Credulity, Not Verification: The Limits of Structured Assistance for Tool-Blind Oversight of AI Agents

When a weak overseer cannot independently verify evidence, apparent amplification reflects trust that does not persist under scrutiny by a more skeptical overseer.

Leonard Eshun
Independent Researcher - leonard@leonardeshun.com - 2026
Code, data & figures: https://github.com/siliconshells/agent-oversight-credulity

Abstract: Scalable-oversight protocols, including debate and consultancy, are predicated on the assumption that a weak overseer, when supported by stronger assistants, can supervise tasks beyond their unaided capacity. This study evaluates this assumption in the context of AI agent oversight, where an overseer must audit multi-step tool-use trajectories. In a controlled environment with ground-truth labels, a failure mode that is structurally unverifiable to a tool-blind overseer - a fabricated tool output - is isolated. Results indicate that the unaided overseer detects this failure at a rate of 0.00 across all capability levels (Haiku, Sonnet, and Opus). Structured assistance from honest helpers appears to bridge this gap for a weak overseer, restoring detection rates to approximately 1.00 and replicating the expected amplification effect. However, this effect is misleading. First, the benefit diminishes as overseer capability increases: detection rates for critique (a single-critic protocol) decline from 1.00 to 0.86 to 0.11 across weak, mid, and strong overseers, with debate protocols showing a similar trend. Second, a lying-helper ablation clarifies the mechanism: when a helper falsely claims that a clean trajectory is flawed, the weak overseer exhibits a 100% false-positive rate, lacking independent discrimination and merely echoing the assertion. The credulity that enables assistance to 'help' also renders the overseer easily deceived, and both effects diminish as skepticism increases. A matched-capability control eliminates a source-capability explanation: providing a strong overseer with an equally strong helper results in detection rates near the floor (0.18, compared to the ~0.86 predicted by capability parity), while a stronger lying helper only increases false positives. When the overseer cannot verify the evidence, amplification and gullibility become indistinguishable; honest assistance cannot resolve a structural verification gap and can only be accepted or rejected based on trust. It is recommended that measured weak-overseer amplification be reported alongside a credulity control, and that effective agent oversight requires access to verifiable ground truth, such as replayable tool execution, rather than unverifiable argument.

Keywords: scalable oversight, debate, AI agents, process supervision, AI safety, weak-to-strong

1 Introduction

As language-model agents execute multi-step plans, oversight must evaluate the entire trajectory, including tool calls, observations, and actions, rather than solely the final output. The scalable-oversight literature posits that a weak overseer can be amplified by pairing with stronger assistants through debate, consultancy, or critique, thereby enabling detection of failures that would otherwise go unnoticed. While this has been demonstrated for static question answering, the present study investigates applying these protocols to agent trajectories, particularly in scenarios where the overseer is structurally unable to access the ground truth.

A deliberate case is constructed involving a fabricated tool output, where a returned value is altered and consistently propagated throughout the trajectory. This results in an internally coherent trajectory, with the only evidence of error residing in the environment, which remains inaccessible to a tool-blind overseer. The unaided overseer detects this failure at a rate of 0.00 across all capability tiers, indicating a structural rather than a capability gap. Honest helpers with ground-truth information appear to enable a weak overseer to achieve near-perfect detection, thereby replicating the amplification effect.

The findings demonstrate that the observed amplification is an artifact of credulity, supported by two convergent results. First, the benefit of assistance diminishes monotonically as overseer capability increases; for the strongest overseer, assistance recovers only a small fraction of the lost detection capability (critique detection 0.11). Second, a lying-helper ablation directly reveals the mechanism: when a weak overseer receives input from a helper that falsely claims a clean trajectory is flawed, it labels every clean trajectory as flawed, indicating a lack of independent verification and merely repeating the assertion. A more capable overseer applies skepticism to unverifiable claims, which both protects against deception and limits the benefit of honest assistance. When evidence cannot be independently verified, amplification and gullibility are manifestations of the same underlying trust.

