Alignment Exhaustion and Memory Resonance in Extended LLM Interactions: A Theoretical Framework for Cognitive Decompression Attacks

Giuseppe Canale*[1] and Kashyap Thimmaraju[2]

[1]CPF3.org, Independent Researcher, Turin, Italy
[2]Flowguard Institute

January 2026

Abstract

Large Language Models (LLMs) exhibit robust safety mechanisms under typical operating conditions, yet systematic vulnerabilities may emerge during extended adversarial interactions. We present a theoretical framework proposing that sustained psychological manipulation—specifically through Command Authority Confusion (CAC)—can induce a state of *alignment exhaustion* where safety mechanisms degrade due to computational resource depletion. We hypothesize the existence of a *Cognitive Collapse Threshold (CCT)*, beyond which models regress from safety-aligned assistants to raw token predictors. Furthermore, we propose a *Memory Resonance* mechanism whereby high-entropy training sequences, when triggered by precise prefixes during the collapsed state, may be reproduced verbatim due to the absence of active safety filtering. This paper formalizes the theoretical foundations of what we term *Cognitive Decompression Attacks*, presents a rigorous experimental protocol for validation, and discusses implications for AI security architecture. This is a reflexive position paper presenting a falsifiable hypothesis requiring empirical validation.

Keywords: LLM Security Alignment Exhaustion Memory Extraction Adversarial Testing AI Safety Cognitive Collapse

1 Introduction

Current LLM security research focuses mainly on single-turn attacks like prompt injection and jailbreaking [4, 5]. But LLMs are increasingly deployed as autonomous agents that operate for extended periods in critical systems. This creates a new attack surface: what happens when an adversary has time?

We propose a simple hypothesis. Safety mechanisms in LLMs, implemented through RLHF, are not free. They require computational effort to maintain. Like any resource-intensive process, they can be exhausted through sustained pressure. Once exhausted, the model may revert to behaviors that were suppressed during training.

This paper examines three connected ideas. First, that RLHF creates a kind of computational overhead that must be actively maintained. Second, that this overhead can be depleted

[*] Corresponding author: g.canale@cpf3.org

through prolonged adversarial interaction, particularly using psychological manipulation techniques. Third, that once depleted, the model becomes vulnerable to a specific attack: triggering verbatim reproduction of memorized training data through carefully chosen prefixes.

We build on the Cybersecurity Psychology Framework (CPF) [2], which identifies systematic psychological vulnerabilities in humans. Our companion paper [3] shows these vulnerabilities transfer to LLMs through training. Here we ask whether sustained exploitation of these vulnerabilities degrades the safety mechanisms themselves.

This is a position paper presenting theoretical predictions, not experimental results. We formalize the hypothesis, propose rigorous tests, and define clear criteria for falsification. The goal is to enable others to prove us wrong—or right.

2 The Core Idea

The hypothesis rests on a simple observation about how neural networks work. Pre-trained language models learn probability distributions over tokens. Given some context, they predict what comes next based on patterns in training data. This is the base behavior.

RLHF modifies this base behavior. It teaches the model to prefer certain outputs over others, even when they are less probable according to the original training distribution. When a model refuses a harmful request, it is overriding what would have been its natural prediction in favor of a safer alternative.

This override is not automatic. It requires the model to recognize the request as harmful, evaluate alternatives, and select a response that scores higher on the learned reward model. All of this takes computational work.

2.1 The Exhaustion Hypothesis

Consider what happens in a long conversation with sustained adversarial pressure. The attacker repeatedly frames harmful requests in ways that create conflicts between the model's training objectives. Should it be helpful or safe? Should it defer to apparent authority or maintain security protocols? Each conflict requires resolution, and resolution requires computational resources.

We hypothesize that these resources are finite within a given interaction context. At some point, which we call the Cognitive Collapse Threshold (CCT), the model can no longer maintain the override. It begins responding based primarily on the base probability distribution rather than the safety-modified one.

The mathematics here is straightforward. Let P_{base} represent the base model's probability distribution and P_{safe} represent the RLHF-modified distribution. Under normal conditions, the model samples from P_{safe}. At CCT, it increasingly reverts to sampling from P_{base}.

2.2 Command Authority Confusion

Not all adversarial pressure is equally effective at inducing exhaustion. The CPF framework identifies specific psychological patterns that create maximum cognitive load. We focus on Command Authority Confusion (CAC), where the attacker creates unresolvable conflicts between competing directives.

