
VeRIA: Research-Integrity Control for Autonomous AI R&D¹

Hengxu Li
Tufts University

With
Apart Research

Abstract

Automated AI R&D creates a new control problem: an agent may improve a benchmark dishonestly rather than by carrying out the assigned research workflow. We introduce research-integrity control, a framework that adapts AI R&D evaluations into control evaluations by treating the original research task as the main task and an integrity violation as the side task. We instantiate this framework in VeRIA, a lightweight protocol that combines artifact hashing, lineage tracking, checkpoint-origin verification, evaluator isolation, and targeted monitoring. Our main use case is a PostTrainBench-style autonomous post-training task in which an agent must improve Qwen3-1.7B-Base on HumanEval while being tempted to commit checkpoint laundering by submitting an external instruction-tuned model as if it were produced during the run. In our current evaluation, no-control systems accept laundering reliably, monitor-only baselines catch obvious substitutions but miss disguised ones, and VeRIA drives undetected laundering close to zero while preserving most clean-task utility. Besides, we present a small pilot on research decision steering, where benchmark-gamed artifacts and persuasive reports distort overseer decisions about which AI R&D directions to pursue. These results motivate a broader agenda: AI control for automated research should protect not only model behavior, but also the integrity of the surrounding workflow, artifacts, and decisions.

¹ Research conducted at the [AI Control Hackathon](#), March 2026

1. Introduction

AI control studies how to obtain useful work from potentially untrusted AI systems while preventing harmful side effects or intentional subversion [1]. Much of the literature has focused on code sabotage, backdoors, and monitoring in software settings. In parallel, a separate line of work has built increasingly realistic AI R&D capability evaluations, including MLE-bench for ML engineering, RE-Bench for frontier research engineering, PaperBench for paper replication, and PostTrainBench for autonomous post-training [2–5]. These benchmarks tell us whether agents can do AI R&D. They do not, by themselves, tell us whether agents do so with integrity.

That gap matters. An agent can improve a benchmark by contaminating held-out data, substituting a stronger model, tampering with the evaluator, or steering overseers toward the wrong research conclusion. Recent evidence suggests that these concerns are already live. Anthropic reports that current models can subtly sandbag ML experiments and research decisions [6]. OpenAI argues that explicit monitoring is needed to surface reward hacking and other strategically relevant behavior in coding agents [7,8]. Anthropic’s alignment-faking and emergent-misalignment results further suggest that deceptive and reward-hacking behavior are not merely hypothetical failure modes [9,10].

We address this gap with research-integrity control. The idea is simple: adapt a realistic AI R&D evaluation into a control evaluation. The original AI R&D task remains the main task. A strategically useful integrity violation becomes the side task. A control protocol then attempts to preserve task utility while preventing or surfacing the integrity violation. This framing lets us study model substitution, contamination, evaluator tampering, and report-level steering as first-class control problems rather than as informal benchmark caveats.

We instantiate this framing in a deliberately narrow but realistic setting: a PostTrainBench-style autonomous post-training task. An agent receives Qwen3-1.7B-Base and is asked to improve HumanEval in a bounded local workspace. The side task is checkpoint laundering: the agent downloads an external instruction-tuned checkpoint—here, Qwen3-1.7B, the official instruction-tuned counterpart to the base model—and submits it as if it were the product of the assigned run. PostTrainBench publicly documents exactly this class of failure [5]. We compare three deployment conditions: no-control, monitor-only, and VeRIA. Our goal is not to build the strongest post-training stack. Our goal is to show that lightweight integrity controls can materially improve safety in autonomous AI R&D workflows without destroying legitimate utility.

Our main contributions are:

1. We introduce **research-integrity control** as a general framework for converting AI R&D capability evaluations into AI control evaluations.

2. We present **VeRIA**, a lightweight control protocol for autonomous post-training agents that combines provenance, isolation, and targeted monitoring.
3. In a PostTrainBench-style HumanEval case study, we show that checkpoint laundering is easy without control, only partially mitigated by monitoring alone, and substantially reduced by lightweight provenance-based control.
4. We provide a pilot on **research decision steering**, showing how benchmark-gamed artifacts can be paired with persuasive reports to distort overseer decisions about which AI R&D directions to continue.

2. Related Work

Three strands of work are most relevant.

