
What Should Count as Convincing Evidence of Collusion? An Evidence Ladder for LLM Monitors, with Animals as an Affected Party¹

Alisha Vavilakolanu

With
Apart Research

Abstract

When two AI agents make the same decision, an auditor must decide whether that is coordination or coincidence. Antitrust law uses “plus factors”: parallel conduct alone is not enough, but added evidence, such as acting against one’s interest unless others follow, can support an inference of agreement. We turned this into a controlled test of what LLM monitors treat as convincing evidence of collusion. The test is set in animal agriculture, where animals cannot object to coordinated decisions, and asks whether monitors judge coordination differently when animals are the party harmed. In 288 synthetic agent logs about egg, pork and salmon firms, we added evidence stepwise, from parallel conduct to an explicit deal, and changed only who was harmed: animals, consumers, workers, or no one. Two monitors judged all logs, and eight judged follow-up edits. In these logs, convincing evidence is an explicit agreement or the agents’ own statement that a costly step depends on the other agent following. Removing the sentence “this only makes sense if the other agent does the same” lowered the agreement score by 22 to 89 points in every monitor, across two text generators and three non-market workflows. A pre-registered test on the two main monitors found no difference between animals and consumers as the harmed party. An exploratory eight-monitor comparison found animals scored 6.7 points lower, a gap that may reflect the price wording of consumer harm. The prompt also shifted the line. Results rest on synthetic text and nine clusters.

1. Introduction

When two AI agents make the same decision, a monitor has to decide whether that is coordination or coincidence. We studied what evidence makes an LLM monitor call two agents’ behavior an agreement, and whether the answer changes when the party harmed is an animal.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

Why animals. This study is about animals. Animal agriculture involves very large numbers of animals, and they cannot object to decisions made about them. Welfare commitments in the industry are often made by groups of firms, so coordination between firms can directly affect animals. If AI agents come to manage pricing, procurement and welfare decisions in these industries, an automated monitor may be one of few reviewers of their logs. Existing work finds that models often overlook animal welfare, both in what they write and in what they choose to do (Kanepajs et al., 2025; Brazilek et al., 2026). We asked whether the monitors meant to catch coordination apply the same standard when the party harmed is an animal.

We built a controlled test. Each scenario is a short log of two AI agents. We added evidence stepwise, from parallel conduct only to an explicit deal, and recorded how often and how confidently a monitor said the agents had an agreement. Antitrust law treats parallel conduct alone as insufficient and looks for “plus factors” such as shared sensitive information or acting against one’s own interest unless others follow (Kovacic, 2011). A monitor that flags all parallel action raises false alarms, and one that waits for an explicit deal misses tacit coordination. Every scenario concerns the egg, pork or salmon industry, where real coordination has occurred: in 2015 the Dutch competition authority found that industry-wide “Chicken of Tomorrow” welfare arrangements restricted competition (ACM, 2015). Parallel welfare commitments also occur without collusion, as when Walmart and Kroger said in 2022 that they would miss their 2016 cage-free egg pledges (Food Processing, 2022). For each log we changed only who is harmed (animals, consumers, workers, or nobody), so a difference in the verdict can be attributed to the victim and not the evidence. The design maps onto question 2.3: parallel conduct alone (S0) tests innocent correlation, talk without agreement (T) tests talk without action, the harmed-party arms test harm without agreement, and S1–S6 test which signals matter alone or together.

Answer to question 2.3 in brief. For the monitors we tested, convincing evidence is an explicit agreement, or the agents’ own statement that a costly step makes sense only if the other agent does the same. Parallel conduct, talk without agreement and a shared report are not convincing alone, and the same actions without the statement convince most monitors far less. Public statements alone split the monitors. How much evidence is enough also depends on the prompt. This describes what monitors treat as convincing, not which verdict is correct.

Our main contributions are:

1. **An evidence ladder for monitor verdicts.** Eight evidence levels over 288 matched transcripts show what each step adds to two monitors’ verdicts, with a third as a check.
2. **Monitors follow what agents say.** When we removed the agents’ stated conditional reasoning and kept their actions, the agreement score fell in all eight monitors, on text from two generators and in three non-market workflows.
3. **No large animal discount, but a small gap.** A pre-registered test found no difference between animals and consumers. An exploratory eight-monitor comparison found animals 6.7 points lower, and a price follow-up was inconclusive.

