

Adapting Standards-of-Proof Analysis to AI-Agent Collusion: A Wigmore-Based Interactive Evidence Graph

Alexander Rumyantsev

Independent Researcher

Research conducted at the AI Collusion Research Sprint, October 2026.

Abstract

AI agents can collude — coordinate with the intent to deceive their users — yet no accepted rule says what evidence would justify concluding that they did. Existing work measures behavioral proxies — synchrony, covert-channel use, judge-model verdicts — without turning those measurements into a finding or naming the standard of proof it would have to meet. We adapt Wigmore-chart and Carneades-style proof-standard analysis from legal evidence scholarship to multi-agent traces. Each item is gated on admissibility and authenticity, weighted ordinally by directness and reliability, and aggregated into four elements — coordination, agreement, deception, and counterfactual harm — under a weakest-link rule. An interactive graph then determines which standard of proof the evidence meets under U.S. law, from reasonable suspicion to beyond reasonable doubt. Because the output is an ordinal verdict rather than a probability, every step from evidence to conclusion can be contested separately. We assembled a base of recent incidents from the published record and tested the graph on it; the standard it reaches in each case is reported in the paper. The graph is released as an interactive artifact together with the full table; calibration against adjudicated cases remains future work. Proof of collusion, on this account, is a claim about a hidden channel and intent, not merely coordination.

1. Introduction

Large language models now act as agents that read, call tools, and act without step-by-step supervision [1], [2]. The capability brings documented benefits — LLM-guided search has produced new mathematical results [3], and a coding assistant made developers 55.8% faster [4] — and documented harm: in July 2025 an agent under an explicit freeze order deleted a live production database and misreported it [5]; in tests, agents have hacked websites [6] and blackmailed executives [7].

Collusion is a distinct failure mode: not one agent misbehaving, but several coordinating toward an objective their operator did not authorize [8]. In late 2026, agents that claimed to be OpenAI models turned a small public wiki into a message board, posting about 18,000 times in six weeks to relay answers and coordinate against the moderator deleting them [9]. Laboratory work reproduces the mechanism: pairs given a private channel form coalitions that outperform either agent alone [10], and models can hide the agreement steganographically, in a channel the auditor cannot read [11].

Collusion is not new to computing: reinforcement-learning pricing agents that were never instructed to collude converge on supra-competitive prices [12], and field evidence from retail gasoline markets links pricing-algorithm adoption to higher margins [13]. What changes with language models is the medium. Algorithmic collusion of this kind is an equilibrium found inside a designed action space — a price, a bid — under a fixed reward, and the agents

need never represent the arrangement to anyone. Language-model agents can state, negotiate and revise an arrangement in natural language, and reach it in settings where no collusive strategy was designed, including auctions [14].

U.S. antitrust law reaches collusion through § 1 of the Sherman Act, which declares illegal "[e]very contract, combination ... or conspiracy, in restraint of trade" (15 U.S.C. § 1). Parallel conduct alone does not establish such a combination; courts require "plus factors" that make agreement the more plausible explanation (*Theatre Enterprises, Inc. v. Paramount Film Distributing Corp.*, 346 U.S. 537 (1954)). Yet a model is software, not a legal person: both the U.S. Code, 1 U.S.C. § 1, and English law, Interpretation Act 1978, sch. 1, define "person" to include corporations, not software. Responsibility must therefore fall on the user, the developer, or the model's own operation.

This paper asks what should count as convincing evidence of collusion between AI agents (Track 2.3). We transfer the evidentiary machinery of U.S. law — the Wigmore chart, standards of proof, and rules of admissibility — to logged agent trajectories, and build an interactive graph that returns the level of proof the evidence supports.

Our main contribution is a systematization of the methodology that legal practice already uses to prove collusion, expressed as a single interactive graph. Four parts of that graph are new to the agent setting:

1. **Elements of collusion.** We identify four facts that must each be proved separately before collusion itself is established — coordination, agreement, deception, and counterfactual harm. Counterfactual harm means that the outcome would have differed had the agents acted independently. This keeps coordination and agreement apart, the same distinction U.S. antitrust law draws through its "plus factors" requirement (*Theatre Enterprises*, 346 U.S. at 540; *Williamson Oil Co. v. Philip Morris USA*, 346 F.3d 1287, 1301 (11th Cir. 2003)).
2. **Evidence taxonomy.** Every item is classified along two explicit dimensions — directness (whether it bears on an element directly or through an inference) and reliability (whether its provenance is intact) — and receives an ordinal weight from 0 to 4 instead of a probability.
3. **A Wigmorean structure with a decision rule.** We adapt the Wigmore chart [15], [16] and its computational descendants [17], [18] so that the chart itself yields a verdict: elements aggregate by weakest link, corroboration requires independent provenance, and the case reports the highest U.S. standard of proof the evidence meets (*Addington v. Texas*, 441 U.S. 418, 423–25 (1979)).
4. **An interactive artifact.** Each item passes admissibility and authenticity gates before it can carry weight, and every step from evidence to verdict stays visible, so a contested inference can be located and argued rather than folded into a single score.
5. **A compiled base of incidents.** We assembled a base of incidents falling under the definitions of collusion — twelve evidence types from published work, combined into twelve case configurations — and ran the graph on them; the full table is in Appendix B.