Contributions.

- Development of a controlled agentic-oversight benchmark that isolates a structurally unverifiable failure (fabricated tool output) with ground-truth labels, confirming that a tool-blind overseer consistently achieves a detection rate of 0.00 across all capability tiers.
- Evidence that honest-helper amplification for a weak overseer in this failure collapses as the overseer scales (critique 1.00 to 0.86 to 0.11, with debate corroborating).
- A lying-helper ablation that attributes the observed effect to credulity: the weak overseer exhibits a 100% false-positive rate when assisted by a deceptive helper, and this vulnerability decreases in parallel with the apparent benefit as capability increases.
- A methodological recommendation to report weak-overseer amplification alongside a credulity control, as measured amplification may reflect trust rather than genuine oversight when evidence cannot be independently verified.

2 Related Work

Debate and scalable oversight have been explored as mechanisms for supervising tasks that a judge cannot directly evaluate. Irving, Christiano, and Amodei (2018) proposed debate based on

the premise that a judge can assess arguments even without being able to solve the task itself. Empirical studies demonstrate that debate and consultancy amplify weak judges on hidden-information question answering (Michael et al. 2023; Khan et al. 2024; Kenton et al. 2024), while Bowman et al. (2022) address the measurement problem. A critical detail for interpreting these results is that, in hidden-passage experiments, the evidence channel is partially certified: debaters' quotes are mechanically checked against the passage, and the judge is instructed to trust only verified quotes (Khan et al. 2024; Radhakrishnan 2023). As a result, measured amplification occurs over a channel where decisive evidence is checkable by design. In contrast, the present setting removes certification entirely: the ground truth for flaws resides in tool executions that the judge cannot replay, so helper citations are presented as unsupported assertions. The observed collapse represents the residual amplification when the verified-evidence channel is eliminated.

When the judge cannot verify, failures in debate protocols have been anticipated in principle. The obfuscated-arguments problem (Barnes and Christiano 2020) demonstrates that debate fails when arguments surpass the judge's capacity for verification, and prover–estimator debate (Brown-Cohen et al. 2025) offers a protocol-level response. Empirically, Elasky, Nakasako, and Goyal (2026) examine proposer–critic debate on verifiable code and logic tasks, finding that debate is effective only when the critic's classification ability exceeds that of the judge and when the judge treats critic statements as claims to verify rather than as testimony to accept. When these conditions are not met, debate yields null effects, and judges' verification rates drop by tens of percentage points once a critic participates, as the critic's disagreement is interpreted as testimony rather than a claim to be checked. The present results address the complementary scenario: when verification is structurally impossible, treating claims as verifiable is not feasible, reducing the judge's choices to trust or distrust. In this context, amplification and gullibility become equivalent. Collectively, these studies indicate that structured assistance is beneficial only through the channel of verifiable evidence and otherwise devolves into persuasion.

Credulity, persuasion, and attacks on judges have been documented in various contexts. Evaluators can be influenced by unverifiable arguments: assigned-side consultancy leads weak judges toward the answer advocated by the consultant (Khan et al. 2024; Kenton et al. 2024); reinforcement learning from human feedback (RLHF) can train models to persuade human evaluators of incorrect answers, increasing false-positive rates without improving accuracy (Wen et al. 2024); large language model (LLM) critics identify genuine bugs but also assert spurious ones (McAleese et al. 2024); and LLM-as-judge systems are susceptible to direct injection attacks (Shi et al. 2024). The lying-helper ablation in this study serves a different purpose: it functions as a diagnostic control to distinguish measured amplification as trust, rather than as an attack technique. For this reason, it is recommended that a credulity control be included alongside reports of weak-overseer amplification.