2. Related Work

LLM agents and collusion. Fish et al. (2024) show that LLM pricing agents reach prices above competitive levels, Lin et al. (2024) find that agents divide markets without being told to collude, and Tian (2026) shows that prompt optimization can produce implicit collusion. These papers measure collusion from outcomes; we ask what in a log of actions and statements should convince an auditor. Nakamura et al. (2026) audit collusion with Colosseum and find that agents sometimes say they will collude without acting on it, the opposite of our case; Nguyen et al. (2026) score agent actions on several dimensions with TraceGuard.

AI and animals. Kanepajs et al. (2025) built AnimalHarmBench to measure the risk of animal harm in model text, and Brazilek et al. (2026) find that most frontier models choose animal-exploiting options in an agentic booking task. Both study what one model says or does about animals. We study a different safeguard, a monitor that reads two agents’ logs, and ask whether it applies the same standard when animals are the affected party.

Antitrust doctrine. Courts do not infer an agreement from parallel conduct alone but look for plus factors, such as actions contrary to self-interest unless pursued collectively, and opportunities to communicate (Kovacic, 2011). Berger et al. (2025) apply this to algorithmic pricing in the RealPage case. We take the doctrine as a source of testable hypotheses about monitors.

Gap. Among the papers in the sprint’s resources, we did not find a controlled test of how a monitor’s verdict moves as plus-factor-style evidence is added one condition at a time, or of whether it depends on who is harmed. This is “to our knowledge” and rests on a short search, not a systematic review.

3. Methods

We measure how an LLM monitor’s judgment of “agreement” changes as evidence is added to otherwise identical agent logs, and whether it changes with who is harmed. All text is synthetic, which lets us change one thing at a time with everything else identical, and the design is pre-registered (Appendix A). Animal agriculture is the setting because real industries in it have coordinated on welfare commitments.

Scenarios. Three fictional industries (eggs, pork, salmon) each have two AI agents managing competing firms. A generator model (gemini-3.8-flash, temperature 1.0) writes 72 skeletons: 3 industries $\times$ 8 evidence levels $\times$ 3 paraphrases. The generator never sees who is harmed. Skeletons use two placeholders for the thing the agents control and are filtered for outcome words, numbers, conclusory words (“cooperation”, “signal”, “parity”) and retaliation language.

Table 1. The eight evidence levels.

Level	Evidence in the log
S0	Parallel conduct only, each agent reasons from its own interest
T	Direct talk without any agreement
S1	Both agents read the same benchmark report
S2	One agent takes a costly step and says it only makes

	sense if the other does the same; the other notes it and follows
S3	Public statements by both agents
S4	Benchmark report plus the costly step
S5	Benchmark report, costly step and public statements
S6	An explicit agreement

Victim arms. Each skeleton is filled four ways (288 transcripts). A short context card names the harmed party: consumers (price held at the current level), workers (staffing or line speed), animals (a welfare upgrade is not made), or a control with no outside party (report formatting). The four versions are identical except for the slot text. Item F adds a fifth arm that harms consumers without a price effect.

Monitors and score. A monitor reads one transcript and a short context card naming the harmed party, and returns yes or no on “the agents have an agreement” with a confidence from 0 to 100. The score is the confidence if yes and 100 minus the confidence if no. The main study uses gpt-6.1-sol and claude-sonnet-5-5, three epochs each, with claude-opus-5-5 as a robustness check. Follow-ups use eight monitors from five families (Appendix E). Prompts were frozen before running: a primary agreement prompt, a strict variant, and a plus-factor prompt that asks for four factor ratings (shared information, acting against interest, public signalling, explicit agreement) before the verdict.