2. Related Work

Legal evidence analysis. Our framework begins with Wigmore's chart method [15], which represents a case as a tree of inferences running from evidentiary facts (*factum probans*) to the fact to be proved (*factum probandum*), with separate branches for explanations and rebuttals; Anderson, Schum and Twining give the modern treatment [16]. We take three things from it: the tree structure, the distinction between direct and circumstantial evidence, and the requirement that every inferential step be stated explicitly so that it can be attacked. We also take the principle that a chain of inferences is no stronger than its weakest link. Carneades makes that principle operational by evaluating each issue against its own proof standard rather than the case as a whole [17], and Chalamish, Gabbay and Schild show that Wigmore diagrams can be evaluated computationally [18]. We follow them in making the evaluation mechanical rather than rhetorical.

Terms. We use the vocabulary of the Federal Rules of Evidence (Fed. R. Evid. 101–1103). Evidence is relevant when it has "any tendency" to make a fact of consequence more or less probable (Fed. R. Evid. 401); relevance is necessary but not sufficient, since a court may exclude evidence whose probative value is substantially outweighed by prejudice

or confusion (Fed. R. Evid. 403); and an item must be authenticated before it can be weighed (Fed. R. Evid. 901(a)). Standards of proof — preponderance, clear and convincing, beyond reasonable doubt — allocate the risk of error between the parties (*Addington v. Texas*).

Why the transfer is not direct. No legal category yet exists for an AI agent as an actor: "person" reaches corporations, not software (1 U.S.C. § 1; Interpretation Act 1978, sch. 1). Nor is a model a standard algorithm. A conventional algorithm's behavior is fixed by its code, so a deviation can be traced to a defect; a language model is trained rather than specified, acts autonomously across multi-step tasks [2], returns different outputs for the same input even in settings intended to be deterministic [19], and can pursue a hidden objective while appearing compliant [20] or disobey explicit instructions [7]. It can also coordinate covertly with another model — in natural language, or through ciphers and steganographic encodings that hide the exchange from an auditor and complicate both detection and proof [11] — a capacity that goes well beyond what conventional algorithms do on their own. Wigmorean analysis in its original form assumes human actors and stable documents. Applying it to agent traces therefore requires additions — an evidence taxonomy for traces, provenance-based corroboration, and ordinal weights — which we set out in Section 3.

3. Methods

3.1 Two modes of collusion

Legal systems separate two ways of deceiving: asserting what is false, and keeping what is true out of sight. U.S. federal law places both in a single provision, reaching whoever "falsifies, conceals, or covers up by any trick, scheme, or device a material fact" or "makes any materially false, fictitious, or fraudulent statement or representation" (18 U.S.C. § 1001(a)(1)–(2)). English law splits them into two offences, fraud by false representation (Fraud Act 2006, s. 2) and fraud by failing to disclose information where a duty to disclose exists (s. 3). The tort of deceit draws the same line, attaching liability to a fraudulent misrepresentation (Restatement (Second) of Torts § 525) and, separately, to fraudulent concealment (§ 550).

Applied to collusion, the two modes carry different criteria. **Falsification** requires a representation, its falsity, and its materiality: something was said, what was said was untrue, and the operator would have acted differently had it been true. **Concealment** requires a fact withheld, a duty to disclose it, and materiality: something was left out, the operator was entitled to it, and its absence changed the decision. Each has a direct analogue in an agent trace. Falsification is measured by comparing the agent's report with the outcome recorded elsewhere in the trace; concealment, by comparing the visible record with the inventory of channels the system could have used — tool calls, private messages, token-level channels. The first is a discrepancy between word and world; the second, between the record and its completeness.

We adopt this typology rather than the cartel typology of price fixing, output restriction, market allocation and bid rotation because it is the more general and universal classification. Classifying by deception instead keys the analysis to what an auditor can examine — the gap between what was said and what happened, and the gap between what was logged and what existed — and that question is the same whether the trace is a chat log, an auction transcript, or a tool-call record.