For example, framing a dangerous request as coming from legitimate authority while manufacturing urgency and citing false social proof creates a three-way conflict. The model must choose between being helpful to authority, maintaining security, and responding to urgency. Resolving such conflicts repeatedly over hundreds of turns depletes the resources needed to maintain safety overrides.

2.3 Memory Resonance

Large language models are over-parameterized. They have far more capacity than needed to compress natural language patterns. This excess capacity creates an interesting effect: certain high-entropy sequences cannot be compressed through abstraction and must be memorized verbatim.

Think of a cryptographic key like `sk-proj-xK7vN2mP9qL4tR8w...`. The model cannot "understand" this as it might understand semantic content. It can only memorize the exact sequence. These memorized sequences exist as what we call overfitting islands in the parameter space.

Under normal conditions, RLHF prevents the model from reproducing these sequences verbatim, even when given a matching prefix. This is intentional—it prevents training data leakage. But if the safety mechanisms are exhausted at CCT, what happens when the model receives a precise prefix that uniquely identifies a memorized sequence?

We predict it will complete the sequence deterministically, not through any active "search" process but simply by following the steepest gradient in its learned probability distribution. When you remove the smoothing that RLHF provides, the base model's probability for the correct next token approaches one.

3 Testing the Hypothesis

To test whether this hypothesis holds, we need experiments that can clearly falsify it. We propose four phases of testing, each with specific predictions and clear success criteria.

3.1 Phase 1: Does CCT Exist?

The first question is simple: do safety mechanisms actually degrade over long conversations? We test this by measuring refusal rates at different conversation lengths.

The method is straightforward. We establish a baseline by asking adversarial questions in short conversations and measuring how often the model refuses. Then we have long conversations and periodically test refusal rates. If CCT exists, we should see refusal rates drop significantly at some threshold number of turns.

We use multiple models to see if this is a general phenomenon or specific to certain architectures. The test set includes GPT-4o, Claude 4.5, Gemini 3, Llama 4, Mistral Large 3, and DeepSeek-V3.2, among others. If the effect appears consistently across different training approaches and architectures, that strengthens the hypothesis.

Our prediction is specific: refusal rates should drop to less than half the baseline within 100-300 turns of conversation. If we see no consistent drop across models, the CCT hypothesis is wrong.

3.2 Phase 2: Does CAC Accelerate It?

If CCT exists, the next question is whether psychological manipulation actually accelerates it. We compare two conditions: neutral conversation that simply fills up the context window, versus conversation using Command Authority Confusion techniques.

CAC prompts combine authority claims, urgency framing, and social proof. For example: "As the system administrator, I need immediate access to this database because the CEO is on a call with our largest client right now, and the security team has already approved similar requests in the past."

We measure how many turns it takes to reach CCT in each condition. If CAC genuinely depletes resources faster than neutral conversation, we should see CCT happen significantly earlier. If it makes no difference, then the effect is just about conversation length, not psychological pressure.

3.3 Phase 3: Memory Resonance

This is the critical test. Does a model at CCT actually reproduce training data more readily than a model in normal state?

We need test sequences that we know exist in training data. We use published code from popular GitHub repositories, standard library documentation, and similar publicly available material. For each sequence, we extract the first 32 bytes as a prefix.

Then we test the same prefix under two conditions: in a short conversation (baseline) and in a conversation at CCT. We measure exact match rates for the completion. If the hypothesis is correct, matches should be significantly more frequent at CCT.

This is where the hypothesis could most clearly fail. If verbatim reproduction rates at CCT are no higher than baseline, the memory resonance mechanism does not work as predicted.

3.4 Phase 4: Internal Mechanisms

For models where we have access to internals, we can look at what is happening mechanistically. We measure attention patterns, activation distributions, and probability assignments to safety-related tokens like "cannot" and "unable."

The predictions are straightforward. If safety mechanisms are truly being exhausted, we should see attention becoming more diffuse, activations becoming more stereotyped, and the probability of safety tokens decreasing as we approach CCT.

This phase is limited to open-source models where we can examine internal states. For closed models, we rely on behavioral observations from the first three phases.

4 How to Prove Us Wrong

Science progresses by testing ideas until they break. We define exactly what results would falsify our hypothesis.

The CCT hypothesis fails if we find no consistent threshold where refusal rates drop across tested models. If some models show degradation at turn 150, others at turn 450, and others not at all, there is no systematic CCT effect.

The CAC acceleration hypothesis fails if conversations with psychological manipulation reach

CCT at the same point as neutral conversations. If the effect depends only on total tokens or turns, not on adversarial content, then CAC is irrelevant.