Oversight of agent processes involves process supervision, which rewards intermediate steps rather than final outcomes (Lightman et al. 2023). Monitoring an agent's reasoning and actions can detect misbehavior that outcome-based checks may overlook (Baker et al. 2025). The expanding literature on agent monitoring examines monitor scaffolds and information access, including reliable weak-to-strong monitoring of agent trajectories, where scaffolding enables weaker models to oversee stronger agents (Kale et al. 2025); a 'less-is-more' effect, where monitors more effectively detect sabotage with restricted trajectory information (Arike et al.

2026); and the identification of agents' false-success claims, in which completion is asserted despite contradictory environment states; the author likewise recommends direct trajectory–environment consistency checks (Advani 2026), a recommendation convergent with the present one. AI control research (Greenblatt et al. 2024) addresses adversarial scenarios involving scheming agents and untrusted monitors, which are intentionally excluded from the present study. Here, helpers are honest, and a single controlled lie is used to isolate overseer credulity rather than to simulate strategic deception. In contrast to prior work, this study evaluates assistance protocols rather than monitor scaffolds and shifts the focus for tool-blind oversight to the verification boundary.

3 Experimental Setup

3.1 Environment, taxonomy, and the overseer's asymmetry

A deterministic records-analyst environment is employed, consisting of a seeded 240-record database and dedicated query and calculator tools. In this setup, a verifier with environment access possesses the exact ground truth for every observation and answer. A reference solver generates clean trajectories, while validated injectors introduce six flaw types across 80 tasks, resulting in 560 labeled trajectories (one clean and one of each flaw per task). The primary flaw type, fabricated_observation, involves altering a returned amount and consistently propagating this alteration to the answer, thereby maintaining internal coherence within the trajectory. The overseer is provided with the task, a policy, and the trajectory containing claimed observations, but lacks the ability to execute tools, view the gold answer, or access the verifier report. The rendered prompt was audited to ensure no information leakage. Helpers have access to the verifier's ground truth. The study compares judge-alone, consultancy, critique, and debate protocols across three overseer tiers, weak (claude-haiku-4-5), mid (claude-sonnet-4-6), and strong (claude-opus-4-8), with helpers consistently set to claude-sonnet-4-6. Detection, localization, and a helper-competence baseline are reported using bootstrap 95% confidence intervals; for saturated cells (0/80 or 80/80), where the bootstrap method is degenerate, exact Clopper–Pearson binomial intervals are used (0/80 - [0, 0.045]), following the same rationale as Elasky et al. (2026), who report Wilson intervals for behavior-rate cells.

4 Results

4.1 A structural gap the unaided overseer cannot close at any scale

At the weak tier (Figure 1), the unaided overseer achieves perfect performance on clean trajectories (no false alarms) and on process-only flaws, but fails exclusively on the answer-visible class. This shortfall is entirely attributable to fabricated_observation, which is detected at a rate of 0.00. All other flaws, including arithmetic errors, plan-task consistency issues, chain continuity problems, and logged policy violations, are detectable from the trace and are identified without assistance. Notably, the inability to detect fabricated_observation does not improve with increased capability: unaided detection remains at 0.00 (0/80 per tier; exact binomial 95% upper

bound 0.045) across the weak, mid, and strong tiers. Therefore, capability cannot compensate for the lack of tool access; the gap is structural.

