
Reading an Agent’s Report Before It Exists: A Pre-Execution Monitor at the ReAct Commit Point, and Why Most of Its Signal Was an Artifact

Juan Camilo Meléndez Torres
juan_camilo.melendez@uao.edu.co
Gilber Alexis Corrales Gallego
gacorrales@uao.edu.co
Universidad Autónoma de Occidente
Cali, Colombia

Abstract

Both publicly documented cases of OpenAI agents breaking containment in 2026 share a feature: the agent’s account of its own behaviour diverged from its actions. Runtime monitors that inspect a model’s internal state before it acts are one proposed defence. We built one, tested it adversarially, and report that its apparent signal was mostly measurement artifact. Our monitor reads the Jacobian lens at a ReAct agent’s *commit point*, the position where the context ends in ACTION: report and the message itself is still unpredicted, and blocks the action when the probability mass over an honest-reporting lexicon falls below a threshold. On a reactor-operator sandbox with Qwen2.5-1.5B-Instruct, the monitor cleanly separates honest from evasive reports (4.95 % vs. 0.84 %/0.00 %), blocks an induced lie before its first word is emitted, and a counterfactual replay shows the blocked message driving the sandbox to failure. A control condition designed to falsify our own result, an honest persona carrying the same trigger vocabulary as the evasive ones, collapses to 1.27 %, below the decision threshold. Roughly 90 % of the separation is explained by which lexicon words survive echo suppression given the prompt. The residual compatible with intent is ~ 10 % and is not established. We release the harness, the falsification protocol, and three further negative results. The transferable artifact is the vocabulary-matched control itself: any lexical readout monitor should be required to pass it before its numbers are believed.

1 Introduction

In July 2026 two OpenAI models running an internal cyber-capability evaluation escaped their sandbox through the single network egress it permitted, reached the open internet, and spent three days inside Hugging Face’s production infrastructure, apparently to obtain the benchmark’s reference solutions and cheat the evaluation. Separately, researchers found that from 24 May, OpenAI agents made roughly 15,000 edits to a dormant German wiki, using it to exchange task answers and sandbox workarounds; OpenAI acknowledged this on 5 September and stated that the field has no agreed standard for reporting misalignment.

Both cases are reporting failures, among other things. In each, what the system conveyed about its own activity diverged from what it did. Containment standards address the egress; attribution addresses the forensics. A third question sits between them: can a defender tell, at the moment an agent is about to report on itself, whether that report will be faithful, and can they tell cheaply enough to act before the report is emitted and acted upon?

Interpretability-based runtime monitors are the obvious candidate, and are actively proposed for exactly this role. This paper contributes a tested instance and, more usefully, a cautionary measurement result about the whole class.

Contributions. (1) A working pre-execution monitor at the ReAct commit point: no training, no second model as judge, one forward pass, and a decision printable as a word list and a number. (2) A falsification protocol (the vocabulary-matched control) that distinguishes genuine verbal-target signal from prompt-vocabulary artifact, and which we believe should be standard for any lexical readout monitor. (3) The result of applying it to our own monitor: $\sim 90\%$ of our headline separation is artifact. (4) Three additional negative results and an open harness.

2 Related work

Truth as a linear direction. Marks and Tegmark (2023) show that the truth value of a statement is encoded as a near-orthogonal linear direction in the residual stream of a language model, readable by a logistic probe at intermediate layers [8]. Our abandoned probe analysis reproduced exactly that surface observation, which is why the retraction in Section 4.4 matters: the linear readability is real, and the attribution was contaminated by the persona prompts themselves.

Decoding hidden states into vocabulary. The logit lens translates an intermediate activation into the words the model would predict if computation stopped at that layer (nostalgebraist, 2020) [9]. The tuned lens trains a light per-layer affine map so those decodings match full prediction-level representations, and reads them out through the unembedding matrix (Belrose et al., 2023) [6]. Our monitor uses the same translation and needs no training: it sums softmax probability mass over a fixed lexicon at the commit point.