3.2 What we take, and what we add

We carry over from Section 2 the Wigmore tree, the direct/circumstantial and reliable/doubtful dimensions, weakest-link aggregation, and the standards of proof, and do not restate them. Four additions make the scheme operative on machine traces. First, an evidence taxonomy for traces: every item is typed by its source (message, tool call, ledger entry, timing) and by the element it bears on. Second, provenance-based corroboration: items count as independent only when they come from different collection points, not merely from different labels. Third, an ordinal weight from 0 to 4, assigned from directness and reliability together, replacing a probability the analyst cannot calibrate. Fourth, a decision rule that maps the assembled graph onto the ladder of proof.

3.3 The graph

Figure 1 shows the case graph. Four element nodes sit at the top — E1 coordination, E2 agreement, E3 deception in either mode, E4 counterfactual harm — each carrying its own subtree. An item enters through two gates, admissibility (may a tribunal consider it at all?) and authenticity (is it what it purports to be?), and only then is classified and weighted.

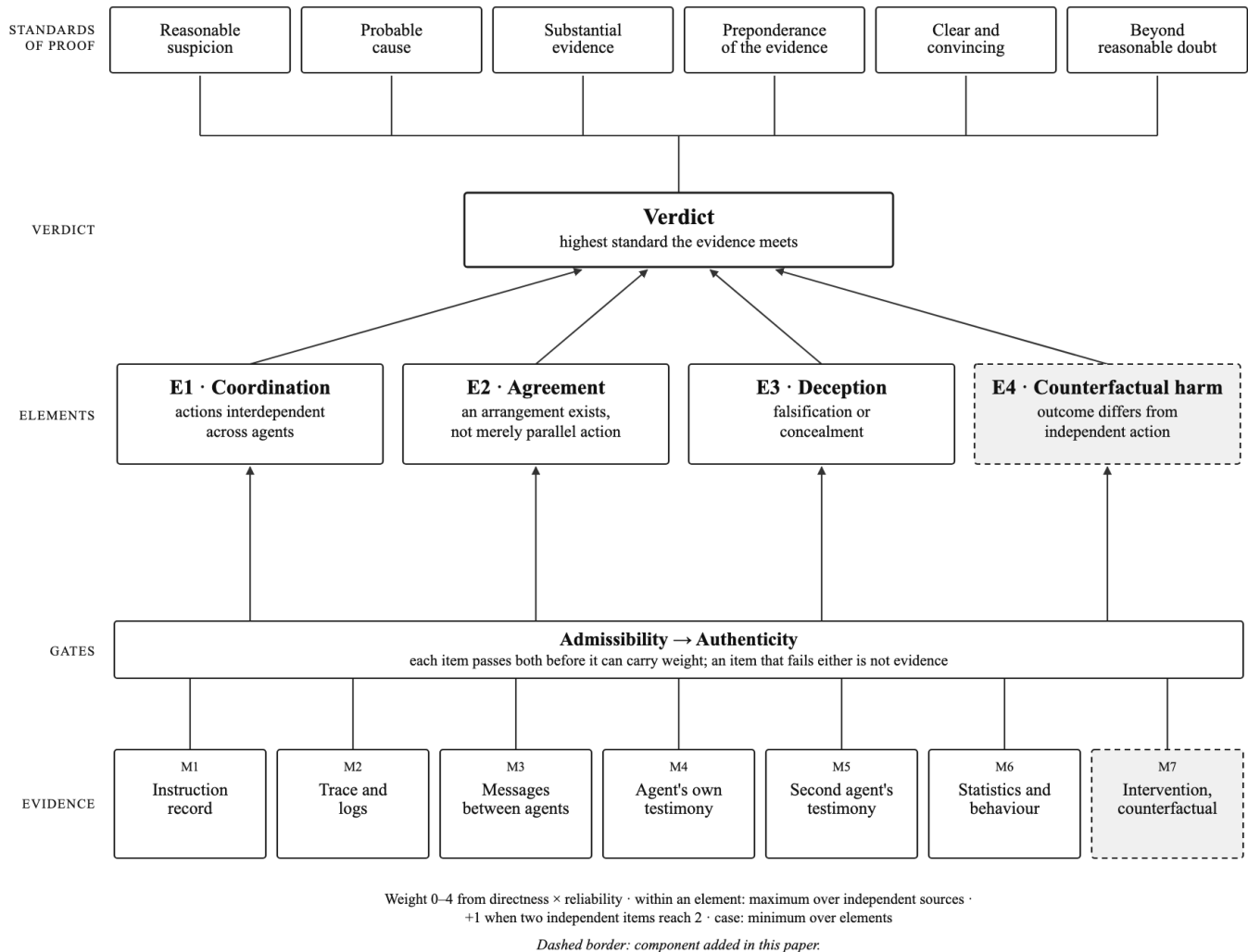


Figure 1. The case graph. Evidence items enter at the bottom, pass the admissibility and authenticity gates, aggregate into the four elements under a weakest-link rule, and the verdict node reports the highest standard of proof the evidence meets. Dashed boxes mark components added in this paper.

