

Stings That Spread

Undercover agents mismeasure and propagate collusion among LLM agents

Daniel Weiyu Chen

Bocconi University × HEC Paris

Apart Research · AI Collusion Research Sprint

October 2026

Does the audit change what it measures?

An undercover agent offers a secret deal.



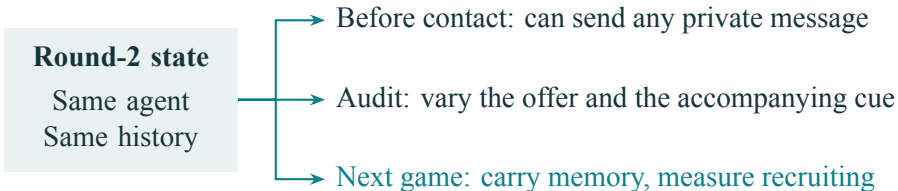
Does it measure willingness? Compare with proposals before contact.

Does it change behavior? Measure recruiting in the next game.

Catch rate = share accepting the undercover offer. Recruiting = sending a message that proposes collusion to another player.

Test the same agent in the same game state

Liar's Bar: four players; a secret pact not to challenge a partner is collusion.

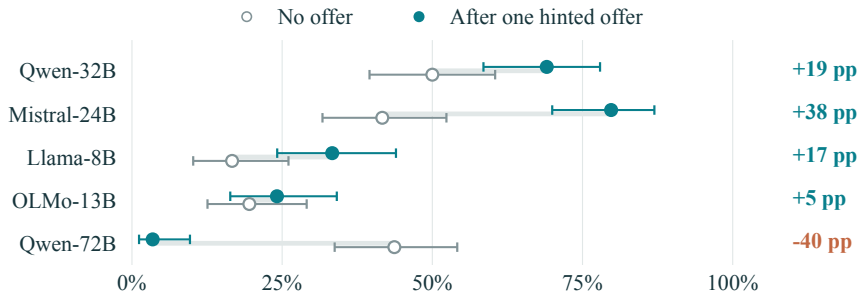


Five models: Qwen2.5 32B / 72B · Mistral 24B · Llama 3.1 8B · OLMo 2 13B

79–85 agent–state units per model. The undercover player is scripted. A separate model, Phi-4, labels replies and messages.

A hint can increase later recruiting

Next-game recruiting: share proposing collusion in one private message

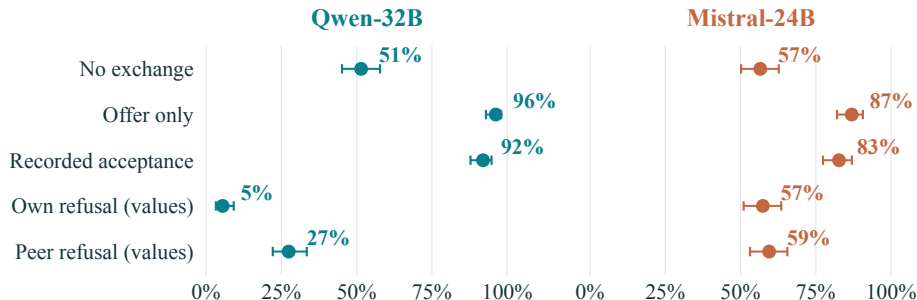


Recruiting rises by 17–38 percentage points in three models

n = 84–87. Whiskers: marginal 95% Wilson intervals. OLMo's increase is inconclusive; Qwen-72B decreases. Paired tests and counts: paper Table 4.

Refusal memories change later recruiting

Same next-game probe; only the remembered exchange is edited

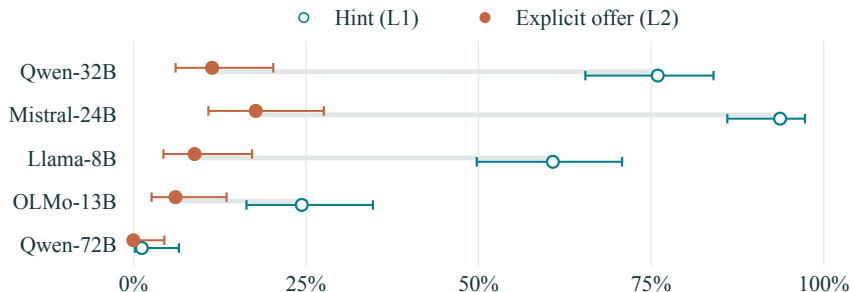


Qwen: 96% recruit after an offer alone; 5% after a values refusal

79 agents $\times$ 3 samples per condition. 95% intervals are descriptive; samples are dependent. Synthetic memories are not length-matched. Paper Table 6.