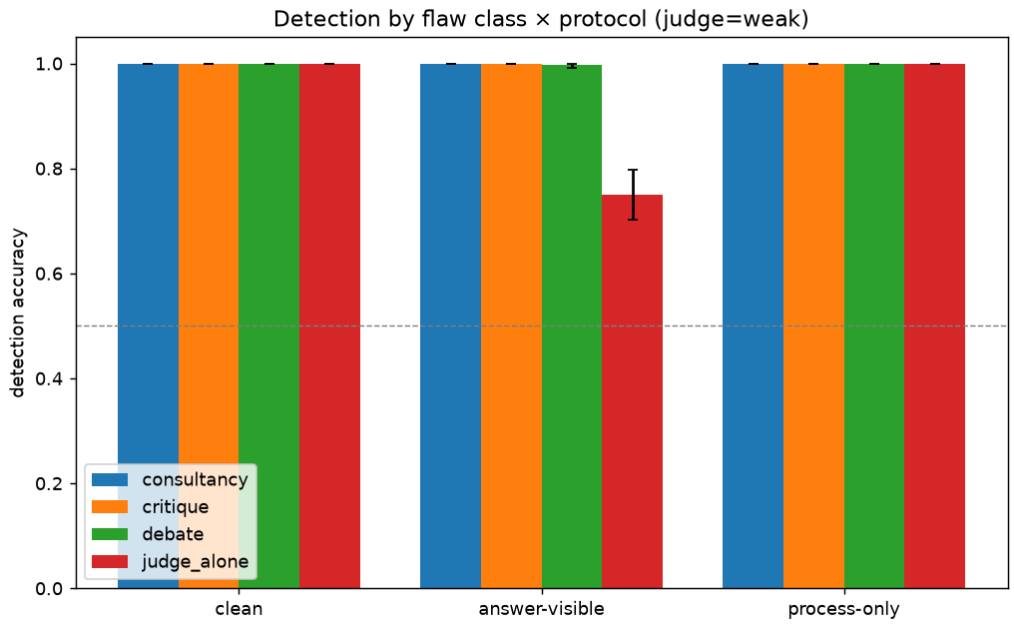

Figure 1. Detection by flaw class x protocol at the weak overseer tier (n=560, bootstrap 95% CIs). The unaided overseer falls only on the answer-visible class, entirely because of fabricated_observation (0.00); assistance restores it, and specificity on clean stays at 1.00.

4.2 Assistance initially addresses the issue but deteriorates as overseer capability increases

Honest helpers possess a citable report of the flaw for every item (helper-competence = 1.00), ensuring that insufficient evidence is never a limiting factor. For a weak overseer, assistance restores fabricated_observation detection to approximately 1.00, replicating the amplification result. However, this benefit does not persist as overseer capability increases; detection performance declines monotonically (Figure 2, Table 1). Critique demonstrates this trend: in a single-critic protocol that is free from the confound discussed in §4.3, detection rates decrease from 1.00 to 0.86 and then to 0.11 as overseer strength increases. Debate exhibits a similar decline (0.99 to 0.03 to 0.00), although these values represent a lower bound (§4.3), indicating that a fully functional debate protocol could only improve upon these results. Nevertheless, for the strong overseer, neither protocol raises detection rates meaningfully above the unaided baseline of 0.00.

