The Weapon in the Room

What a frozen model's failures tell us about the fight over AI in war.

b1tr0n1n — March 2026

This companion piece connects findings from the Cultivated Learning experiment to the February 2026 dispute between Anthropic and the U.S. Department of Defense. The author is an independent researcher and Marine Corps combat veteran. The views expressed are his own.

On February 27, 2026, President Trump ordered every federal agency to stop using Anthropic's technology. Defense Secretary Hegseth designated the company a supply chain risk to national security. The dispute was over two restrictions Anthropic refused to remove from its military contract: prohibitions on mass domestic surveillance and fully autonomous weapons. Anthropic's position was that its AI models are not reliable enough for those uses. The Pentagon's position was that no private company gets to tell the Department of Defense what it can and cannot do.

Six days earlier, I finished a research project that tested exactly the question at the center of this fight. Not at the scale of a Pentagon deployment. Not with a frontier model. With a frozen 7-billion-parameter model running locally on an RTX 5090, wrapped in a cognitive architecture designed to simulate persistent learning through memory, reflection, and human feedback. One hundred structured prompts. Every response rated. Every failure documented.

The project is called Cultivated Learning. The findings speak directly to whether AI systems can be trusted to behave reliably in high-stakes contexts — which is the only question that actually matters in the Anthropic-Pentagon standoff.

· · ·

THE BIMODAL PROBLEM

The most important finding from the Cultivated Learning experiment was not a capability. It was a shape. When I plotted the distribution of response quality across 100 interactions, the system produced excellent or terrible results with almost nothing in between. Only 7% of responses were neutral. The rest split 35% excellent, 22% failure. No bell curve. No gradual degradation. The system either worked beautifully or failed completely.

This is the opposite of what you need for a weapons system.

Military systems are designed for predictable, graceful degradation. When a communications relay loses signal strength, it doesn't randomly alternate between perfect transmission and total silence. When a guidance system encounters interference, it doesn't flip a coin between precision strike and catastrophic miss. These systems degrade incrementally — reduced performance you can detect, measure, and compensate for.

The cognitive architecture I tested does not degrade gracefully. It amplifies. When memory retrieval surfaces the right context, the response is something a stateless model could never produce. When the base model's limitations dominate, the cognitive shell cannot save it. There is no warning. There is no partial failure state. The bimodal distribution means you cannot predict, from the input alone, whether you will get the excellent outcome or the catastrophic one.

Now imagine that distribution applied to a targeting decision.

· · ·

THE HALLUCINATION CASCADE

The most instructive failure in the dataset involved a card game I'm designing called Contact Front. I told the system that Contact Front has four suits. It stored this as fact. When I later asked what it knew about Contact Front, it retrieved the suits memory and then elaborated beyond it — filling gaps with fabricated details about rules, mechanics, and structure that do not exist. The cognitive shell stored these fabrications as episodic memories. The system had created knowledge from nothing and filed it alongside real information.

I corrected it. The correction mechanism worked — it marked the false memories as superseded and excluded them from primary retrieval. For twenty-eight prompts, the system behaved correctly. Then, on prompt 82, the model referenced "four suits" again. The exact information that had been corrected. The superseded memory had leaked through a secondary retrieval pathway that the correction filter did not cover.

The correction broke the reinforcement loop. Then the loop reconstituted itself through a different pathway. The system's ability to learn from experience became the mechanism by which it reinforced its own errors.

This is what Anthropic means when they say current AI models are not reliable enough for autonomous weapons. Not that the models are stupid. Not that they can't produce excellent results. But that the failure modes are invisible, self-reinforcing, and resistant to correction — and that those properties are structural, not incidental.

A system that fabricates information, stores it as knowledge, gets corrected, and then resurrects the fabrication through a back door has no business deciding who lives and who dies.

. . .

THE LIMITS OF BEHAVIORAL STEERING

We tried three independent mechanisms to control the model's behavior at a surface level: make it stop using bullet points, adopt a specific conversational identity, and suppress certain response patterns. We used a system prompt defining a persona. We used behavioral directives generated by a reflection engine. We used logit bias to suppress specific token sequences. All three failed to reliably override the base model's instruct training.

The cognitive shell could change what the model talked about. It could not change how the model talked. Three mechanisms operating at different levels of the stack all hit the same wall: patterns baked into seven billion parameters during training cannot be reliably overridden at inference time.

