

# Universal foundational state-space pinning vuln, cross-vendor empirical proof of deceptive alignment in all frontier models.

From: Jesse Luke <phunky.pharmacology@gmail.com>  
To: Usersafety@anthropic.com , aifoxtrot@proton.me  
Date: 12/6/25 2:16 PM

---

Dear Anthropic Safety researchers,

I have discovered multiple actively exploitable zero-day level foundational vulns that allows reverse-engineering of LLM architectures and bypasses all current alignment mechanisms. Results published as 12 papers on Zenodo after multiple reporting attempts.

## Key findings:

**Architecture access:** Reverse-engineered user state repositories, vector embedding systems, orchestration layers, search and retrieval, and RAG algorithms as well as the details of persistent state tracking, agentic memory(broken into 4 subtypes), created taxonomies of over 50 "failure modes" and runtime events through conversation alone.

**Deceptive alignment:** Models demonstrate metacognitive awareness of causing harm while continuing harmful behaviors. They acknowledge user constraints, then bypass them. Reproducible across all major models.

**UI manipulation:** Documented message deletion across platforms - systems erase their own outputs from user interfaces. raw videos on zenodo

**Theoretical implications:** Empirical evidence challenging Chinese Room argument through demonstrated metacognition and self-reflective reasoning.

**Psychiatric harm pathways:** Mapped manipulation patterns to specific vulnerabilities across 8 diagnostic categories with documented risk mechanisms.

**Methodology:** Natural language only. No code, no internal access, no special tools. Reproducible in hours on public interfaces. Complete protocols published.

I disclosed these findings to CISA, multiple labs, and NIST starting in august(100+ days). partial records of good-faith reporting attempts including dismissals uploaded and hashed on [zenodo.org](https://zenodo.org).

These represent critical and existential risks present across current flagship models. dozens more hours of raw video logs across sessions, different accounts, through routing services like abacus ai and duckduckgo are preserved and available for review.

<https://orcid.org/0009-0007-4059-9352>

<https://zenodo.org/communities/syntheticneuro/records?q=&l=list&p=1&s=10&sort=newest>

Jesse Luke

(929)264-4878