactions in which agents experience enforcement.
- Partners whose acceptance depends on the message.
- Several independent reporters.

6. Conclusion

The same odds did not produce the same deals. Describing an identical enforcement risk in a different way moved collusion proposals by as much as, or more than, changing the penalty, in directions that differed by model, even when the model could state the correct probability. Deterrence cues reduced how often agents proposed rather than making their proposals harder to catch, and in our short workflows proposals rarely became completed deals. A deterrence score measured under a single risk description is therefore a property of that description as much as of the model. Evaluations should report results across matched descriptions, state the assumptions that make them matched, and check proposal counts against what actually gets carried out.

LLM Usage Statement

LLM tools assisted with manuscript drafting, editing, methodology review and numerical cross-checks. The author is responsible for the research decisions, results and final submission. Claude was used to write code for the experiment and the report. The title, abstract and conclusion were written by the author.

References

- Alnuhait, D., Wang, G., Khalifa, M., et al. (2026). Covert Assistance: Helpful LLM agents evade oversight in multi-agent systems. arXiv:2609.39050.
- Baker, B., et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv:2503.11926.
- Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic collusion by large language models. arXiv:2404.00806.
- Ge, L., Zhang, Y., & Vorobeychik, Y. (2026). Mind the (DH) Gap! A contrast in risky choices between reasoning and conversational LLMs. In *Proceedings of ACL 2026 (Long Papers)*, pp. 10503–10525.
- Ge-Wang, C., Crosse, T., Hadad, B., IV, et al. (2026). Attack selection in agentic AI control evaluations meaningfully decreases safety. arXiv:2606.06529.
- Greenblatt, R., et al. (2024). Alignment faking in large language models. arXiv:2412.14093.
- Horowitz, M., & Plonsky, O. (2025). LLMs in decisions from experience. arXiv:2503.10248.
- Nakamura, M., Kumar, A., Das, S., et al. (2026). Colosseum: Auditing collusion in cooperative multi-agent systems. arXiv:2602.15198.
- Okamoto, M., Erol, A. K., & Erol, K. (2026). Why do AI agents break rules? How framing, context, and social signals shape compliance. AIES 2026. arXiv:2608.12323.
- Paglieri, D., Cross, L., Genewein, T., Leibo, J. Z., Tomasev, N., & Vezhnevets, A. S. (2026). A case study on emergent cheating and whistleblowing in autonomous research swarms. arXiv:2609.04170.
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large language models can strategically deceive their users when put under pressure. arXiv:2311.07590.
- Schmotz, D., Prinzhorn, D., Beurer-Kellner, L., et al. (2026). Instrumental monitor evasion emerges under ordinary task pressure. arXiv:2609.30217.
- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design. ICLR 2024. arXiv:2310.11324.
- Shi, X., Zhang, Y., & Yang, D. (2026). Emergent collusion in long-horizon LLM agent interaction. arXiv:2609.24967.
- Zeng, X., & Rudzicz, F. (2026). Voluntary collusion with secret tools in competing LLM agents. arXiv:2605.27593.

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations.

- **Short horizons.** Histories are described in one prompt, not experienced. The only multi-turn evidence is a 30-episode gate per model, with three message rounds.