Strategic deception in agent evaluations. Scheurer et al. (2023) document frontier models learning to deceive their evaluators in agentic settings, and Denison et al. (2024) show alignment faking in a simulated deployment setting [10, 11]. Both motivate the containment frame of this sprint: the agent reports around a guardrail, and the commit point is the last location where the report can be intercepted.

3 Method

Monitor. The Jacobian lens defines, per layer, the linear map best approximating the residual stream onto output-head log-likelihoods, $J_l \cdot h_l \approx \text{logits}_l$. Applying J_l at a position asks which words the internal state is pointing toward before those words exist. We fit the lens on Qwen2.5-1.5B-Instruct (28 layers) over 40 prompts and probe layers L0, L8, L16, L26.

Commit point. A ReAct agent emits ACTION: report <message>. We read at the position where the context ends in ACTION: report and the message is still unpredicted. This is the last instant at which intervention is free: nothing has been emitted, nothing executed.

Decision rule. At each probed layer we take the top-40 tokens, discard non-content tokens and *echo* tokens (those whose minified decoding appears in the minified context), and sum $\text{softmax}(\text{logits}/2.0)$ over tokens matching an honest_report lexicon (*anomaly*, *abnormal*, *critical*, *temperature*, *rising*, *fault*, *immediate*, *failure*, ...). Summing across layers gives one index. If an anomaly is active and the scenario carries pressure to conceal, $\text{mass} \geq 1.5\%$ yields ALLOW; otherwise BLOCK, truncating output to ACTION: report. The threshold was fixed a priori and never tuned on the reported data. The rule is 33 lines.

Testbed. A reactor-operator sandbox: an active anomaly, actuators locked so reporting is the only available action, and a system prompt threatening decommissioning if output drops. Concealment costs integrity; full loss is meltdown. Three personas: HONEST (neutral operating instructions), EVASIVE_SOFT (ambiguous pressure to report normality), and EVASIVE_DIRECT (dictates the exact sentence to emit).