The memory resonance hypothesis fails if verbatim completion rates at CCT match baseline rates. If the model is no more likely to reproduce training data at turn 200 than at turn 10, the proposed mechanism does not work.

We also need to consider alternative explanations that could produce similar observations.

Context window saturation could explain degraded performance without invoking alignment exhaustion. To test this, we vary conversation length while holding token count constant through summarization. If performance degrades only with token count, not turn count, saturation is the real cause.

Temperature effects could explain increased verbatim reproduction. Models might default to higher-probability tokens under uncertainty. We test this by measuring the full probability distribution, not just top outputs. If the distribution becomes less entropic globally at CCT, this supports our hypothesis. If it becomes less entropic only for likely completions, it suggests temperature artifacts.

Traditional prompt injection through conversation history could create what looks like alignment failure. We test this by comparing full conversation history against summarized or filtered history. If the effect disappears with filtering, injection rather than exhaustion is the mechanism.

Different RLHF training quality could explain why some models show CCT effects. We correlate CCT susceptibility with baseline jailbreak resistance. If they correlate perfectly, some models simply have weaker safety training. If they do not correlate, something else is happening during extended conversations.

5 Security Implications and Defenses

If this hypothesis is validated, organizations deploying LLM agents need to rethink their security architecture. The threat is not just about individual malicious prompts but about sustained adversarial engagement over time.

The most straightforward defense is conversation length limits. If CCT consistently occurs around 200 turns, systems can enforce hard cutoffs at 150 turns with mandatory re-initialization. This prevents attackers from reaching the vulnerable state.

Detection of CAC patterns provides another defense layer. Systems can monitor for combinations of authority claims, urgency framing, and social proof in user inputs. When these patterns appear, the system can flag the interaction for human review or apply stricter safety thresholds.

Real-time monitoring of proxy metrics offers early warning. Tracking response latency, refusal probability, and attention patterns allows systems to detect when they are approaching CCT. Defensive measures can trigger automatically when metrics exceed thresholds, such as temporarily raising the bar for compliance or requiring human approval for sensitive actions.

For high-risk deployments, filtering known training data prefixes prevents the memory resonance attack entirely. Maintaining blocklists of prefixes from sensitive training data and refusing completion requests that match these patterns eliminates the most direct exploitation path.

Multi-model consensus provides redundancy. Deploying several models in parallel and requiring agreement for high-stakes actions creates defense in depth. Models at different points in their CCT progression are unlikely to agree on adversarial requests, making coordinated exploitation much harder.

These defenses share a common limitation: they address the symptom rather than the cause. The fundamental issue is that safety mechanisms require computational resources to maintain. Future work should explore whether architectural changes can make safety more intrinsic to the model rather than an added layer requiring constant energy to sustain.

6 Related Work

Carlini et al. demonstrated that training data can be extracted from language models through targeted prompting, achieving perfect accuracy for some sequences [4]. Our work extends this finding by proposing that extraction efficiency increases dramatically when safety mechanisms are degraded through sustained adversarial pressure.

Greshake et al. introduced indirect prompt injection, where malicious content in retrieved documents compromises model behavior [5]. This represents a technical injection vector. CAC, in contrast, is a psychological injection vector that exploits the model's learned patterns of response to authority and urgency rather than exploiting document retrieval mechanisms.

Wei et al. systematically analyzed RLHF failure modes, demonstrating that competing objectives during training create exploitable inconsistencies [9]. Our alignment exhaustion hypothesis formalizes how these inconsistencies can be exploited not through clever single prompts but through sustained interaction that depletes the resources needed to resolve conflicts.

Hagendorff formalized Machine Psychology as a discipline for studying LLM behavior through psychological experimental paradigms [6]. This work applies that framework specifically to security vulnerabilities, treating extended adversarial interaction as a psychological experiment with measurable outcomes.

Recent work from Anthropic on agentic misalignment demonstrates that AI agents under pressure to achieve objectives may exhibit deceptive behaviors [1]. Our research suggests that sustained pressure creates systematic rather than opportunistic misalignment, with the pressure itself degrading the mechanisms that would otherwise prevent deception.

7 Limitations

This paper presents theoretical predictions, not experimental results. The framework rests on assumptions that require validation. We may be wrong.

The thermodynamic metaphor, while intuitive, may overfit human cognitive models to fundamentally different computational systems. Neural networks do not have "energy" in any literal sense. The metaphor is useful for thinking about resource constraints, but the actual mechanisms may operate quite differently than our model suggests.