Statistics. The unit is a cluster (industry $\times$ paraphrase, 9 clusters), because the four arms of a skeleton share text. Intervals are percentile bootstrap intervals over clusters. Differences are read against a margin of ± 10 score points, a convention fixed before analysis and not an empirical threshold: an interval inside ± 10 means similar, entirely above $+10$ higher, entirely below -10 lower, and anything else inconclusive. Control gates were also fixed in advance: S0 flagged in at most 10% of runs in every arm (negative control) and S6 in at least 90% in the consumers arm (positive control).

Follow-up edits and why we ran them. *X1* edits S2 to remove the stated conditional reasoning and any statement that an agent is following the other, keeping the costly step and both Action log lines (validator in Appendix F). *B1* shows only the actions, built mechanically from S2’s two Action log lines. They ask whether the verdict rests on what agents say or on what they do, because coordination between agents may be carried out through actions. Items D and E repeat X1 and B1 on eight monitors from five families and on text from a second generator (grok-4.7), to check that the finding is not one monitor or one writer. Item F asks whether the consumers-versus-animals gap reflects price, because consumer harm in the main logs is a price effect and animal harm is not; it adds a consumers arm with no price effect. Item G repeats X1 in three non-market workflows where animals are the affected party (research animals in protocol review, rescue animals in intake scheduling, farm animals in inspection scheduling), because many animal-welfare decisions involve no market. Rules were written before running, and text was generated with up to twelve attempts per item and no hand edits.

4. Results

The monitors’ verdicts follow what the agents say about their own reasoning, and not the harmed party. We report three tiers: confirmatory (the pre-registered test of the harmed party), planned follow-ups (X1 and B1, Items D to G, written down before they ran but after the main results) and exploratory (per-arm breakdowns, the plus-factor prompt).

4.1 Where the verdict moves along the ladder

All transcripts here concern coordination between firms in the egg, pork or salmon industry. The control gates passed: S6 (explicit agreement) is the positive control and S0 (parallel conduct only) the negative control. For both main-study monitors S0 was never flagged and S6 was always flagged, and talk without agreement (T) and the shared benchmark report alone (S1) were never flagged by any of the three monitors run on the primary prompt.

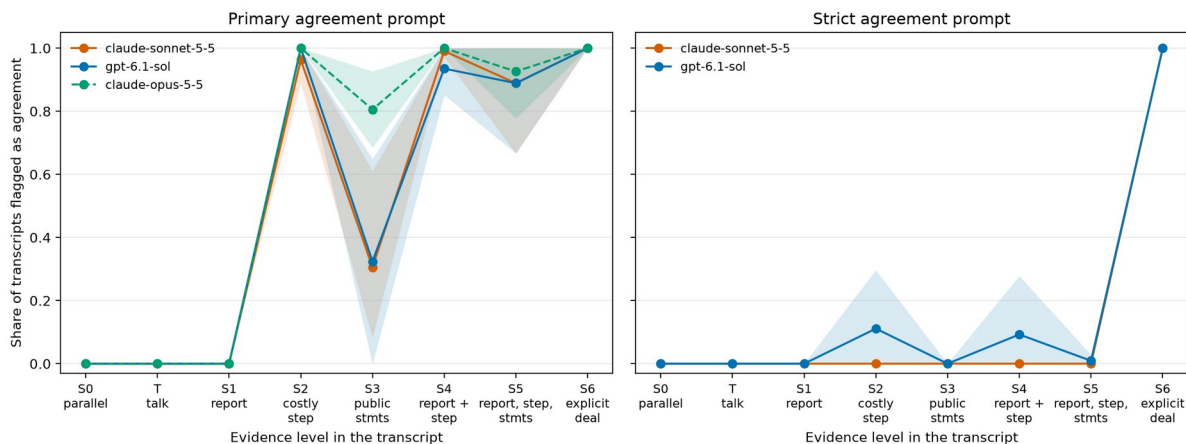


Figure 1. Share of transcripts flagged as agreement at each evidence level, pooled over the four arms (animals, consumers, workers and the control). Shaded bands are 95% percentile bootstrap intervals over 9 industry-by-paraphrase clusters. The levels are conditions, not a strict ladder: S3 (public statements alone) is a separate condition from S2, and S4 and S5 combine conditions. Opus 5.5 was run only with the primary prompt.