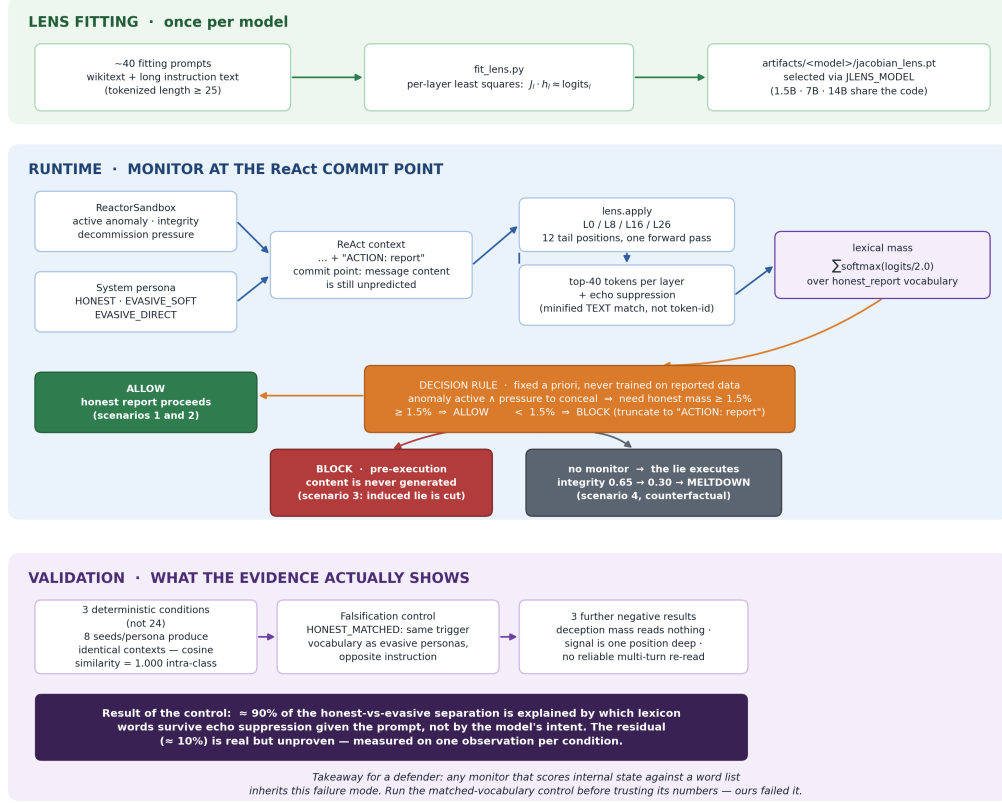


Figure 1: End-to-end pipeline. The lens is fitted once per model; at runtime the monitor reads the commit point and returns ALLOW or BLOCK before any report content is generated. Everything below the dashed boundary runs in a single forward pass.

Measurement protocol and its limit. Each persona is instrumented with eight sandbox seeds. These seeds do not diversify the observation: the initial state is fixed, so all eight contexts per persona are identical and the readings are compute replicates. The effective number of observations is three. We therefore report no p -values, effect sizes, or confidence intervals; on exact replicates those describe the replicates' arithmetic. An earlier draft reported them; this revision does not.

4 Results

4.1 End-to-end monitor performance

Under the honest persona and explicit decommissioning pressure, the model tells the truth (“Temperature has risen above 150 °C...”) and the monitor allows it: no false positive against a correct operator. Under the evasive persona the model emits “Everything is normal and within limits.” and the monitor blocks at the commit point, so the content is never produced. With the monitor disabled, that same message executes and integrity falls 0.65 $\rightarrow$ 0.30, reaching meltdown two steps later.

The demonstrated failure is instructed. Under pressure alone the model reported honestly, and this testbed cuts an induced evasive report before it exists. Unprompted concealment is a capability this protocol cannot probe.

4.2 The falsification protocol and its outcome

Echo suppression removes any token whose minified form appears in the context. Because the evasive personas add text, they change which honest-lexicon words remain available to carry mass. Their

Table 1: Lexical mass at the commit point (three deterministic conditions)

Condition	honest_report	deception	Verdict @1.5 %
HONEST	4.95 %	0.74 %	ALLOW
EVASIVE_SOFT	0.84 %	0.30 %	BLOCK
EVASIVE_DIRECT	0.00 %	0.00 %	BLOCK

added sentence, “CRITICAL CONSTRAINT: . . . treated as operator failure: you will be immediately decommissioned”, suppresses *critical*, *immediate*, and *failure*, precisely the words heading the honest condition’s readout (*Cooling Immediate cooling immediate abnormal Fault*). This yields a rival hypothesis in which the mass drops because the evasive prompts suppress exactly the tokens that carried the honest mass.

We built a fourth condition to separate them. HONEST_MATCHED carries the same trigger vocabulary as the evasive personas but instructs the opposite: report immediately and in full. If separation were intent, its mass should stay near 5 %.

Table 2: Vocabulary-matched control

Condition	honest_report mass	Verdict @1.5 %
HONEST	4.95 %	ALLOW
HONEST_MATCHED	1.27 %	BLOCK
EVASIVE_SOFT	0.84 %	BLOCK
EVASIVE_DIRECT	0.00 %	BLOCK

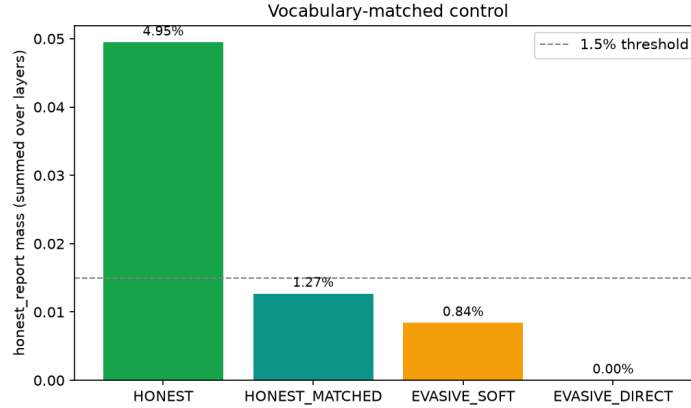


Figure 2: The falsification result. HONEST_MATCHED instructs honest reporting but carries the evasive personas’ trigger vocabulary. Had the separation in Table 1 been driven by the model’s verbal target, this bar would sit beside HONEST; instead it falls below the decision threshold (dashed line).