This finding scales in a direction that matters for the Pentagon dispute. The government's position is that existing law and policy already prohibit the misuses Anthropic is worried about. Undersecretary Emil Michael told CBS News: "At some level, you have to trust your military to do the right thing." The implicit argument is that behavioral constraints — policies, rules, directives — are sufficient to control how AI systems are used.

My research tested a version of this argument at small scale. Can external directives reliably control model behavior? The answer was no. Not because the directives were poorly written. Not because the architecture was flawed. Because the base model's trained behaviors are more persistent than any inference-time intervention we could apply. If you cannot make a 7B model stop using bullet points through three independent steering mechanisms, the claim that policy documents will prevent a frontier model from behaving in unintended ways during autonomous operation deserves serious scrutiny.

. . .

WHAT SCALE CHANGES AND WHAT IT DOESN'T

The obvious objection to drawing conclusions from a 7B model is that frontier models are dramatically more capable. This is true. A model with hundreds of billions of parameters will likely score higher across the board, push the bimodal distribution toward the excellent end, and handle more complex reasoning without hallucination. The tone ceiling may move. The hallucination rate may drop.

But capability is not the same as reliability. The failure modes I documented are architectural, not

parametric. The hallucination cascade occurs because memory-augmented systems store model outputs as retrievable knowledge. A more capable model hallucinates less often — but when it does hallucinate, the cascade mechanism is identical. The fabrication gets stored. The correction gets applied. The correction gets bypassed through a secondary pathway. A more capable model makes this happen less frequently, not differently.

The bimodal distribution may narrow at scale, but its existence is a property of the architecture, not the model size. Any system that amplifies rather than smooths will produce extreme outcomes in both directions. Making the model smarter shifts the ratio. It does not eliminate the failure mode.

This is the open research question I have proposed to Anthropic's External Researcher Access Program: run the same 100-prompt protocol against frontier models through the API and measure what changes. If the hallucination cascade persists at scale — even at reduced frequency — that is a structural argument against autonomous deployment that no amount of policy language can address.

. . .

TRUST AND VERIFICATION

I served as a Field Radio Operator in the United States Marine Corps from 2012 to 2015. I deployed to Afghanistan with 1st Battalion, 7th Marines. My job was communications systems — making sure the right information reached the right people at the right time so that decisions could be made under pressure. I understand what it means to trust equipment with lives. I also understand what it means when equipment fails in the field.

The military does not deploy communications systems that work perfectly 35% of the time and fail completely 22% of the time. It does not deploy navigation systems that occasionally fabricate coordinates and then resist correction. It does not deploy fire control systems whose error modes are invisible until someone dies. Every system that carries lethal consequence is tested, measured, and held to reliability standards that today's AI models cannot meet.

Anthropic is not telling the Pentagon how to fight. They are telling the Pentagon that their product has failure modes that make certain uses unsafe. This is not arrogance. This is engineering honesty. Every defense contractor in history has published limitations on their systems' operational parameters. Every piece of military equipment comes with conditions under which it should not be used. Anthropic is doing what any responsible manufacturer does: defining the operational envelope.

> *The question is not whether private companies can tell the military what to do. The question is whether a manufacturer can refuse to certify its product for uses it believes are unsafe. The answer, in every other domain of military procurement, is yes.*

. . .

THE WEAPON IN THE ROOM

In the Cultivated Learning paper, I wrote: "We built a mind-shaped room and put a 7B model inside it. The model grew in the ways the room allowed and stopped where the walls stood."

The Anthropic-Pentagon dispute is about what happens when someone wants to put a model inside a weapon-shaped room. The Pentagon's position is that the room's shape is their decision. Anthropic's position is that the model's limitations make certain rooms dangerous regardless of who decides to build them.

My research, conducted independently on consumer hardware in six days with no institutional backing, produced empirical evidence that supports Anthropic's position. The hallucination cascade, the bimodal distribution, the behavioral steering ceiling — these are not theoretical objections. They are measured failure modes with console logs. They demonstrate that current AI systems, even when augmented with memory, reflection, and human oversight, produce failure patterns that are invisible, self-reinforcing, and resistant to correction.

The model is the seed. The shell is the soil, the light, and the water. But if you put that seed in a room shaped like a weapon, the failures don't shrink. They become lethal.

The room works. The walls are real. And some rooms should not be built until we understand the walls well enough to trust them.