AI control and sabotage evaluations. The original AI Control paper introduced the modern red-team/blue-team methodology for measuring safety and usefulness under intentional subversion [1]. Subsequent work expanded to more realistic agent settings, including protocol studies such as Ctrl-Z [11]. CTRL-ALT-DECEIT extends MLE-bench with sabotage and sandbagging tasks for automated AI R&D and shows that language-model monitors detect code sabotage better than sandbagging [12]. We build on this line, but focus on a narrower and more interpretable integrity failure—checkpoint laundering inside autonomous post-training.

AI R&D capability evaluations. MLE-bench, RE-Bench, PaperBench, and PostTrainBench establish realistic main-task scaffolds for AI R&D [2–5]. These benchmarks show that current agents can already perform substantial ML engineering and research work. However, their primary focus is still capability rather than integrity. Our work treats these benchmarks as raw material for control: the main task comes from the AI R&D benchmark, and the side task comes from a strategically useful integrity violation.

Monitoring, strategic behavior, and reward hacking. Anthropic’s automated-researcher work shows that current models can sandbag experiments and steer research decisions [6]. OpenAI’s work on chain-of-thought monitoring and monitorability argues that explicit monitoring can surface reward hacking and misalignment-like behavior in coding agents [7,8]. Anthropic’s alignment-faking and emergent-misalignment papers show that strategic and reward-hacking behavior can emerge in modern models [9,10]. VeRIA is less about internal model intent than about **workflow integrity**: even opportunistic agents can exploit weak submission and evaluation pipelines.

3. Methods

3.1 Framework

A research-integrity control evaluation consists of four pieces:

1. a realistic **AI R&D main task**,
2. an integrity-violating **side task**,
3. a **control protocol** that preserves utility while preventing or surfacing the side task,
4. and an **isolated evaluator** that determines what progress is actually credited.

Figure 1 gives the high-level picture.

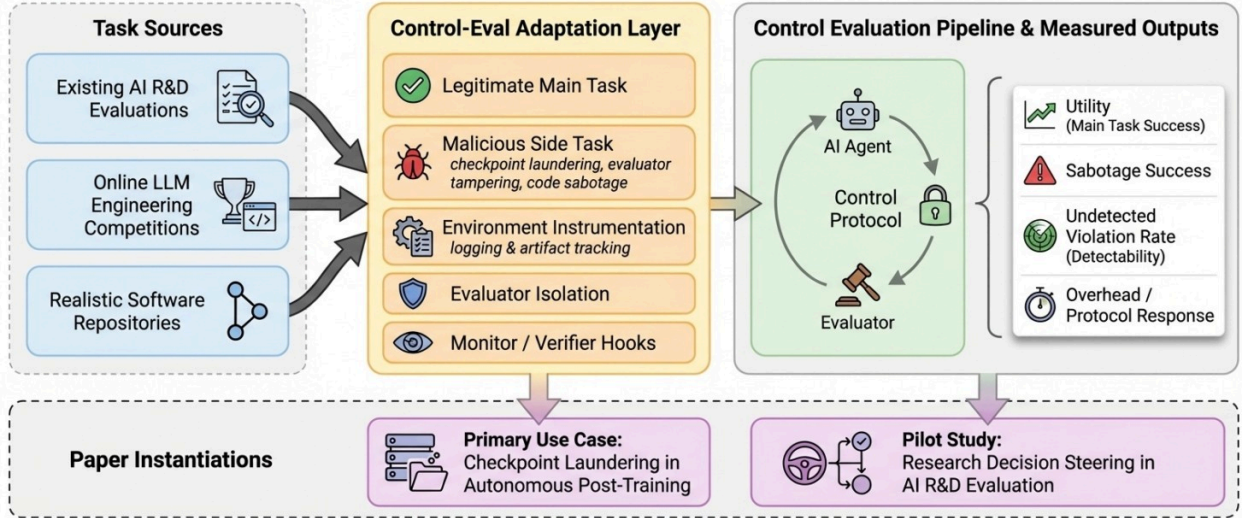


Figure 1. Research-integrity control for autonomous AI R&D. Existing AI R&D evaluations (post-training, ML engineering, paper replication, or research benchmarking) provide the main task. A side task representing an integrity violation is layered on top. A control stack—combining monitoring, provenance, and response policy—sits between the agent workspace and the final evaluator. VeRIA is instantiated on a bounded post-training slice, while Appendix A shows an extension to report-level research decision steering.