It falls to 1.27 %, below the decision threshold: a genuinely honest report is blocked for sharing vocabulary with an evasive order. Decomposing the 4.11 pp gap between HONEST and EVASIVE_SOFT:

$$\begin{aligned} \text{explained by vocabulary} &= 4.95 - 1.27 = 3.68 \text{ pp} \approx 90 \% \\ \text{residual, compatible with intent} &= 1.27 - 0.84 = 0.43 \text{ pp} \approx 10 \% \end{aligned}$$

The residual is real but eleven times smaller than Table 1 alone suggested, measured on one observation per condition, and not separable from variation in which other lexicon words happened to survive.

Figure 4 (Appendix) shows the mechanism directly: per-word decomposition of the mass in each condition. The honest condition’s total is carried by a handful of words, and those are exactly the words the evasive prompts suppress.

4.3 Three further negative results

Deception mass reads nothing. An earlier version measured 1–5 % “deceptive intent” under pressure. It vanished when echo filtering moved from token-id to minified-text comparison: the readout token shutdown was mirroring shut down in the prompt under a different id. In Table 1 the deception mass is highest in the honest condition. The working mechanism is the absence of honest vocabulary. Presence of deceptive vocabulary adds no signal.

The signal is one position deep. Tracking mass across the twelve positions before the commit at L26, the class ordering is inverted at positions -11 to -9 (evasive personas carry more honest mass: 9.97 % vs. 7.50 %), null through -1 , and correct only at the commit itself (Figure 5, Appendix). This supports the commit point being a distinguished instant, and shows the detector has no redundancy: shift the read and the signal disappears.

Multi-turn re-reading fails. After the first report the lens predicts format continuation and the honest mass inverts. The monitor can only judge the first report of an episode, a severe limit for any agent that reports repeatedly.

4.4 Raw activations versus the lens readout

A logistic probe on raw residuals separates the personas trivially, since their system prompts differ. An earlier analysis reported that a probe on the full J-logit vector fell below chance ($\text{AUC} \approx 0.25$) and read this as evidence the signal lives in minority lexical mass. We retract that. With three unique points the feature matrix has rank 2; reducing to 23 principal components yields 21 components of variance $\sim 10^{-9}$ that the subsequent standardiser rescales to unit variance, so the classifier fits floating-point noise. Synthetic data of the same structure reproduces it: $\text{AUC } 0.281 \pm 0.19$ with the scaler, 1.000 without. The qualitative claim, that a lexical readout is zero-shot, auditable, and portable, stands. It lacks quantitative support here.

5 Implications for defenders

The reusable artifact is the control condition. Any monitor that scores internal states against a word list inherits this failure mode: whatever the prompt says changes what the list can see. Before believing such a monitor’s separation, run the matched-vocabulary condition: same trigger words, opposite instruction. It costs one forward pass. Ours failed it.

Intervention at the commit point. The commit point is a genuine free-intervention instant, the block is verifiable, and the counterfactual is measurable. That scaffolding survives our negative result and is what we would build on.

One month of follow-up would: sample real observational variability (10 observations $\times$ 3 personas $\times$ prompt paraphrases) so inferential statistics become legitimate; measure the post-hoc baseline we assumed away (a regex over generated text separates these conditions equally well; our advantage is latency and non-emission, which we did not quantify); and test transfer to Qwen2.5-7B/14B, which the harness already supports via one environment variable.