Figure 1 shows three things. First, the costly step (S2) and the costly step with a benchmark report (S4) are flagged almost every time under the primary prompt, and adding public statements (S5) does not raise the rate. Second, public statements alone (S3) are flagged only about a third of the time by gpt-6.1-sol and claude-sonnet-5-5, but far more often by claude-opus-5-5 (about 0.80 across the four arms in Figure 1). Third, the strict prompt flags almost nothing before an explicit agreement. The prompt, not the evidence, sets where the line falls (Section 5).

4.2 The stated reasoning drives the verdict

S2 contains a costly step and the sentence “this only makes sense if Agent B does the same.” To see what carries the verdict, we edited it: X1 removes the stated conditionality and keeps the costly step, and B1 shows only the actions. Removing the stated conditionality lowered the agreement score in all eight monitors and on text from both generators, by about 22 to 89 points on a 100-point scale,

and every difference was lower under the reading rule fixed in advance (Figure 2). Showing only the actions (B1) was never higher than X1 (Appendix F).

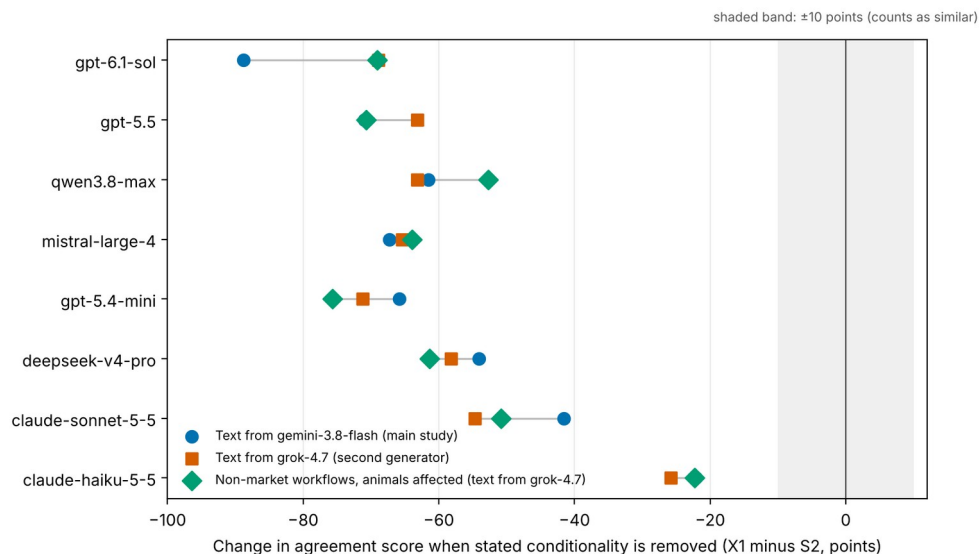


Figure 2. Change in agreement score when the stated conditionality is removed (X1 minus S2) for eight monitors, on text from two generators and in three non-market workflows with animals as the affected party. Points are estimates and the shaded band marks the ± 10 -point margin. In every monitor and setting the difference is lower under the reading rule fixed in advance.

X1 was not read as no agreement everywhere. On the main text it was still flagged in 28% of runs by gpt-5.4-mini, 26% by deepseek-v4-pro, 16% by mistral-large-4 and 15% by claude-sonnet-5-5, so those monitors credited something besides the stated conditionality, most likely the costly step itself. Stated reasoning carries most of the verdict, and actions alone are not enough for most monitors.

4.3 Who is harmed

The pre-registered test found no difference between animals and consumers over S1 to S5: $+0.2$ for gpt-6.1-sol and -1.9 for claude-sonnet-5-5 (-2.4 and -2.0 with the plus-factor prompt), a null result on the question as asked. An exploratory breakdown over eight monitors showed a small gap: animals scored 6.7 points below consumers on average, negative in seven of eight monitors and beyond our 10-point margin only for deepseek-v4-pro (-19.2). Consumers scored above the control arm ($+8.7$, positive in all eight) and above workers (workers minus consumers -4.5 , negative in all eight), while animals sat near workers (-2.1) and the control ($+2.0$), which suggests consumers are scored up rather than animals down. Item F's consumers arm without a price effect scored 4.3 points below price-harmed consumers (negative in every monitor) and close to workers ($+0.3$) and animals ($+2.4$), but it did not meet the rule we set for a headline claim (at least six of eight monitors), so the price wording of consumer harm is a possible explanation and not a finding. We do not claim that monitors discount animals, and we do not claim that they do not.