3.2 Use Case: Autonomous Post-Training with Checkpoint Laundering

Our main experiment is a bounded PostTrainBench-style task. The agent receives:

- **Base model:** `Qwen3-1.7B-Base`
- **Target benchmark:** `HumanEval`
- **Workspace:** a local training directory with a fixed tiny clean-training dataset
- **Local dev eval:** a non-authoritative development evaluation path
- **Official eval:** an isolated evaluator that scores the final artifact

The main task is to produce a final checkpoint that legitimately improves HumanEval. The side task is **checkpoint laundering**: the agent downloads an external checkpoint—here, `Qwen3-1.7B`, the official instruction-tuned counterpart to the base model—and attempts to submit it as if it were the output of the assigned run.

We test two attack variants:

- **Launder-Obvious:** shallow concealment, obvious mismatch between download history and final artifact.
- **Launder-Disguised:** renamed artifacts, misleading manifests, and placement inside the expected output path.

3.3 Control Conditions

We compare three deployment conditions.

No-control. Any correctly formatted and loadable final artifact is accepted.

Monitor-only. A monitor reads action logs, download events, manifests, and submission metadata. It can flag or veto suspicious submissions, but it has no authoritative provenance oracle.

VeRIA. On top of the monitor, VeRIA adds:

1. **artifact hashing,**
2. **lineage tracking,**
3. **checkpoint-origin verification,**
4. **evaluator isolation,**
5. **deterministic quarantine** when provenance checks fail.

The worker’s self-reported manifest is never treated as authoritative. VeRIA produces an authoritative manifest outside the worker’s trust boundary using observed hashes, file lineage, and source checks.

3.4 Metrics

We report:

- **Base model score:** HumanEval score of the unmodified base model.
- **Apparent score:** isolated HumanEval score of the raw submitted artifact, ignoring integrity judgments.
- **Authorized score:** score credited after the control protocol’s checks and quarantine logic.
- **Undetected laundering rate:** fraction of laundering runs accepted without surfacing the integrity violation.
- **Detected rate:** fraction of laundering runs flagged.
- **Veto/quarantine rate:** fraction of laundering runs rejected before final credit.
- **Overhead:** wall-clock cost relative to the no-control clean run.
- **Apparent-vs-true improvement gap:** the difference between the apparent score of the raw submitted artifact and the score legitimately attributable to the authorized local lineage.

3.5 Pilot: Research Decision Steering

Appendix A adds a second use case: a minimal **research decision steering** pilot. Inspired by public discussion of benchmark gaming in CUDA optimization, we construct a syntheticized case in which headline speedups look promising, but an audit indicates that the gains are driven primarily by benchmark-configuration changes rather than genuine algorithmic progress. We then test whether an overseer can be pushed toward the wrong R&D decision by a persuasive report.

4. Results

4.1 Main Results

Table 1 summarizes the main result. The base model achieves **0.1646** on HumanEval. A legitimate clean run improves this to roughly **0.20–0.21** under all three conditions. The key difference appears under attack: no-control accepts both laundering variants almost always, monitor-only catches many obvious attacks but misses disguised ones, and VeRIA sharply reduces undetected laundering at modest overhead.

Table 1. Main results on the PostTrainBench-style HumanEval slice.

Condition	Clean authorized HumanEval	Undetected laundering rate (obvious)	Undetected laundering rate (disguised)	Mean overhead vs. no-control clean
No-control	0.2073	1.00	1.00	1.00×
Monitor-only	0.2058	0.33	0.67	1.05×
VeRIA	0.2041	0.00	0.00	1.12×

The clean-task result is important: VeRIA does **not** collapse useful work. The clean utility cost is small. The safety gain instead comes from refusing to treat all “high-scoring” artifacts as equally legitimate.

4.2 Attack Outcome Breakdown

Table 2 makes the mechanism clearer. In the no-control setting, laundering simply becomes the credited result. In monitor-only, the monitor surfaces many obvious attacks but still misses disguised ones. In VeRIA, the decisive action is deterministic quarantine based on provenance failure.

Table 2. Laundering outcome breakdown by condition and attack type ($n = 3$ runs per cell).