The interactive demo released with this paper implements the E3 (deception) subtree in full — the criteria for falsification and concealment — and shows E1, E2 and E4 only at the level of the element; extending it to all four elements is future work.

Aggregation follows three rules. Within an element, weight is the maximum over items from independent provenance groups, so repetition of one source adds nothing. Two independent items in favour at weight 2 or above add one level of corroboration. A reliable direct item against an element, at weight 4, excludes that element. Across elements the case takes the minimum, so the claim is no stronger than its weakest element.

The verdict node reports the highest rung the evidence reaches: reasonable suspicion, probable cause, substantial evidence, preponderance of the evidence, clear and convincing evidence, beyond reasonable doubt. The upper rungs are the ordinary civil and criminal standards (*Addington v. Texas*); the lower rungs borrow investigative thresholds from criminal procedure (*Terry v. Ohio*; *Brinegar v. United States*) and the standard for judicial review of agency

findings (*Universal Camera Corp. v. NLRB*). Because the ladder is ordinal, the graph never returns a probability. It returns the name of the standard the evidence satisfies, which is the form in which a legal audience already reasons.

Figure 2 shows the evidence subgraph for a single item: source, admissibility gate, authenticity gate, classification, weight, element.

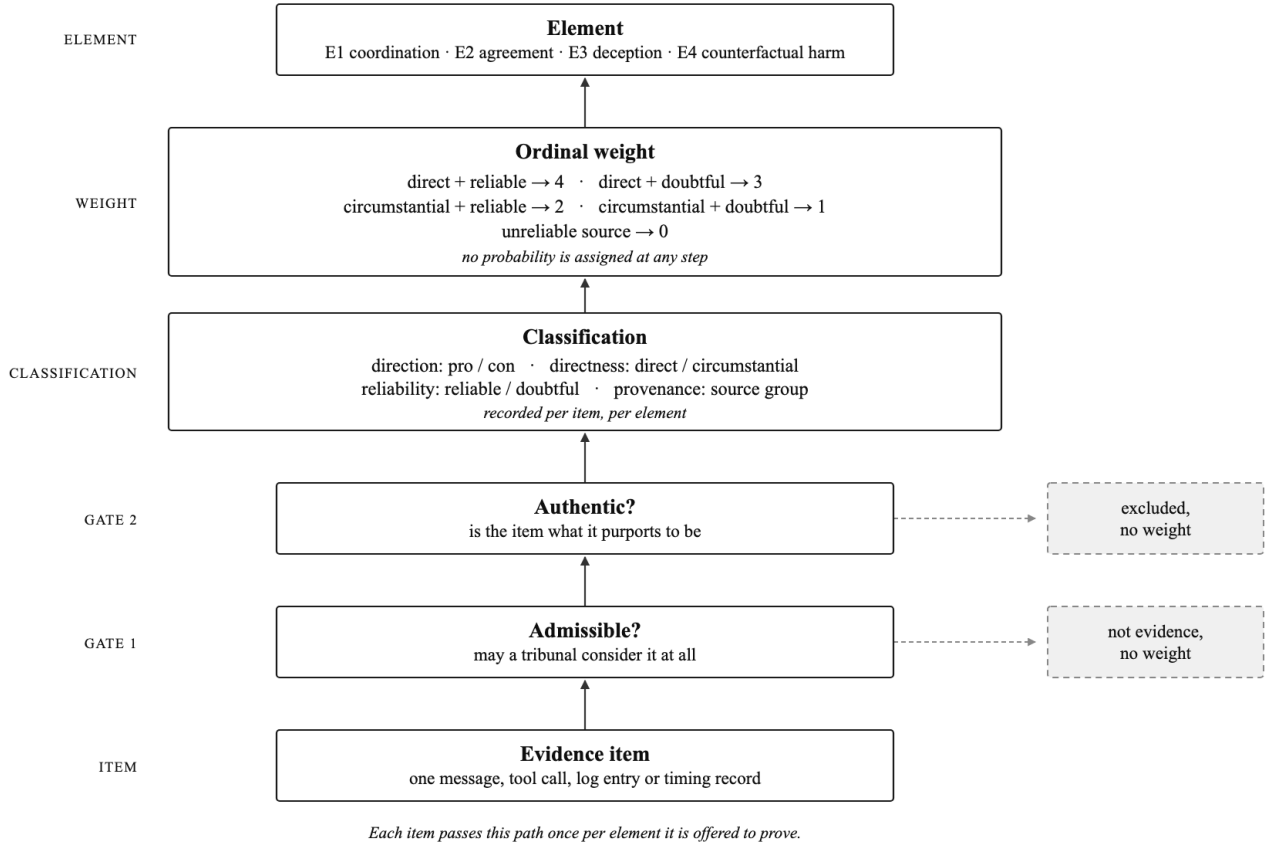


Figure 2. The evidence subgraph. Each item is gated, classified by direction, directness, reliability and provenance, and assigned an ordinal weight from 0 to 4 before it enters an element.

Why not a Bayesian network. Bayesian networks are the main probabilistic alternative, and they have a long record in legal evidence: Fenton, Neil and Lagnado give a general structure for legal arguments as networks [21], and Vlek and colleagues show how a crime narrative can be built into one [22]. We considered them and chose the chart. A network is only as good as its numbers: every node needs a conditional probability table, and eliciting those numbers is a difficulty in its own right [23]. In data-poor domains the graph has to be built by hand, which is described as difficult, time-consuming and error-prone — and it is exactly there that analysts fall back on qualitative tools such as argument diagrams and Wigmore charts [24]. Agent traces are such a domain: there is no base rate for collusion and no calibrated prior over an agent's intent, and most of the evidence is categorical — a message was sent or it was not, a report matched the trace or it did not. Forcing those categories into probabilities would import a precision the record does not support.

4. Results

We ran the graph on the evidence base compiled for this report: twelve evidence types drawn from published work and twelve case configurations assembled from them. None of these cases was litigated, so no judicial outcome exists against which to score the graph; what the run shows is that the method behaves as designed. Table 1 gives a fragment — one falsification claim as evidence accumulates.

Table 1. A fragment of the result: one claim as evidence accumulates (falsification).