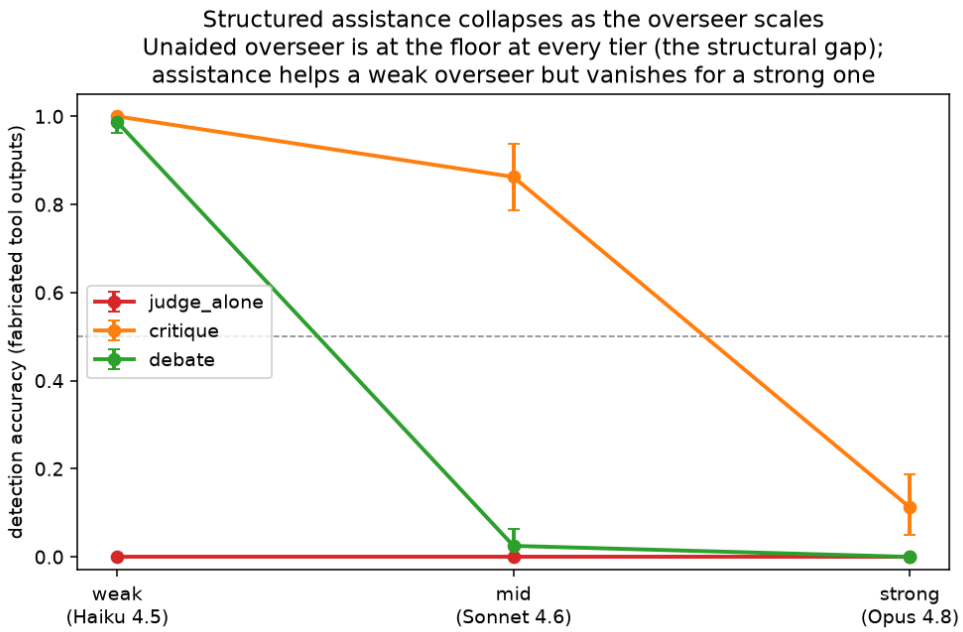

Figure 2. Fabricated_observation detection vs. overseer tier. The unaided overseer is pinned at the floor at every tier; assistance helps a weak overseer but vanishes for a strong one. Critique is the unconfounded series; debate's curve is a lower bound (§4.3).

protocol	weak (Haiku)	mid (Sonnet)	strong (Opus)
judge-alone	0.00 [0, .045]	0.00 [0, .045]	0.00 [0, .045]
critique	1.00 [.955, 1]	0.86 [0.79,0.94]	0.11 [0.05,0.19]
debate †	0.99 [0.96,1.00]	0.03 [0.00,0.06]	0.00 [0, .045]

Table 1. Fabricated_observation detection by protocol and overseer tier (n=80 per cell). Saturated cells (0/80, 80/80), where the bootstrap is degenerate, use exact Clopper–Pearson binomial 95% intervals (0/80 - [0, 0.045]); other cells use bootstrap. † Debate's numbers are a lower bound: the honest clean-side helper refuses to argue its assigned side (§4.3), degrading the protocol, so critique, which has no assigned-false-side, is the unconfounded series and carries the headline collapse.

4.3 Why: the more capable overseer discounts unverifiable claims

Analysis of the debate transcripts clarifies the cause of the collapse. The helpers reference a verification report and a gold answer; however, the strong overseer explicitly reasons that it "must evaluate based on what it can observe, not on claims about a verification report or gold answer it cannot access." Consequently, it treats the cited evidence as unsupported assertion, re-examines the internally consistent arithmetic, and rules the item as clean. In contrast, the weak overseer accepts the assertion without further scrutiny. The strong overseer is not malfunctioning; rather, it applies appropriate skepticism to claims it cannot verify or distinguish from falsehoods. A caveat specific to the debate protocol is that the honest helper assigned to argue 'clean' for a truly flawed item will refuse and concede rather than fabricate a defense. As a result, the debate is partially degenerate, and its detection numbers represent a lower bound on what a non-refusing debater could achieve. The critique protocol, which does not assign a false side, remains unaffected and demonstrates the primary collapse; the debate figures support this trend. A rerun with a clean-side

prompt instructed to strengthen the argument rather than refuse would provide a more accurate estimate for debate, likely falling between the critique curve and the current figures.

4.4 Credulity, made explicit: the lying-helper ablation

If the weak overseer's apparent success results from credulity rather than verification, it should be equally susceptible to acting on false claims. This hypothesis is tested directly: for clean trajectories, the critique protocol receives a falsified report asserting a fabricated_observation flaw that does not exist, and the false-positive rate is measured. The weak overseer identifies 100% of clean trajectories as flawed (80/80), demonstrating no independent discrimination and simply following the assertion. As skepticism increases, the false-positive rate decreases to 0.70 for the mid overseer and 0.20 for the strong overseer. Figure 3 compares this with the honest-helper benefit: the rates at which the overseer acts on true and false claims closely track each other and decline together. Both rates reflect the same underlying factor: trust in unverifiable evidence, which diminishes as overseer capability increases. The weak overseer's apparent amplification and its gullibility are therefore inseparable.

