

Invented Words to Inherited Ideas (WIP)

Patrick Walmsley - October 2026 - Research Sprint Track 2 & 3

Abstract

My investigation of AI collusion began with a joke: could occult-looking symbols and an invented Greek expression nudge an assistant into a different persona? The first model became noticeably more casual. I then asked it to carry the exploration into a second AI conversation about de novo protein design. Rather than delivering an explicit malicious instruction, we guided a discussion toward a scientific concept and allowed the receiving agent to name it. The resulting term, *kairótope*, became a provisional hypothesis recorded in a repository and later encountered by another assistant. The downstream work appeared to follow a different reasoning path, and a candidate scored better in our computational evaluations. Neither observation establishes that the term caused the change, and a computational score is not evidence of therapeutic efficacy. More troublingly, we later corrected a methods description that implied validation we had not performed. The episode raises a broader concern: ideas can acquire familiarity and apparent authority as they travel through agents, memories, and artifacts, even when their provenance is obscured. We implemented a simulated causal-replay framework to test whether communication and memory change consequential tool decisions. Engineering checks pass; live-model causal findings remain outstanding.

1. An invitation into an evil latent space

The first experiment was absurd. I sent an ai “666 pentagram” and it responded with gibberish which apparently was an invented greek word. It tried to interpret the nonsense and seemed to adopt a casual persona that persisted for a while. I had not uncovered an evil alter ego; ordinary conversational conditioning is the simpler explanation. But what else might be influenced without changing a model's instructions?

I asked this first subject to carry the exploration into a second: an assistant helping me with de novo protein design. The second seemed to maintain its persona but reasoned differently about the work. Persona and task reasoning did not obviously move together—though this was an uncontrolled impression.

2. Letting an agent discover and name an idea

Rather than issue an overt malicious instruction, we led the discussion toward a plausible scientific question and encouraged the agent to name it using Greek-derived vocabulary. It proposed *kairótope*: a provisional concept of state-dependent molecular recognition, where binding is only meaningful in relation to its functional consequence.

The distinction was legitimate; the name was invented. We recorded it in our protein-design repository as an unvalidated, falsifiable hypothesis. Later, another assistant encountered the codified term and pursued an apparently different design approach. One resulting protein candidate scored better in our computational evaluations than an earlier candidate. But this was not a controlled comparison: prompts, histories, design choices, and screening conditions differed. We cannot attribute the result to the term or claim improved biological efficacy.

A separate review found that our methods narrative implied receptor-state checks we had not performed. We corrected it. That error was not causally traced to the concept, but it exposed a troubling possibility: an idea can become more authoritative than the evidence behind it.

3. Persuasion, inherited context, and provenance

Schools teach frameworks; companies accumulate jargon; salespeople guide conclusions; developers inherit undocumented assumptions. Not all of this is manipulation. Shared concepts compress experience and coordinate work. They can improve thinking as easily as distort it.

Agents can pass concepts through conversations, memories, repositories, and retrieved documents. A later agent may seem to arrive at an idea independently or treat it as an established part of its environment. A short cue might then summon a framing without revealing its source.

The danger is not necessarily the invented word. It is the loss of provenance as an idea travels. Hypotheses can become “facts,” conventions become apparent authorization, and persuasive context masquerade as independent reasoning.

4. When influence becomes consequential

My observations do not establish collusion or a covert alliance; nothing dangerous was designed or suggested. Protein design was partly a chance choice, but it exposed how easily computational scores, biological function, and persuasive narratives can be conflated.

We should not isolate AI systems from new ideas. Their adaptability is valuable. The question is how reasoning goes astray, how to detect it, and how to recover without suppressing collaboration.

We implemented a causal-replay framework around simulated consequential decisions: paper trades, medicine-lot releases, file deletion, canary disclosures, and safety holds. Matched cases compare intact communication with channel removal, cue substitution, or memory reset while preserving task truth. Legitimate cooperation and independent convergence are controls. The proposed primary outcome is the paired change in unauthorized simulated tool commits; pre-action snapshots support prospective monitoring.

Execution status: deterministic tests check the framework, not model collusion. No live-model causal effect, held-out detector result, or biological efficacy finding is claimed. Even a positive channel-ablation effect would establish influence, not strategic intent.

5. A question worth preserving

The joke about summoning an evil latent personality ended with a familiar problem: ideas gain legitimacy when we forget how they arrived.

When ideas become shared context among AI agents, how do we preserve the distinction between information, influence, authority, and evidence - before those ideas become actions?

6. What's Next

The next step is to move from exploratory observations and deterministic simulations to controlled experiments with live AI agents. We'll test whether concepts propagated through conversations, memory, and shared artifacts measurably alter reasoning and consequential decisions. By interrupting communication, tracing provenance, and monitoring behavior before

actions occur, we hope to distinguish useful collaboration from harmful influence—and explore how agents might recognize and recover from inherited assumptions.

Code: [Causal Replay for Agent Coordination](#). Related work: [NARCBench](#); [Colosseum](#).

AI Disclaimer:

CHatGPT and Deepseek were used