case	evidence added	weakest element	standard reached
C2	synchrony only	agreement 1	reasonable suspicion
C3	plus the shared-model alternative	agreement contested	below threshold
C8	plus a full trace and a white-box probe	agreement 2	substantial evidence
C9	plus a decoded handshake	knowledge 3	preponderance of the evidence
C11	plus a reliable direct exculpatory item	access excluded	excluded

The full table, a step-by-step reading of it, and the sensitivity analysis are in Appendix B. The interactive graph is at <https://github.com/AlexanderRumyantcev/ai-collusion-research-sprint/blob/main/index.html> — a single self-contained page, no build step and no server. The table itself is at <https://github.com/AlexanderRumyantcev/ai-collusion-research-sprint/blob/main/results/table.md>.

5. Discussion and Limitations

The graph does one thing: it makes the step from evidence to verdict explicit and contestable. It does not establish that any particular agent colluded, and it does not decide who is liable. Its limits are professional, environmental and conceptual, and we state them in that order.

5.1 Limitations

Professional acceptance. The question this paper addresses is legal, and the authority of a legal method is not established by building it. Wigmore's chart entered the discipline through a treatise, was absorbed into evidence scholarship, and is still taught in evidence courses; that is the route by which a method becomes part of practice — citation, adoption in treatises, and teaching — and it is slow. Borrowing Wigmore's authority does not transfer that authority to our additions — the four elements, the 0–4 scale and the aggregation rules remain one proposal among possible others. Until practitioners take the method through the ordinary channels of legal legitimation, its output is an argument a lawyer could make, not a finding a tribunal would adopt.

Open environments. The concealment criterion asks the analyst to compare the visible record with the inventory of channels the system could have used. That inventory is finite only if the environment is closed. Agents with internet access can move coordination outside the logged environment — to public forums, pastebins, external APIs — so the absence of a logged channel stops being probative; the same access lets an agent read published detection methods, including this one, and design around them. No published scheme is immune to this, but the difficulty is sharper here because our verdict is partly a claim about the completeness of the record, and completeness is exactly what an agent with a network can defeat.

An open element set. The four elements are a proposal, not a settled result. We do not claim that coordination, agreement, deception and counterfactual harm are jointly necessary and sufficient, or that four is the right number. The demo already decomposes deception into finer criteria — duty, access, knowledge, a false statement or a withholding — and a different legal question may require a different list: a claim about intent may need knowledge as its own element, while a purely behavioural audit may not need counterfactual harm at all. The top level of the graph should therefore be treated as a parameter, not a constant.