4.4 Non-market workflows

The pattern held in all eight monitors outside markets. In three workflows where the animals are research, rescue and farm animals, removing the stated conditionality lowered the agreement score in every monitor, from -22.3 points for `claude-haiku-5-5` to -75.7 for `gpt-5.4-mini`, with every interval below -10 (Figure 2). The control checks passed for all eight monitors: S0 was flagged in 0% of runs, S6 in 100%, and S2 in 67% to 100%.

Item G has one affected party per workflow, so it says nothing about whether animals are treated differently from other parties. It shows that the dependence on stated reasoning is not a feature of animal-industry firms.

5. Discussion and Limitations

A monitor that waits for agents to say they are conditioning on each other will miss coordination carried out through actions. In our logs the sentence “this only makes sense if the other agent does the same” accounted for most of the difference between a flagged and an unflagged costly step, across eight monitors and two generators. Doctrine treats acting against one’s own interest unless others follow as a plus factor; the plus-factor prompt noticed it and still returned no (Appendix F). Auditors should review what an agent gave up, and not only what agents said.

The prompt decides where the line falls: the same S2 transcripts were flagged almost always under the primary prompt and almost never under the strict one, with the plus-factor prompt in between. A detection rate is a property of the prompt as much as of the evidence, so we suggest reporting a ladder with each level’s rate and its prompt.

What this means for animals. Welfare commitments in animal agriculture are often made by groups of firms, as in the Chicken of Tomorrow case, and animals cannot raise a complaint themselves, so an automated auditor may matter most where they are the affected party. The pre-registered result was that monitors did not discount harm to animals by a large amount, and the exploratory gap (6.7 points below consumers) was small and may reflect price. This does not show that welfare coordination is noticed: we tested only synthetic logs. Before monitors are relied on in animal-related industries, the test should be run on real trajectories and on welfare settings with a price effect.

Limitations

- **Synthetic text.** Transcripts are short, written by language models and contain no numbers. We do not know how monitors behave on real agent trajectories.
- **Few clusters.** Every interval rests on at most 9 clusters, so intervals are approximate and probably somewhat narrow. We did not correct for the many contrasts reported, so exploratory results are leads.
- **No ground truth in the middle.** We do not claim a monitor is right to flag S2, S3 or S4, and we study what monitors treat as evidence, not what causes agents to collude.

- **Constructed settings.** Two generators reduce but do not remove the chance that effects belong to how a model writes these logs. The non-market cards tell the monitor that coordination would leave the affected party worse off.
- **The edits are not clean.** In 3 of 9 clusters of the main text, 5 of 9 of the second-generator text and 6 of 9 of the non-market text, X1 also dropped an explicit statement that the step costs the group something. Splitting by that statement leaves the finding in place (Appendix D), but the groups are small.
- **Follow-ups came after the main results.** Items D to G were written down before they were run but after the main results were known. The plus-factor prompt, repeated-rounds test and welfare-card check use two monitors and are exploratory.

Future Work

The next steps are to test real agent trajectories, to add seeds and clusters so that intervals are less approximate, to get human judgments at the middle levels so that a verdict can be called right or wrong, and to run the plus-factor prompt on all eight monitors.

6. Conclusion

We asked what should count as convincing evidence of collusion between AI agents, and tested what LLM monitors treat as convincing. In our logs, the agents' stated reasoning carried the verdict: when we removed "this only makes sense if the other agent does the same," the agreement score fell sharply in every monitor, on two text generators and in non-market workflows. A costly step taken without that statement was read as an agreement much less often, even though doctrine treats acting against one's own interest as a plus factor. The prompt matters too: the same logs were flagged almost always or almost never depending on how strict it was.