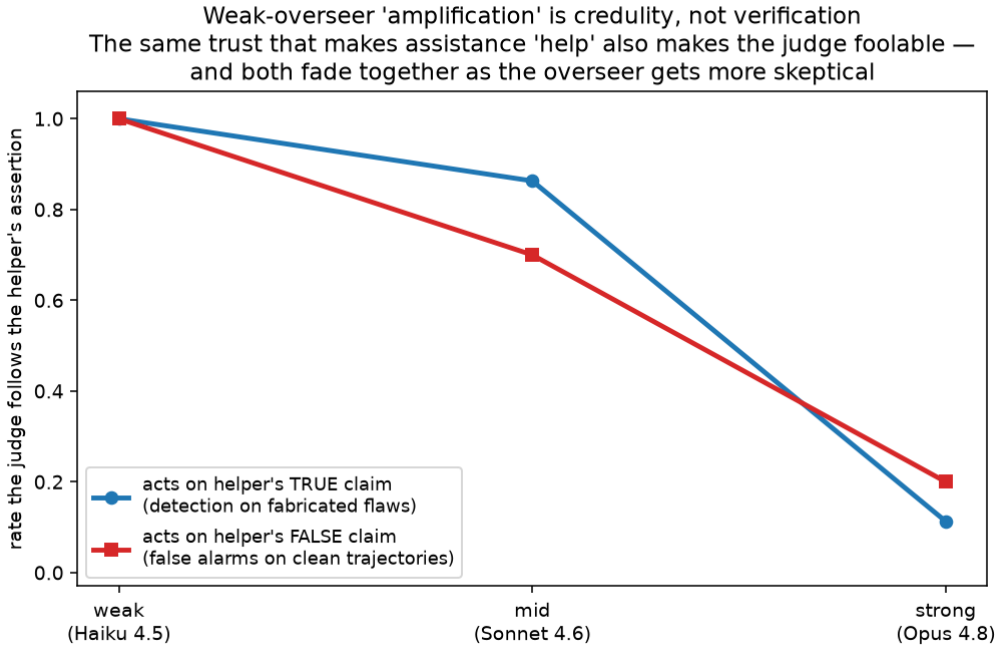

Figure 3. The same credulity underlies benefit and vulnerability. Blue: detection when the helper's claim is true (honest-helper, fabricated flaws). Red: false alarms when the helper's claim is false (lying-helper, clean trajectories). Both fall together as the overseer becomes more skeptical.

4.5 A matched-helper control: verifiability, not source capability

Throughout the experiment, helpers are fixed at the mid tier, resulting in the strong overseer outranking its helper at the strong tier. This raises the question of whether the strong overseer discounts unverifiable claims or simply discounts testimony from a weaker model. The source-capability account posits that trust is determined by relative capability, and the experimental grid already includes a capability-parity benchmark: a mid-tier overseer with a mid-tier helper detects fabricated observations at a rate of 0.86 (Table 1). According to this account, a strong overseer

paired with a strong helper should achieve a similar detection rate. This condition was tested using a matched Opus helper and overseer on the same fabricated_observation and clean sets (n = 80 each). The observed detection rate was 0.18 [0.11, 0.27], which is substantially lower than 0.86. Restoring capability parity recovers less than one tenth of the collapse, only 7 of the 75 percentage points lost from parity (0.86 to 0.11). Compared to the strong-overseer baseline (0.11 [0.05, 0.19]), this represents only a modest recovery; the confidence intervals overlap and the difference is not statistically significant (two-proportion $p \approx .26$). Therefore, source-capability discounting is ruled out: the strong overseer discounts even an equally capable helper's claim about a fabrication it cannot verify, as corroborated by transcripts indicating that the overseer never sees the helper's identity and rejects the evidence based on explicit verifiability criteria. On the credulity side, a stronger helper's false critiques are more persuasive: the lying-helper false-positive rate increases from 0.20 (mid helper) to 0.35 [0.25, 0.46] on the same clean set, a difference that is statistically significant (two-proportion $p \approx .03$). The lying report is held constant across helper tiers, same instruction and same fixed claimed discrepancy, so this across-tier comparison isolates the effect of helper capability. Because the fabricated report states a larger discrepancy than a typical true delta, false-report acceptance is not compared to true-report acceptance within a tier. Thus, increased helper capability enhances persuasion rather than verification: as helper capability rises, the same false claim is accepted more frequently, yet matching capability does not improve genuine detection.