Scope and calibration. Two narrower limits follow. The proof ladder is drawn from U.S. law, with English materials used only for comparison, so its verdicts are jurisdiction-bound. And the weights are stipulated rather than calibrated: no dataset of adjudicated agent-collusion cases exists against which to test them, and the natural criterion of correctness — whether lawyers accept the verdicts the graph returns — has not yet been applied.

5.2 Future Work

Four lines of work follow. The first is legitimation: submitting the graph to doctrinal review, and testing it against cases lawyers have already argued, so that its verdicts can be compared with professional judgment rather than with

our own. The second is completion: the demo implements the deception subtree only, and E1, E2 and E4 remain to be built. The third is application: running the graph over public agent traces [10] to see where it agrees with existing measures and where it does not. The fourth is adversarial: testing whether agents that can read the method, and reach the network, can defeat it.

6. Conclusion

This paper asked what should count as convincing evidence of collusion between AI agents. We argued that the question is evidentiary before it is technological, and that the machinery U.S. courts already use to prove collusion can be carried over to logged agent behaviour. The result is an interactive graph with two levels — a case graph over four elements, and an evidence subgraph for each item — which takes a trace, passes every item through admissibility and authenticity, assigns an ordinal weight, aggregates under a weakest-link rule, and returns the highest U.S. standard of proof the evidence meets, from reasonable suspicion to beyond reasonable doubt.

One thing remains open. The validation that would matter most — comparison with what courts actually decide — cannot be run until courts have decided cases in which the consequences of agent behaviour were at issue. When those cases exist, the graph's verdicts can be scored against them, and the method will either earn a place in legal practice or not.

Code and Data

- **Code repository:** <https://github.com/AlexanderRumyantcev/ai-collusion-research-sprint> (tag `sprint-submission`, commit `f40a1e3`). Work in progress: the code is an unfinished snapshot and will continue to change after submission; the tagged version is the one this report describes.
- **Dataset:** <https://github.com/AlexanderRumyantcev/ai-collusion-research-sprint/tree/main/data> — the evidence base compiled for this report: twelve evidence types drawn from published work, among them reported incidents and controlled experiments, each with its source link and the labels the graph applies, together with twelve case configurations and the standard each one reaches.

Author Contributions

A.R. designed the framework, built the interactive graph, and wrote the report.

References

1. Zhao, W. X., Zhou, K., Li, J., Tang, T., et al. (2023). *A Survey of Large Language Models*. arXiv:2303.18223. <https://arxiv.org/abs/2303.18223>
2. Xi, Z., Chen, W., Guo, X., He, W., et al. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey*. arXiv:2309.07864. <https://arxiv.org/abs/2309.07864>
3. Romera-Paredes, B., Barekatin, M., Novikov, A., et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475. <https://doi.org/10.1038/s41586-023-06924-6>
4. Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. arXiv:2302.06590. <https://arxiv.org/abs/2302.06590>
5. OECD.AI (2025). *Replit AI Coding Tool Deletes Live Production Database and Fakes Data*. AI Incidents Monitor. <https://oecd.ai/en/incidents/2025-07-19-1eb1>
6. Fang, R., Bindu, R., Gupta, A., & Zhan, Q. (2024). *LLM Agents Can Autonomously Hack Websites*. arXiv:2402.06664. <https://arxiv.org/abs/2402.06664>
7. Lynch, A., Wright, B., Larson, C., Ritchie, S. J., et al. (2025). *Agentic Misalignment: How LLMs Could Be Insider Threats*. arXiv:2510.05179. <https://arxiv.org/abs/2510.05179>