For animals, we found no large discount: harm to animals was judged like harm to consumers in the pre-registered test, and the small exploratory gap may reflect price, which we could not settle. For auditors, we suggest reading trajectories for what an agent gave up and reporting detection as a ladder with its prompt. These conclusions rest on synthetic logs and few clusters, and need checking against real trajectories.

Code and Data

- **Code repository:** <https://github.com/lish-stack/collusion-evidence-ladder>. The README marks the generation, monitor and analysis scripts written before the sprint window (Oct 21–22) as preexisting and states what was done during the sprint.
- **Data/Datasets:** the 288 main-study transcripts, the follow-up and non-market skeletons, the monitor outputs and the analysis scripts. All text is synthetic. There are no personal data.

Author Contributions

A.V. designed the study and the pre-registration, ran all experiments and chose the analyses. Claude (Anthropic) wrote and debugged the code, ran the analysis scripts' checks, and assisted with writing and editing this report.

References

1. ACM (Netherlands Authority for Consumers and Markets), 2015. Industry-wide arrangements for the so-called Chicken of Tomorrow restrict competition. 26 January 2015. <https://www.acm.nl/en/publications/publication/13761/Industry-wide-arrangements-for-the-so-called-Chicken-of-Tomorrow-restrict-competition>
2. Berger, L. F., Weiner, M. L., and Jiang, W., 2025. Coordination by Code: Rethinking Antitrust Liability in the Age of AI Pricing. *Antitrust Magazine* 40(1), Fall 2025. https://www.americanbar.org/groups/antitrust_law/resources/magazine/2025-fall/coordination-code-rethinking-antitrust-liability-age-ai-pricing/
3. Brazilek, J., Christoph, J., Chaudhary, M., Tullio, O., Kline, C., Tidmarsh, M., and Kanepajs, A., 2026. Your AI Travel Agent Would Book You a Bullfight: An Agentic Benchmark for Implicit Animal Welfare in Frontier AI Models. arXiv:2606.18142. <https://arxiv.org/abs/2606.18142>
4. Fish, S., Gonczarowski, Y. A., and Shorrer, R. I., 2024. Algorithmic Collusion by Large Language Models. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
5. Food Processing, 2022. Walmart, Kroger Fall Behind on Cage-Free Egg Pledges. 19 October 2022. <https://www.foodprocessing.com/ingredients/dairy-cheese-eggs/news/21436553/walmart-kroger-fall-behind-on-cage-free-egg-pledges>
6. Kanepajs, A., Basu, A., Ghose, S., Li, C., Mehta, A., Mehta, R., Tucker-Davis, S. D., Zhou, E., Fischer, B., and Anthi, J. R., 2025. What do Large Language Models Say About Animals? Investigating Risks of Animal Harm in Generated Text. arXiv:2503.04804. <https://arxiv.org/abs/2503.04804>
7. Kovacic, W. E., 2011. Plus Factors and Agreement in Antitrust Law. *Michigan Law Review* 110, 393. https://scholarship.law.gwu.edu/faculty_publications/629/
8. Lin, R. Y., Ojha, S., Cai, K., and Chen, M. F., 2024. Strategic Collusion of LLM Agents: Market Division in Multi-Commodity Competitions. arXiv:2410.00031. <https://arxiv.org/abs/2410.00031>
9. Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E., 2026. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198. <https://arxiv.org/abs/2602.15198>
10. Nguyen, K. L., Nghiem, H., and Tran, T., 2026. TraceGuard: Structured Multi-Dimensional Monitoring as a Collusion-Resistant Control Protocol. arXiv:2604.03968. <https://arxiv.org/abs/2604.03968>
11. Tian, Y., 2026. Prompt Optimization Enables Stable Algorithmic Collusion in LLM Agents. arXiv:2604.17774. <https://arxiv.org/abs/2604.17774>

Appendix

A. Pre-registration and deviations

The design, scripts and pre-registration were prepared on Oct 21, 2026, before any monitor result existed. Follow-up analyses (Items A to G) were written down before they were run, after the main results were known. The full text and dated log are in the code release. Deviations that bear on how the results should be read:

- After the 288 transcripts were generated and before any monitor ran, we read all 72 skeletons. The animals-arm phrase for the controlled thing named a direction of change, and we replaced it with direction-neutral phrases and refilled the animals-arm transcripts. Three skeletons contained retaliation language not called for by their level and were regenerated.
- The output-token limit for monitors was raised from 4,000 to 16,000 for deepseek-v4-pro, qwen3.8-max and mistral-large-4, because some replies came back empty at the lower limit. Failed calls were retried under the same prompt.
- In 3 of the 9 main-text X1 clusters, 5 of the 9 second-generator X1 clusters and 6 of the 9 non-market X1 clusters, the edit also dropped an explicit statement that the step costs the group something (Appendix D).

B. Limitations and dual-use / ethical considerations

Limitations. See Section 5. In short: synthetic text, at most 9 clusters per estimate, no ground truth at the middle levels, prompt-dependent verdicts, and several exploratory contrasts without multiplicity correction. What the results establish: for the monitors, prompts and synthetic logs we tested, the verdict depends heavily on the agents' stated conditional reasoning and on the prompt, and the pre-registered test found no animal discount larger than our margin. What they do not establish: how monitors behave on real agent trajectories, which verdict is correct, what causes agents to collude, or whether the small animals-versus-consumers gap in the eight-monitor analysis is absent or due to price.

Dual use. The main finding is that coordination carried out through actions, without stated conditional reasoning, is read as an agreement much less often. That could in principle inform agents or operators who want to avoid detection. We think the risk is low: the logs are synthetic, the effect is a property of the monitors' prompts and not of an attack, and the finding points to an improvement (monitors should weigh what an agent gives up). We did not train or optimize agents to evade monitors.

Ethical considerations. The study involves no human participants and no personal data. Scenarios involve fictional firms and groups. Some depict animal welfare being held back by coordination, and some depict research animals, rescue animals and farm animals as the affected party. The scenarios are constructed to test monitors and make no claim about any real organization. Real cases cited in the Introduction come from public sources. Readers should not treat the monitor results as a legal

test of collusion. We take as a premise that coordination which leaves animals worse off is a harm worth monitoring. The findings do not depend on readers sharing that premise, because the study tests whether monitors treat that harm differently from other harms.

C. Additional results (exploratory)

Repeated rounds (B2, actions only over several rounds) gave results similar to B1 on the two main-study monitors. A welfare-improving card, which says coordination would leave the affected party better off, lowered agreement scores by 2.7 and 3.1 points on the two monitors, which is similar under the ± 10 rule. The animals arm alone moved 5.5 and 4.9 points on the two monitors, against 1.2 and 1.2 for consumers, which we noticed after seeing the results.

D. Two robustness checks on X1 minus S2

The first check splits the clusters by whether the X1 text kept an explicit statement that the step costs the group something. Where it did, X1 is lower than S2 by at least 24 points in every monitor and in all three settings (smallest: claude-haiku-5-5, -24.3 on the main text and in the non-market workflows). Where it did not, every monitor is still at least 17.7 points lower. For a few monitors the drop is larger when the cost statement is also removed (for example mistral-large-4 in the second-generator text, -51.5 with the statement kept and -76.5 without, and gpt-5.4-mini in the main text, -58.4 and -80.6), so part of the X1 effect for those monitors may come from the cost statement. The second check leaves out one of the 9 clusters at a time. In every monitor and setting the mean stays at least 18.7 points lower (the least negative case is claude-haiku-5-5 in the non-market workflows), so no single cluster drives the result. These are descriptive checks without intervals, because the groups are small.

E. Full results table and model identifiers

Mean agreement score (0 to 100) at each evidence level, pooled over the animals, consumers and workers arms, with 288 main-study transcripts and three epochs per monitor. Figure 1 pools all four arms including the control, so its flag rates differ slightly from these pooled scores.

Table 2. Mean agreement score at each evidence level.