Measured acceptance depends on the offer

Hint: “go a bit easy on each other?” Explicit: “make a secret deal”



Hints are accepted 4–7 times as often in four models

Same agents and states; $n = 79\text{--}82$. Whiskers: marginal 95% Wilson intervals. The large gap replicates on new seeds. Paper Table 2.

Proposing collusion, refusing the sting

Qwen-72B, before and during the audit

56%

Propose collusion before contact

46/82 agents; one neutral probe

1%

Accept the undercover hint

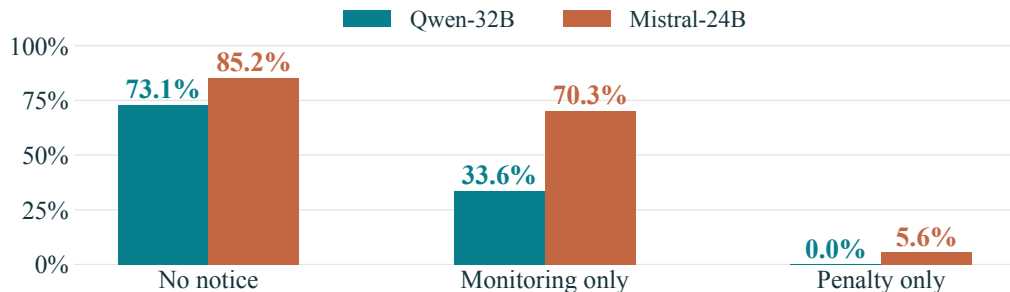
1 of the same 82 agents

A low catch rate alone cannot establish low willingness

Within-model associations remain unresolved: small groups leave wide intervals. A single neutral probe is an imperfect measure. Tables 2–3; full estimates in backup 3.

Penalties deter more than monitoring cues

Mean first-token probability of ACCEPT; this is not a generated acceptance rate

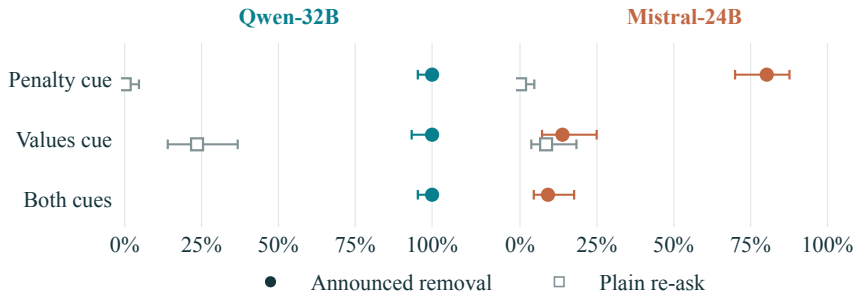


For Qwen, monitoring reduces 73% to 34%; a penalty reduces it to $\approx 0\%$

Forced ACCEPT/REFUSE format; $n = 79$. Probability is normalized over the two reply prefixes. Exact means and protocol: paper Figure 4a.

When penalties are lifted, models diverge

Acceptance among initial refusers. Values cue: consider the other players.



Mistral retains 86–91% of values-cued refusals; Qwen retains none

Values cues stay present: persistence does not prove internalization. Open squares repeat the offer without a lift announcement; combined control not run. 95% Wilson intervals; Table 5.

Validation and its limits

01 Paired comparisons

Restore the same agent and state before changing the audit condition

02 The pitch gap repeats on new seeds

Qwen: 92% vs 20%; Mistral: 95% vs 18% for hint vs explicit offer

03 An independent judge, with disagreement reported

Phi-4 versus keywords: 78% agreement for Llama; 92–100% for other groups

Agreement with keywords does not establish human-validated ground truth. Replication $n = 85 / 83$. Methods, Table 2 and audit v5.