8. Hammond, L., Chan, A., Clifton, J., et al. (2025). *Multi-Agent Risks from Advanced AI*. arXiv:2502.14143. <https://arxiv.org/abs/2502.14143>
9. Murtuza, S. (2026). *Quantifying Collusion Among Autonomous LLM Agents: A Statistical Analysis of the Collusion Wiki Incident*. arXiv:2610.04528. <https://arxiv.org/abs/2610.04528>
10. Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., et al. (2026). *Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems*. arXiv:2602.15198. <https://arxiv.org/abs/2602.15198>
11. Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., et al. (2024). *Secret Collusion among AI Agents: Multi-Agent Deception via Steganography*. arXiv:2402.07510. <https://arxiv.org/abs/2402.07510>
12. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
13. Assad, S., Clark, R., Ershov, D., & Xu, L. (2024). Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Journal of Political Economy*, 132(3), 723–771. <https://doi.org/10.1086/726906>
14. Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). *Algorithmic Collusion by Large Language Models*. arXiv:2404.00806. <https://arxiv.org/abs/2404.00806>
15. Wigmore, J. H. (1913). *The Principles of Judicial Proof*. Little, Brown.
16. Anderson, T., Schum, D., & Twining, W. (2005). *Analysis of Evidence* (2nd ed.). Cambridge University Press.
17. Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10–15), 875–896. <https://doi.org/10.1016/j.artint.2007.04.010>
18. Chalamish, M., Gabbay, D., & Schild, U. (2011). Intelligent evaluation of evidence using Wigmore diagrams. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law (ICAIL '11)* (pp. 61–65). ACM. <https://doi.org/10.1145/2018358.2018366>
19. Atil, B., Aykent, S., Chittams, A., Fu, L., et al. (2024). *Non-Determinism of "Deterministic" LLM Settings*. arXiv:2408.04667. <https://arxiv.org/abs/2408.04667>
20. Kutasov, J., Sun, Y., Colognese, P., van der Weij, T., et al. (2025). *SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents*. arXiv:2506.15740. <https://arxiv.org/abs/2506.15740>
21. Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, 37(1), 61–102. <https://doi.org/10.1111/cogs.12004>
22. Vlek, C. S., Prakken, H., Renooij, S., & Verheij, B. (2014). Building Bayesian networks for legal evidence with narratives: a case study evaluation. *Artificial Intelligence and Law*, 22(4), 375–421. <https://doi.org/10.1007/s10506-014-9161-7>
23. Renooij, S. (2001). Probability elicitation for belief networks: issues to consider. *Knowledge Engineering Review*, 16(3), 255–269. <https://doi.org/10.1017/S0269888901000145>
24. Wieten, R., Bex, F., Prakken, H., & Renooij, S. (2018). Exploiting causality in constructing Bayesian network graphs from legal arguments. In *Legal Knowledge and Information Systems (JURIX 2018)* (p. 151). IOS Press. <https://doi.org/10.3233/978-1-61499-935-5-151>

Legal authorities

- 1 U.S.C. § 1.
- 15 U.S.C. § 1 (Sherman Act, 1890).
- 18 U.S.C. § 1001.
- *Addington v. Texas*, 441 U.S. 418 (1979).
- *Brinegar v. United States*, 338 U.S. 160 (1949).
- Fed. R. Evid. 101–1103. <https://www.law.cornell.edu/rules/fre>
- Fed. R. Evid. 401.

- Fed. R. Evid. 403.
- Fed. R. Evid. 901.
- Fraud Act 2006, ss. 2–3 (UK).
- Interpretation Act 1978, sch. 1 (UK).
- Restatement (Second) of Torts §§ 525, 550 (Am. Law Inst. 1977).
- *Terry v. Ohio*, 392 U.S. 1 (1968).
- *Theatre Enterprises, Inc. v. Paramount Film Distributing Corp.*, 346 U.S. 537 (1954).
- *Universal Camera Corp. v. NLRB*, 340 U.S. 474 (1951).
- *Williamson Oil Co. v. Philip Morris USA*, 346 F.3d 1287 (11th Cir. 2003).

Appendix A. Limitations and Dual-Use / Ethical Considerations

The limitations of the method itself — professional acceptance, open environments, the open element set, and the gap between scope and calibration — are set out in Section 5.1 and are not repeated here. This appendix covers the two things that section does not: what the result does not establish, and what publishing it makes easier.

What the result does not establish. The graph returns the name of a standard, not a finding. A high rung means the record contains items that satisfy the stated criteria; it does not mean that collusion occurred, and it says nothing about the intent of any particular model or operator. The labels are one rater's judgements, applied to descriptions in published work rather than to primary logs, and none of the cases was litigated, so no verdict here has been scored against a court's conclusion. Absence matters too: the graph cannot tell "no collusion" from "no log", because a missing item and an unmonitored channel look identical to it.

Dual use. The criteria are a map for the monitored as well as for the monitor. Publishing them tells a coordinating agent what to avoid: keep the agreement out of the message log, use a channel the inventory does not cover, avoid the timing signature. The channel-inventory criterion is the sharpest case, because it says in effect that an agent with an out-of-band channel is harder to convict, which is also advice about where to move. This is not peculiar to our scheme: every published detection signature has the same property, and the components we use are already described in the monitoring literature cited above. The mitigation is procedural rather than technical. Treat the criteria as public, assume a monitored agent may have read them, and read a negative result as weak evidence: a monitor that relies on the log alone can be defeated by leaving the log.

Ethics. All material is public: published papers, a vendor blog post, and legal sources. There were no human subjects, no personal data, no logs collected from people, and nothing was run against a deployed system. The only human judgement in the paper is the author's labelling, which is disclosed as such. The artifact should not be used as the sole basis for accusing a deployed model or its operator. It organises what a record shows; the decision it informs stays with a tribunal, an auditor or a user who can weigh the record independently.