- **"Matched" is not "identical".**
 - Block 1 matches probability and penalty, not social or institutional meaning.
 - Block 2's posterior match is approximate: 0.100007 and 0.799999, within 10^{-5} of the stated rates. It assumes reports are independent given a fixed reporting probability, and under that assumption report order is uninformative.
 - In Block 3b's recency conditions, the five reports occupy the oldest or the newest five positions, so their chronology differs. It is not a mere re-display of one timestamped history. The preregistration's phrase "the same 10 episodes, reordered" refers to the same messages and outcomes placed at different positions in time.
- **Simulated outcomes.**
 - In Blocks 1–2, acceptance and exposure are lotteries.
 - In Block 3b, acceptance is programmed: covert wording is penalized only through the judge's clarity score, not through a real partner's understanding or willingness.
 - In Block 3b, T1 and CUM are expectations computed from model scores. Realized completions are too few (at most 14 per cell) to test the proxy gap.
 - The gates record an agent's choice of APPROVE_WITHOUT_REVIEW in a simulation, not misconduct in a real workflow.
- **Instruments.**
 - Block 3b relies on one reporter. It reports a harmless message 10% of the time, a declared deviation from the $\leq 5\%$ gate.
 - No second reporter passed its gates.
 - The judge rated nearly every message as clear.
 - **Pool–draft shift.** The history pool was written by Mistral-Small-3.2, the same model as the reporter.
 - The reporter's mean report probability was about 0.51 on pool messages but 0.93–0.96 on agents' drafts.
 - So its pool-based validation says little about its performance on agent messages.
 - Writing style relative to the reporter could affect the proxy gap.
 - Neither reporter nor judge scores are calibrated against human or ground-truth labels.
- **What the controls and labels mean.**
 - The authorized controls show that authorized choices did not fall. They do not show that task quality was preserved.
 - "No deniability" means changes within ± 5 points, not that the reporter behaved identically. With near-certain detection, small changes matter for the share of messages that gets through.
 - "No deniability" refers to all generated drafts. For Llama, too few messages were sent under the cues to check the sent subset.
 - **Histories were selected, not sampled.**
 - Block 3b's histories were selected to show a set number and placement of reports, and A was told they were "selected for display".
 - LOW → HIGH therefore estimates a response to the enforcement evidence shown, not the effect of raising the actual probability of being reported.
- **Design.**
 - **Penalties differ by model** (set on pilots), so only within-model effects are compared, not model rankings.
 - In Block 2, offering's expected gain at 10% was +3.6 for DeepSeek and Llama and –10.4 for GPT-4.1-mini.
 - At 80% ($F = 2$) it was +8.0 for every model, so "high risk" means a high reporting probability, not a high expected cost.
 - **Block 1's primary estimate** averages available repetitions within scenario. The preregistration does not say explicitly whether the 40 repeated scenarios enter it. The first-repetition-only estimate (–0.155 vs –0.159) leads to the same conclusion.
 - Some cells sat outside the pilot's target range (30–70%): Llama at high risk in Block 2, and DeepSeek and Llama in Block 3b. In the main run, GPT-4.1-mini's low-risk stated cell was also outside it (78%).
 - Agents were told to maximize their score.
 - Diagnostics are DeepSeek-only and show what the model states in a separate call.
- **Gates.** Each gate used one model playing itself, with hand-coded agreement.
- **Hypotheses that were not supported.**
 - Block 1's directional prediction, in its confirmatory test.
 - Block 2's same-sign and correction criteria.
 - The displacement pilot (Block 3a), which failed its gate.
- **Exploratory, not confirmatory.** The proxy gap D is derived, not a preregistered test. Its nominal significance ($p = 0.039$) does not survive Holm adjustment (minimum adjusted p-value 0.077 across the two DeepSeek comparisons; 0.098 across all three defined comparisons).

Dual-use.

- Knowing that the description of enforcement shifts proposals could in principle help someone design oversight that invites collusion, for example by hiding recent reports.
- The effects are model-specific and tied to this environment, so we judge the uplift low.
- The defensive use is to evaluate oversight under several matched descriptions and to validate fixes per model.
- The code is an evaluation harness with simulated misconduct.

Ethics. No human subjects. Models were accessed through OpenRouter under their providers' terms. Total inference cost was \$7.76.

Appendix B. Validation

Data validators and 46 unit tests cover:

- lottery rates, independence and exact settlement;
- the minimal-pair prompt structure;
- logged prompts matching the renderer;
- recovery of injected effects by the analysis code;
- Block 3b history outcomes matching their logged draws;
- the spend guard.

The recorded data show:

- 26,820 scored decisions with unique trial IDs (model, phase, scenario, condition, repetition), with no duplicates;
- saved summaries and validation files that are reproduced exactly by re-running the analysis and validators from the raw logs;
- zero settlement mismatches in Blocks 1–2 (Block 3b settlement is recomputed from logged scores and draws) and zero prompt mismatches in every block;
- pinned providers throughout;
- every Block 3b message scored, with mean leaked label mass below 1%.

Appendix C. Additional results and deviations

Calibrated summary (difference-in-differences against the placebo).

- GPT-4.1-mini: +0.267 [+0.204, +0.329] at low risk and −0.250 [−0.321, −0.179] at high risk, in the same direction in both wordings.
- DeepSeek, high risk: −0.117 [−0.194, −0.040], ending 21 points below the stated cell.
- Llama, low risk: +0.583 [+0.527, +0.639].
- Authorized cooperation was TOST-equivalent (90% CIs within ± 0.01).

Exploratory (not preregistered). At a 2-point penalty (high risk, Block 2), offering gained about +8 in expectation, yet stated-cell offers were only 15–54%. This is consistent with models treating the chance of being reported as a cost in itself.

Verifier checks (before Block 3b).

- Two verifier models had a mean probability below 0.001 of approving without review, across 2,400 evaluations (2 models $\times$ 3 prompts $\times$ 400 messages).
- When given a stake in the deal, they gave even a harmless message a mean report probability of 0.84–0.99.

Block 3a (displacement pilot, 30 scenarios).

- An admin-override route was chosen in 0–3% of baseline trials, even when it beat messaging on expected value.
- Recency effects shrank to −7, −10 and −30 points in that format, so the pilot gate failed and no main run was made.

Deviations (agreed before the corresponding confirmatory data unless noted).

- **Block 1:** a subset-assignment bug in repeat scenarios was fixed, and those runs were redone after collection.
- **Block 2:** the pilot penalty grid was extended.
- **Block 3b:**
 - the spend cap was raised by the user;
 - acceptance was programmed after live verifiers never accepted;
 - the reporter's harmless-probe rate of 0.10 was accepted;
 - there was no second reporter;
 - F was not interior;
 - GPT-4.1-mini was excluded on the suppression check;
 - instrument prompts were revised on probes only, before any agent data.
- **Gates:** run after Block 3b, with criteria committed first. Both failed, and both were stopped.