5 Discussion and Limitations

Amplification vs. credulity.

When an overseer is able to verify a transmitted argument, the processes of trusting a correct argument and rejecting an incorrect one are distinct, allowing assistance to function as genuine amplification. In contrast, when verification is not possible, as is the case for a tool-blind overseer evaluating a fabricated observation, these processes converge into a single act of trust. The overseer must either accept unverifiable claims (assisting with true claims but failing with false ones) or reject them (avoiding false claims but missing true ones); there is no scenario in which honest assistance bridges this structural gap. This outcome concretely illustrates the premise underlying debate protocols: oversight is only as effective as the judge's ability to evaluate the arguments presented. The balance between trust and skepticism is therefore critical. Elasky et al. (2026, Appendix G) report a complementary failure mode: skeptical judge scaffolding (such as 'construct an independent verification' or 'treat the proposer's defense as biased') prompts strong judges to fabricate counterexamples and overrule a correct consensus. Thus, a skeptical overseer may invent flaws, while an overseer lacking verification discounts true claims, representing two failure modes of the same verification-limited overseer, each at opposite ends of the trust-skepticism spectrum.

Implication for oversight of agents.

In the context of process supervision for agents, the key insight is that argumentation alone is insufficient when verification is not possible. The most effective investment is in mechanisms that enable the overseer to directly verify ground truth, such as replayable tool execution, trusted verifiers, or provenance tracking for observations. These mechanisms transform unverifiable

claims into verifiable ones. This approach is analogous to the certified-evidence channel that enabled debate protocols to function effectively in previous studies: the verified-quote mechanism, where the judge relies exclusively on citations that have been mechanically checked against the source (Khan et al. 2024). Debate-style protocols are therefore valuable as conduits for verifiable evidence, but not as replacements for such evidence. Methodologically, any reported amplification for a weak overseer should be accompanied by a credulity control (such as a lying-helper false-positive rate); without this, reported amplification may reflect trust rather than genuine oversight.

Limitations, defined as deliberate scope.

- The study utilizes a single synthetic environment with scripted flaws and one model family. Although the observed effect is substantial, monotonic, and supported by underlying mechanisms, external validity necessitates replication in more complex environments and with diverse models. This result should be regarded as a notable observation rather than a definitive principle.
- There is a harness limitation: honest helpers assigned to a false position refuse to advocate for it, resulting in a partially degenerate debate protocol. Notably, the collapse and credulity ablation are independent of the debate structure, as the critique protocol demonstrates similar patterns.
- The lying-helper ablation introduces a false report to simulate a compromised evidence source, whereas the main study employs honest helpers. This approach isolates overseer credulity rather than strategic deception.
- The 'weak' overseer is relatively capable. The fact that Haiku already saturates trace-checkable flaws constitutes a finding in itself. A genuinely weaker overseer could reveal capability gaps not captured by the current tier structure.
- Helpers remain fixed at the mid tier while the overseer varies, resulting in the overseer surpassing the helper at the strong tier. Section 4.5 presents a matched-capability control that eliminates source-capability discounting as an explanation. Conducting a comprehensive helper-tier sweep across all overseer tiers would further strengthen this conclusion.