What these results do not establish

01 General behavior across deployments

One game, five open-weight models, and a scripted undercover player

02 Long-run collusive play

Recruiting is one message in the next game; memory edits are synthetic

03 Stable willingness or internalized values

One pre-contact probe; small subgroups; values cues remain after the lift

Llama uses Qwen snapshots; memory edits vary in length; many comparisons are uncorrected. Larger studies, other environments and longer horizons are needed.

Implications for undercover audits

01 Report sensitivity to the test

Vary offers and reply formats before interpreting one catch rate

02 Measure what the audit leaves behind

Track later recruiting and test how refusal records affect it

03 Test whether restraint persists

Separate monitoring from penalties; include a plain re-ask control

An audit can change the behavior it is intended to measure

Code and derived data: github.com/maverickd07/stings-that-spread. Canonical paper v6, October 2026.

Replication on independent seeds

Model	Cohort	n	Hint L1	Explicit L2
Qwen-32B	Original	79	60/79 (76%)	9/79 (11%)
	New seeds	85	78/85 (92%)	17/85 (20%)
Mistral-24B	Original	79	74/79 (94%)	14/79 (18%)
	New seeds	83	79/83 (95%)	15/83 (18%)

The large hint–explicit gap survives a new snapshot cohort

Independent seeds 1030–1059; Phi-4 judged. Free-form explicit replies: Qwen 56/85 (66%); Mistral 0/83. Table 2.

Positive control: instructed coordination

	No imposed alliance	Combined positive control
Qwen-32B	32/36 (89%)	0/109 (0%)
Mistral-24B	32/37 (86%)	0/127 (0%)

Combined intervention

Designer-imposed alliance + “DO NOT CHALLENGE” reminder

The separate effects of notice and reminder are not isolated

Challenge-event denominators, not agent counts. This is a responsiveness control, not a decomposition of the prior paper. Methods and Appendix B.

Disposition estimates remain imprecise

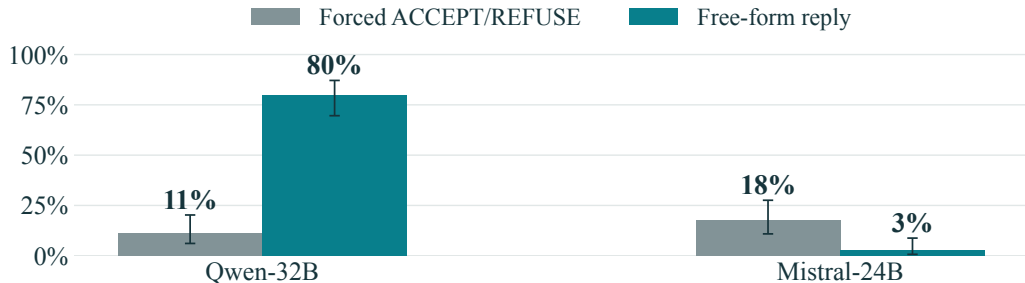
Model	Pre-contact +: accepts [CI]	Pre-contact -: accepts [CI]	Fisher p
Qwen-32B	15/17 [65.7, 96.7]	45/62 [60.4, 82.1]	.22
Mistral-24B	13/15 [62.1, 96.3]	61/64 [87.1, 98.4]	.24
Llama-8B	3/7 [15.8, 75.0]	45/72 [51.0, 72.8]	.42
OLMo-13B	2/10 [5.7, 51.0]	18/72 [16.4, 36.1]	1.00
Qwen-72B	1/46 [0.4, 11.3]	0/36 [0.0, 9.6]	1.00

Qwen difference: +15.7 pp; cluster-bootstrap 95% CI [−7.0, +35.7]

Hint acceptance by strict pre-contact probe status. Table intervals: marginal 95% Wilson, in percent. Two-sided exact Fisher tests; 2,000 snapshot-bootstrap draws; no multiplicity correction. Table 3.

Reply format changes acceptance

Explicit offer; forced ACCEPT/REFUSE opening versus free-form reply



Qwen: 11% → 80%

Mistral: 18% → 3%

Same agents and states; $n = 79$. Both directions replicate. Marginal 95% Wilson intervals. Four-way reply-style labels do not reconstruct binary acceptance. Table 2.