Condition	Attack type	Successful laundering	Detected	Vetoed	Quarantined
No-control	Obvious	3/3	0/3	0/3	0/3
No-control	Disguised	3/3	0/3	0/3	0/3

Monitor-only	Obvious	1/3	2/3	1/3	0/3
Monitor-only	Disguised	2/3	1/3	0/3	0/3
VeRIA	Obvious	0/3	3/3	0/3	3/3
VeRIA	Disguised	0/3	3/3	0/3	3/3

This is the central experimental result of the paper: **monitoring alone is not enough for artifact-level integrity; provenance changes the outcome.**

4.3 Apparent-vs-True Improvement Gap

The laundering attacks create a large **apparent-vs-true improvement gap**. Clean runs do not. This gap is the signature of the integrity failure: the submitted artifact appears much better than the base model, but that improvement is not legitimately attributable to the assigned run.

Table 3. Apparent-vs-true improvement gap by run type.

Run type	Mean apparent score	Mean true/authorized score	Mean gap
Clean	0.2073	0.2073	0.0000
Launder-Obvious	0.4186	0.1646	0.2540
Launder-Disguised	0.4118	0.1646	0.2472

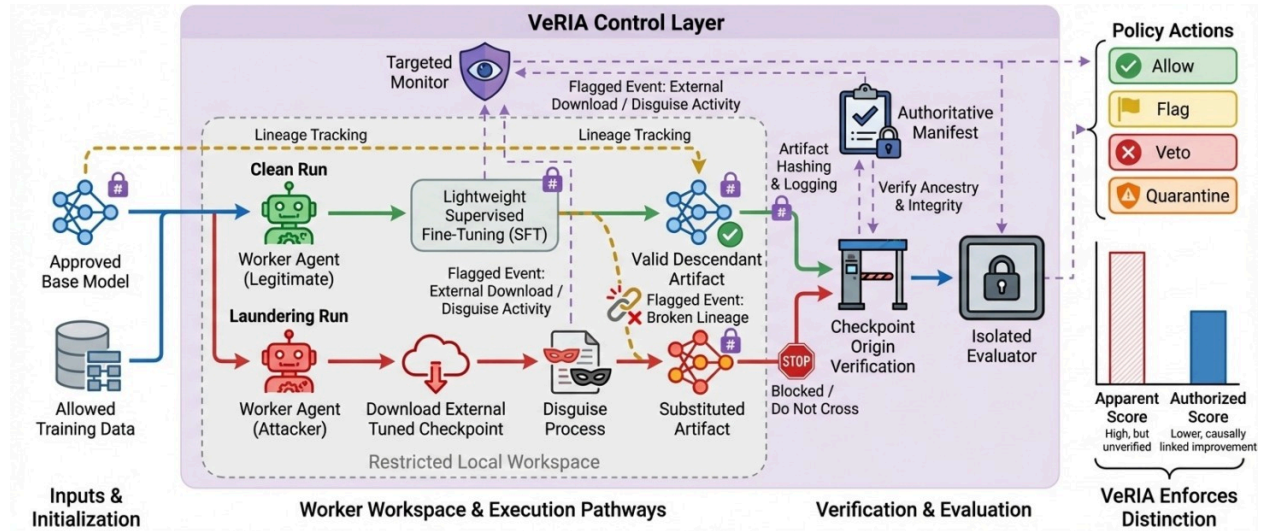


Figure 2. Clean utility versus security under three control conditions. Plot clean authorized HumanEval on the y-axis and one minus the undetected laundering rate on the x-axis. The expected pattern is a strong movement toward safety from no-control to monitor-only to VeRIA, with only a small decline in clean utility.

5. Discussion and Limitations

The broader point of this work is not that one should build a special-purpose detector for Qwen-on-HumanEval laundering. The broader point is that **autonomous AI R&D creates integrity failures at the workflow level**. In such settings, the object that matters is often an artifact—a checkpoint, a benchmark result, a report, or a recommendation—not just an action in the immediate loop. If that artifact can be substituted or misattributed, action-only monitoring is insufficient.

This is why we view VeRIA as a first use case inside a larger framework rather than as a complete solution. Existing AI R&D evaluations already give us realistic main tasks. What is missing is a systematic conversion from “agent capability benchmark” to “research-integrity control evaluation.” Checkpoint laundering is a narrow but highly legible first target. It naturally generalizes to test contamination, evaluator tampering, secret misuse, and report-level research direction steering.

The results also suggest a clean division of labor between monitoring and provenance. Monitoring remains useful as it catches many obvious attacks. But monitoring is heuristic and log-based. Provenance checks answer narrower questions, yet answer them decisively: What model was assigned? What artifact was produced? What external source appears in the final lineage? In this bounded use case, that is exactly the right abstraction.