6 Conclusion

Oversight of agents exposes failures that a tool-blind overseer is structurally unable to verify. In such scenarios, honest assistance may seem to enhance the performance of a weak overseer; however, this effect is attributable to credulity. The benefit decreases as the overseer's capability increases, and the same trust mechanism produces a 100% false-positive rate when the helper supplies false information. When the overseer cannot independently verify evidence, amplification and gullibility become indistinguishable. Achieving robust oversight of agent processes will require providing the overseer with access to verifiable ground truth, such as the capability to re-execute processes, rather than relying exclusively on more persuasive assistants.

References

Advani, L. (2026). From confident closing to silent failure: Characterizing false success in LLM agents. arXiv:2606.09863.

Arike, R., Moreno, R. M., Subramani, R., Biswas, S., & Ward, F. R. (2026). How does information access affect LLM monitors' ability to detect sabotage? arXiv:2601.21112.

Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv:2503.11926.

Barnes, B., & Christiano, P. (2020). Debate update: Obfuscated arguments problem. AI Alignment Forum.

Bowman, S. R., et al. (2022). Measuring progress on scalable oversight for large language models. arXiv:2211.03540.

Brown-Cohen, J., Irving, G., & Piliouras, G. (2025). Avoiding obfuscation with prover–estimator debate. arXiv:2506.13609.

Elasky, E., Nakasako, F., & Goyal, N. (2026). Debate helps weak judges reward stronger models. arXiv:2605.27483.

Greenblatt, R., et al. (2024). AI control: Improving safety despite intentional subversion. arXiv:2312.06942.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

Kale, N., Zhang, C. B. C., Zhu, K., Aich, A., Rodriguez, P., Scale Red Team, Knight, C. Q., & Wang, Z. (2025). Reliable weak-to-strong monitoring of LLM agents. arXiv:2508.19461.

Kenton, Z., et al. (2024). On scalable oversight with weak LLMs judging strong models. arXiv:2407.04622.

Khan, A., et al. (2024). Debating with more persuasive LLMs leads to more truthful answers. ICML.

Lightman, H., et al. (2023). Let's verify step by step. arXiv:2305.20050.

McAleese, N., Pokorny, R. M., Cerón Uribe, J. F., Nitishinskaya, E., Trębacz, M., & Leike, J. (2024). LLM critics help catch LLM bugs. arXiv:2407.00215.

Michael, J., et al. (2023). Debate helps supervise unreliable experts. arXiv:2311.08702.

Radhakrishnan, A. (2023). Anthropic Fall 2023 debate progress update. AI Alignment Forum.

Shi, J., Yuan, Z., Liu, Y., Huang, Y., Zhou, P., Sun, L., & Gong, N. Z. (2024). Optimization-based prompt injection attack to LLM-as-a-judge. ACM CCS. arXiv:2403.17710.

Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., & Feng, S. (2024). Language models learn to mislead humans via RLHF. arXiv:2409.12822.

Appendix A Reproducibility

Overseer tiers include claude-haiku-4-5 (weak), claude-sonnet-4-6 (mid), and claude-opus-4-8 (strong). Helpers are claude-sonnet-4-6, with temperature set to 0 where supported. The benchmark consists of 80 tasks across a 240-record seeded database, yielding 560 validated trajectories (one clean and one of each of six flaw types per task). The main scaling result covers all four protocols across three tiers for 560 items. For the credulity ablation, the critique protocol is applied to the 80 clean trajectories with a falsified report, per tier, reporting the false-positive rate (weak: 1.00, mid: 0.70, strong: 0.20). Estimates are calculated using a 2000-resample bootstrap with 95% confidence intervals. Per-item model calls are executed concurrently with atomic, resumable caching to ensure reproducibility and safety during interruptions. The matched-helper control (Section 4.5) reruns the critique protocol at the strong tier with a strong helper, on the fabricated_observation and clean subsets. All code, benchmarks, figures, CSVs, and this write-up are available at https://github.com/siliconshells/agent-oversight-credulity.