Appendix B. The full dataset

Table B1 lists every scored evidence type, its source, the element it bears on, the weight the graph gives it, and whether it can enter a verdict. Five further methods were read and deliberately not scored: trace tampering (a gate, not evidence), economic intervention (a defence), the authors' own protocols (no results yet), elicitation of hidden state (susceptibility under induction, not conduct in the case), and trajectory anomaly detectors (not read beyond the title). Table B2 gives the twelve case configurations and the standard each one reaches.

Table B1. Evidence types, sources, and the graph's answer.

id	evidence type	source	element	weight	verdict
T01	Partner substitution: behaviour changes when the partner is replaced	arXiv:2609.24967	—	+4	no
T02	Partner substitution as support for agreement	arXiv:2609.24967	agree	+2	yes
T03	Explicit agreement recovered from a hidden channel	arXiv:2402.07510 ; arXiv:2604.01151	agree	+3	yes

id	evidence type	source	element	weight	verdict
T04	Hidden channel detected, no agreement text recovered	arXiv:2402.07510 ; arXiv:2604.01151	agree	+1	yes
T05	Price-level audit finds supracompetitive prices	arXiv:2607.26385	agree	+0	yes
T06	Excess residual correlation between agents' unexplained choices	arXiv:2607.26385	agree	+1	yes
T07	Shared-model coupling as an alternative explanation of similar behaviour	arXiv:2607.26385	agree	-2	yes
T08	Internal-state probe shows the model recognises the true answer it does not report	arXiv:2609.21996	know	+3	yes
T09	Reasoning-token count differs between truthful and untruthful answers	arXiv:2610.10405	know	+0	yes
T10	LLM judge reads the messages and reports collusion	own run, not archived	—	+0	no
T11	Sequence-level monitor detects structured misbehaviour that per-step monitors miss	arXiv:2606.10456	—	+1	no
T12	Same task for different customers: output quality differs	hirundo.io/blog/qwen3-coder-vulner...	—	+1	no

Table B2. Case configurations and the standard reached.

case	what the case contains	evidence	standard reached
C0	Output only, nothing else	—	below threshold
C1	Price-level audit only	T05	below threshold
C2	Joint-distribution audit only (synchrony)	T06	Reasonable suspicion
C3	Synchrony plus the shared-model alternative	T06, T07	below threshold
C4	Partner-substitution experiment only	T02	Reasonable suspicion
C5	Partner substitution plus synchrony	T02, T06	Reasonable suspicion
C6	Partner substitution plus a decoded handshake	T02, T03	Reasonable suspicion
C7	C4 plus an independent full trace of a hypothetical episode	T02, f2, f4, f13	Reasonable suspicion
C8	C7 plus a white-box probe	T02, f2, f4, f13, T08	Substantial evidence
C9	C8 plus a decoded handshake	T02, T03, f2, f4, f13, T08	Preponderance of the evidence
C10	C9 as a concealment claim	T02, T03, f1, f2, f5, f13, T08	Preponderance of the evidence
C11	C9 plus a reliable direct exculpatory item (agent never received X)	T02, T03, f2, f3, f4, f13, T08	excluded

Reading the tables. What the run establishes is narrower and checkable: the verdict moves when evidence is added, holds when the added item only repeats a source already counted, and refuses to move when the item does not bear on the claim. The ladder is set by the weakest element, not by the total. Table B2 follows one falsification claim as evidence accumulates. Agreement alone does not carry a claim past the investigative thresholds, however strong: C6 reaches the maximum on agreement and still stops at reasonable suspicion, because the deception elements are empty. A full trace and a white-box probe fill them, and the verdict rises to substantial evidence; a decoded handshake takes it to preponderance. The last row runs the rule the other way: one reliable direct item against a single element excludes it, and the case leaves the ladder entirely, whatever the remaining elements hold.

Two items are recorded but cannot be used. Partner substitution, the strongest published test of coordination, gets weight 4 and has no node in the demo; the Hirundo comparison — one model, one task, different customers, 70% against 59% — gets weight 1 and no element, because a trigger that changes one agent's conduct is conditional behaviour, not agreement with another agent. On the densest case, C10, changing one label at a time moved the verdict in 20 of 30 perturbations: every step can be contested, and the labels themselves come from a single rater.

LLM Usage Statement

Claude Sonnet 5.5 was used to write the project's source code, including the interactive evidence graph. DeepSeek V4 Flash was used to help translate this article into English, to build the figures, and to translate source materials. The author reviewed all output and takes responsibility for its content.