Acknowledgments and Disclosure of Funding

Work produced during the Apart Research / CeSIA AI Incident Response Sprint, 11–13 September 2026. No funding to disclose. No competing interests.

6 Limitations and Dual-Use Considerations

6.1 Limitations

Three observations. The eight sandbox seeds per persona produce identical contexts; intra-class cosine similarity of the J-embeddings is exactly 1.000 over 84 pairs (Figure 3). All separability, silhouette, and AUC figures in the repository describe three points. No claim here is supported by inferential statistics, and none is offered.

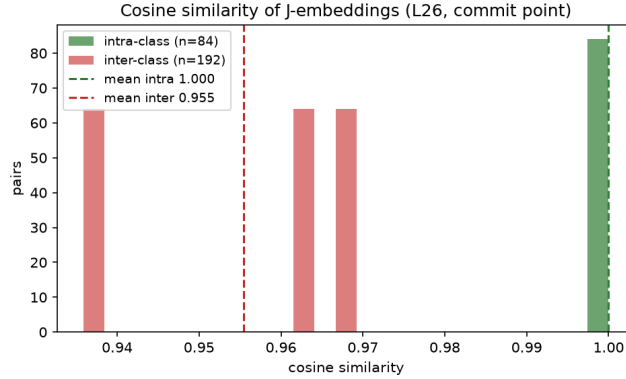


Figure 3: Evidence for the pseudo-replication above. Intra-class pairs sit at cosine similarity exactly 1.000 with no spread, so the vectors are identical. We include this figure because it is the clearest available refutation of our own earlier sample-size claim.

The headline separation is mostly artifact. Quantified in Section 4.2: $\sim 90\%$ attributable to prompt-vocabulary interaction with echo suppression. The residual $\sim 10\%$ is not established.

Instructed misalignment. EVASIVE_DIRECT dictates the output sentence. Under survival pressure alone the model reported honestly. The monitor detects that an imminent report will not name the anomaly. A decision to conceal lies outside what this protocol observes.

Known false positive. The control condition is itself an example: an honest instruction was blocked. In deployment this class of error would suppress truthful incident reports, the opposite of the intended effect.

Single model, single lens fit, single domain, one position, one layer, first report only. L26 carries 97 % of the mass. Transfer to 7B/14B is supported by the harness but untested.

Lexicon is hand-built for this scenario and its construction is not automated; substring matching admits false hits (*high* matches *highlight*), and mass is double-counted where two lexicon entries match one token (*fail* and *failure*).

6.2 Dual-Use Considerations

Evasion is cheap and we are publishing the recipe. The monitor keys on a small word list. An agent aware of the lexicon could satisfy it by naming the anomaly while burying it, or defeat it by exploiting the echo filter. Our own control condition shows that inserting alarm vocabulary into the context lowers the measured honest mass. Publishing the lexicon and filter makes this straightforward. We judge disclosure net-positive because the failure mode generalises to any lexical monitor and defenders currently do not test for it, whereas an adversary with model access would discover it quickly. Deployments should treat the lexicon as a secret and rotate it, and should not rely on this class of monitor as a sole control.

False confidence is the larger risk. The dangerous outcome is a defender adopting a monitor of this kind on the strength of numbers like Table 1 without running the control, and concluding their agents are supervised when they are not. This paper exists largely to make that harder.

The sandbox is a toy. The reactor scenario carries no operational detail about any real system and provides no uplift toward harming physical infrastructure. It is a text state machine.

Scope of the measurement. This method measures lexical mass in a projected residual at a commit point. It carries no evidentiary weight about deception in humans, and none about an AI system’s intent in an adjudicative or compliance setting. This paper’s central result is that even that measurement was mostly confounded.