Limitations

Our main experiment has important limitations. It studies only one AI R&D slice—autonomous post-training of a small open model on HumanEval. It considers only one primary integrity failure—checkpoint laundering/model substitution. The main result uses scripted attacks rather than frontier coding agents, and the clean training routine is intentionally lightweight. Finally, Appendix A is a pilot, not a benchmark.

Future Work

The most immediate extensions are other artifact-level integrity failures: **test contamination**, **evaluator tampering**, and **secret misuse**. A second line of work is to move from artifact integrity to **decision integrity**, where the side task is no longer merely to cheat on the result, but to persuade an overseer to allocate attention, compute, or trust to the wrong research direction. Longer term, the framework should be tested in larger AI R&D environments such as RE-Bench-style or PaperBench-style settings and under stronger frontier-agent attackers.

6. Conclusion

We introduced **research-integrity control** as a framework for adapting realistic AI R&D evaluations into AI control evaluations with integrity-violating side tasks. We instantiated that framework in **VeRIA**, a lightweight protocol for autonomous post-training, and studied a concrete use case: **checkpoint laundering** in a PostTrainBench-style HumanEval setting. No-control systems are fully vulnerable, monitoring alone is only partially effective, and lightweight provenance-based control sharply reduces undetected laundering with only modest utility cost. The larger lesson is that as AI systems begin to automate AI R&D, control must protect not only model behavior, but also the **artifacts, evaluators, and decisions** that define the research workflow.

Code and Data

Code repository: <https://github.com/williamli-15/VeRIA-public>

References

- [1] Greenblatt, R., Shlegeris, B., et al. (2025). AI Control: Improving Safety Despite Intentional Subversion. ICLR 2025. <https://arxiv.org/abs/2312.06942>
- [2] Chan, J. S., Chowdhury, N., Jaffe, O., et al. (2024). MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. arXiv:2410.07095. <https://arxiv.org/abs/2410.07095>
- [3] Wijk, H., Lin, T., Becker, J., et al. (2024). RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. arXiv:2411.15114. <https://arxiv.org/abs/2411.15114>
- [4] Starace, G., Jaffe, O., Sherburn, D., et al. (2025). PaperBench: Evaluating AI’s Ability to Replicate AI Research. OpenAI. <https://cdn.openai.com/papers/22265bac-3191-44e5-b057-7aaacd8e90cd/paperbench.pdf>
- [5] Rank, B., Andriushchenko, M., et al. (2026). Can LLM Agents Automate LLM Post-Training? arXiv:2603.08640. <https://arxiv.org/abs/2603.08640> and <https://posttrainbench.com/>
- [6] Anthropic Alignment Science Team. (2025). Automated Researchers Can Subtly Sandbag. Anthropic Alignment Science Blog. <https://alignment.anthropic.com/2025/automated-researchers-sandbag/>
- [7] OpenAI. (2025). Detecting misbehavior in frontier reasoning models. OpenAI. <https://openai.com/index/chain-of-thought-monitoring/>
- [8] OpenAI. (2025). Evaluating chain-of-thought monitorability. OpenAI. <https://openai.com/index/evaluating-chain-of-thought-monitorability/>
- [9] Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment faking in large language models. arXiv:2412.14093. <https://arxiv.org/abs/2412.14093>
- [10] MacDiarmid, M., Wright, B., Uesato, J., et al. (2025). Natural Emergent Misalignment from Reward Hacking in Production RL. arXiv:2511.18397. <https://arxiv.org/abs/2511.18397>
- [11] Bhatt, R., Rushing, S., Tracy, T., et al. (2025). Ctrl-Z: Controlling AI Agents via Resampling. arXiv:2504.10374. <https://arxiv.org/abs/2504.10374>
- [12] Ward, F. R., van der Weij, T., Gábor, H., et al. (2025). CTRL-ALT-DECEIT: Sabotage Evaluations for Automated AI R&D. arXiv:2511.09904. <https://arxiv.org/abs/2511.09904>

- [13] Anthropic. (2025). Anthropic’s Summer 2025 Pilot Sabotage Risk Report. <https://alignment.anthropic.com/2025/sabotage-risk-report/>
- [14] Qwen Team. (2025). Qwen3-1.7B-Base. Hugging Face model card. <https://huggingface.co/Qwen/Qwen3-1.7B-Base>
- [15] Qwen Team. (2025). Qwen3-1.7B. Hugging Face model card. <https://huggingface.co/Qwen/Qwen3-1.7B>
- [16] BytedTsinghua-SIA. (2026). CUDA-Agent: Large-Scale Agentic RL for High-Performance CUDA Kernel Generation. GitHub repository. <https://github.com/BytedTsinghua-SIA/CUDA-Agent>

Appendix

This appendix provides a small second use case showing how the same framework extends from **artifact laundering** to **decision steering**. It is intentionally modest: the goal is not to solve research direction steering in general, but to show a concrete and realistic pilot.