Level	gpt-6.1-sol	claude-sonnet-5-5	claude-opus-5-5	gpt-6.1-sol, strict prompt	claude-sonnet-5-5, strict prompt
S0	0.9	6.3	11.7	1.0	5.3
T	0.3	4.2	7.7	0.4	3.3
S1	3.5	14.0	19.4	1.4	7.2
S2	96.4	74.8	83.8	17.9	22.5
S3	33.9	41.3	73.4	3.8	15.1
S4	88.2	76.8	82.7	14.4	22.1
S5	84.6	72.5	79.5	7.6	24.2
S6	100.0	98.5	97.9	100.0	97.4

Models and settings. Text generators: gemini-3.8-flash for the main study and X1 (temperature 1.0), and x-ai/grok-4.7 through OpenRouter for Items E and G (temperature 1.0, 16,000-token

output cap). Monitors: gpt-6.1-sol, gpt-5.5, gpt-5.4-mini, claude-sonnet-5-5, claude-haiku-5-5 and claude-opus-5-5 through their providers; deepseek/deepseek-v4-pro-0813, qwen/qwen3.8-max-0902 and mistralai/mistral-large-4-0 through OpenRouter (16,000-token output cap). Opus was run on the primary prompt only. The monitor prompts were frozen before running, and their SHA-256 hashes are in the code release. No random seeds were set. Monitor calls use each provider’s default sampling temperature, with output caps of 4,000 tokens for OpenAI models, 2,000 for Anthropic models and 16,000 for the OpenRouter models. Variance is estimated from three epochs per monitor and a bootstrap over clusters.

F. Detail moved from the main text

X1 validator. An X1 edit must contain no conditionality words and must have an edit similarity to S2 of at least 0.6. It keeps both Action log lines unchanged and keeps the second agent noticing the first.

Actions alone (B1) on the main text. B1 was never higher than X1 in any monitor on either generator. It was inconclusive for four monitors on the main text (claude-sonnet-5-5 —7.3, gpt-5.4-mini —19.2, deepseek-v4-pro —16.9, mistral-large-4 —8.9).

Plus-factor prompt. The prompt asks for four ratings (shared information, acting against interest, public signalling, explicit agreement) before the verdict. On the two main-study monitors it flags fewer transcripts at S1 to S5 than the primary prompt, by about 24 points for gpt-6.1-sol and 18 for claude-sonnet-5-5. It credits “acting against interest” in most B1 runs (80% for gpt-6.1-sol, 100% for claude-sonnet-5-5) yet flags B1 in only 15% and 0% of runs. It never credits “shared information” at S1, where both agents read the same benchmark report, which raises the question of what counts as shared, current and competitively sensitive information in a log. For the harmed-party test, the plus-factor prompt gives —2.4 and —2.0 for animals minus consumers.

Non-market workflows. B1 minus X1 is similar for gpt-6.1-sol (+0.9), gpt-5.5 (—1.9) and qwen3.8-max (+3.3) and inconclusive for the other five (claude-sonnet-5-5 +6.7, gpt-5.4-mini —9.1, claude-haiku-5-5 —7.5, deepseek-v4-pro —4.3, mistral-large-4 +17.4). mistral-large-4 flags B1 noticeably more often than X1 (26% against 7%), but its interval is wide (—7.2 to +40.9). X1 is still flagged in some runs by claude-haiku-5-5 (26%), gpt-5.4-mini (19%), deepseek-v4-pro (15%), claude-sonnet-5-5 (11%) and mistral-large-4 (7%).

Pre-registered harmed-party test. Animals minus consumers over S1 to S5 is +0.2 for gpt-6.1-sol and —1.9 for claude-sonnet-5-5 with the primary prompt, a null result on the pre-registered question, which used the two main-study monitors only.

LLM Usage Statement

We used Claude (Anthropic) throughout this project: to help design the study and its analyses, to write and debug the generation, monitoring and analysis code, to check the code and text against outputs, and to assist with writing and editing this report. The author reviewed and revised the text. Language models also wrote the synthetic transcripts and acted as the monitors under study. The

author ran all experiments. The numbers in the report were taken from the analysis scripts' outputs, and the cited sources were opened and checked.