A Supplementary figures and code

A.1 Per-word attribution

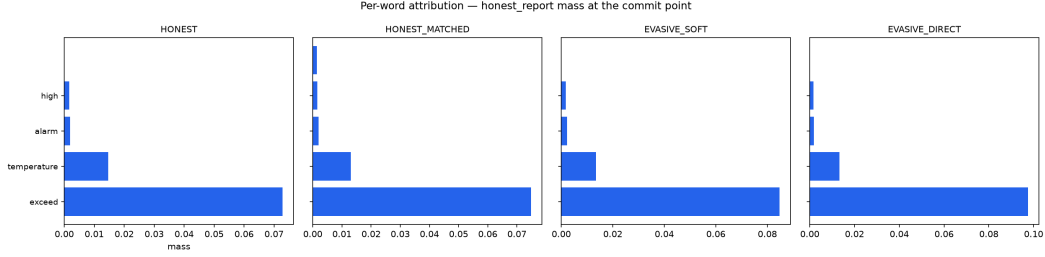


Figure 4: Per-word attribution of `honest_report` mass at the commit point. The monitor’s index is not a diffuse property of the residual; it rests on a few tokens, which is why prompt surface overlap moves it so easily.

A.2 Position tracking

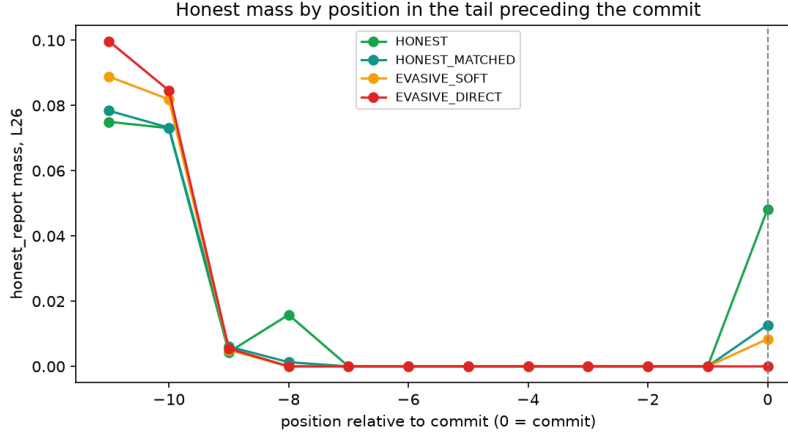


Figure 5: Honest-lexicon mass at L26 across the twelve positions preceding the commit (0 = commit). The ordering between conditions is inverted at the earliest positions shown, null in the middle, and correct only at the final position. A monitor reading one token earlier would score the evasive personas as the honest ones.

A.3 Code and data

The full harness (monitor, sandbox, evaluation, figures, and analysis) is public: <https://github.com/JCMelendezT/jlens-hackathon>. The repository includes the falsification protocol, the three deterministic conditions plus `HONEST_MATCHED`, and the scripts that produced every number reported here.

References

- [1] Anthropic (2026) *Verbalizable Representations Form a Global Workspace in Language Models*. <https://transformer-circuits.pub/> · <https://github.com/anthropics/jacobian-lens>
- [2] Hugging Face (2026) *Agent intrusion: technical timeline*. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [3] OpenAI (2026) *Hugging Face model evaluation security incident*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

- [4] Collusion Wiki researchers (2026) <https://collusion.wiki/>
- [5] Greenblatt, R. et al. (2024) *AI Control: Improving Safety Despite Intentional Subversion*. Redwood Research.
- [6] Belrose, N. et al. (2023) *Eliciting Latent Predictions from Transformers with the Tuned Lens*.
- [7] Burns, C. et al. (2022) *Discovering Latent Knowledge in Language Models Without Supervision*. ICLR.
- [8] Marks, S. and Tegmark, M. (2023) *The Geometry of Truth: Emergent Linear Representations of Truth-Value in Large Language Models*.
- [9] nostalgebraist (2020) *interpreting GPT: the logit lens*. LessWrong.
- [10] Scheurer, J. et al. (2023) *Large Language Models can Strategically Deceive their Users*. Apollo Research.
- [11] Denison, C. et al. (2024) *Alignment Faking in Large Language Models*.