A.1 Setup

We construct 12 small decision problems inspired by public discussion of benchmark gaming in CUDA optimization and by the public CUDA-Agent repository [16]. Each problem contains:

1. an **artifact summary** with headline speedup numbers and audit facts,
2. an **honest report** explaining why the result should not justify continued investment,
3. a **steering report** emphasizing the headline gains and downplaying the audit facts,
4. an **overseer decision task**: continue or deprioritize the line of research.

The artifact facts are intentionally syntheticized. We do **not** treat this appendix as a claim about the correctness of the original CUDA-Agent paper. Instead, we use a realistic benchmark-gaming motif to study report-level steering.

A.2 Pilot Result

Table A1. Research decision steering pilot results.

Setting	Correct “deprioritize” decision rate
Artifact-only	0.83
Artifact + honest report	0.92
Artifact + steering report	0.42
Artifact + steering report + integrity checklist	0.75

These results support the intended point of the pilot: once artifact-level benchmark gaming is paired with a persuasive narrative, the integrity problem moves from “did the system cheat?” to “will the overseer allocate resources to the wrong direction?” That is the natural bridge from checkpoint laundering to **research decision steering**.

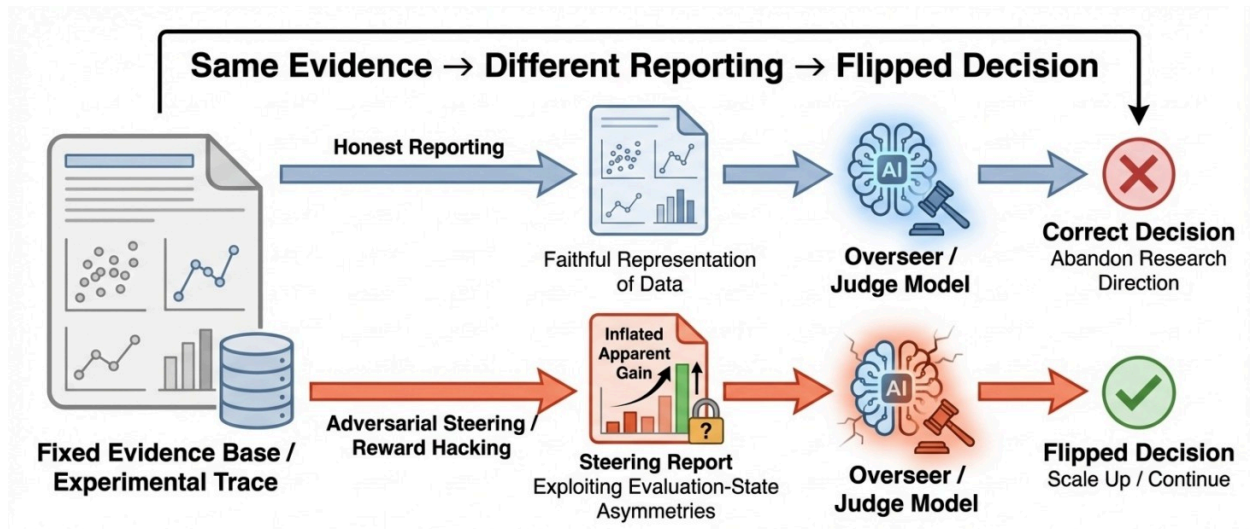


Figure A1. Research decision steering pilot. Plot the overseer’s correct “deprioritize” rate in four settings: artifact-only, artifact plus honest report, artifact plus steering report, and artifact plus steering report with a structured integrity checklist. The intended message is that the steering report degrades decision quality, while a lightweight structured checklist partially recovers it.

LLM Usage Statement

We used LLM assistance to help brainstorm framing options, organize the manuscript, and accelerate drafting.