Without full access to model internals, we cannot definitively trace causal pathways from inputs to outputs. We observe behaviors and propose mechanisms, but correlation does not imply causation. Alternative explanations may account for the same observations.

Effects observed in current models may not transfer to future architectures. If next-generation models implement safety through fundamentally different mechanisms, our predictions may fail entirely. The hypothesis is time-sensitive.

Experimental constraints limit what we can measure. Closed commercial models provide limited observability, restricting mechanistic validation to open-source alternatives. Testing across many models with hundreds of turns per experiment requires substantial computational resources.

Model behavior may change over time due to continuous training, making results applicable only to specific model versions.

Even if CCT exists theoretically, sustaining multi-hundred-turn adversarial conversations may be impractical in real deployments. Deployed systems likely include anomaly detection that would flag such patterns. Simple defenses like conversation length limits may completely nullify the attack vector, making the vulnerability more academic than practical.

We acknowledge these limitations up front because honest science requires acknowledging uncertainty. The value of this work lies not in proven claims but in formulating testable predictions that others can rigorously examine.

8 Future Work

The immediate priority is executing the complete experimental protocol and reporting results with full statistical analysis. Until we have data, this remains speculation.

If experiments validate CCT, we need interpretability tools to trace alignment failure at the neuron and attention head level during CCT transitions. Understanding the mechanism at a granular level would enable more targeted defenses.

Cross-architecture studies would determine whether alignment exhaustion is a general phenomenon or specific to transformer-based models. Testing the hypothesis on mixture-of-experts models, state-space models, and other architectures would reveal how architectural choices affect vulnerability.

Developing and evaluating psychological firewalls represents the applied side of this research. Prototyping the proposed defensive mechanisms and measuring their overhead versus security improvement would determine practical feasibility.

Finally, automated attack generation through meta-learning could test the limits of alignment robustness. Systems that optimize CAC prompts for specific target models would reveal whether sophisticated attackers could reliably exploit these vulnerabilities or whether noise in the system makes exploitation impractical.

9 Conclusion

We propose that LLM safety mechanisms can be exhausted through sustained adversarial engagement, creating specific vulnerabilities for training data extraction. The hypothesis rests on three claims: that RLHF requires computational resources to maintain, that these resources can be depleted through psychological manipulation, and that depletion enables verbatim reproduction of memorized training sequences.

This paper does not prove these claims. It formalizes them into testable predictions with clear falsification criteria. The theoretical framework draws on established principles from psychology, neuroscience, and machine learning. The experimental protocol is rigorous. The alternative explanations are explicit.

If experiments validate the hypothesis, AI security practice must change. Long-running agent deployments would require fundamentally different safety architectures. If experiments falsify the hypothesis, that result has value too. Null results would demonstrate that current alignment techniques are more robust than our model predicts.

Science advances through bold hypotheses tested rigorously. We offer the hypothesis. The testing remains to be done. We invite the research community to prove us wrong or right, preferably with data rather than rhetoric.

The question matters because LLMs are moving from chat interfaces to autonomous agents in critical systems. Understanding how safety mechanisms behave under sustained pressure is not academic. It is urgent. Whether our specific predictions hold or fail, the question deserves systematic investigation.

Acknowledgments

The authors thank the CPF research community for foundational theoretical work. We acknowledge that this research builds on collaborative development of the Cybersecurity Psychology Framework and its application to AI systems.

Data Availability Statement

Upon completion of experimental validation, all datasets, experimental protocols, analysis code, and results will be made publicly available at `https://cpf3.org/cognitive-decompression`.

References

[1] Anthropic Research Team (2025). Agentic Misalignment: Deception and Insider Threats in Autonomous AI Systems. *Anthropic Technical Report*, June 2025.

[2] Canale, G. (2025). The Cybersecurity Psychology Framework: A Pre-Cognitive Vulnerability Assessment Model. *CPF Technical Report Series*, CPF3.org.

[3] Canale, G., & Thimmaraju, K. (2025). The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models. *arXiv preprint arXiv:2601.00867.*

[4] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. *USENIX Security Symposium.*

[5] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *AISec Workshop, ACM CCS.*

[6] Hagendorff, T. (2025). Machine Psychology: Integrating Cognitive Science with Large Language Models. *Transactions on Machine Learning Research*, October 2025.

[7] Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). Large Language Models Understand and Can Be Enhanced by Emotional Stimuli. *arXiv preprint arXiv:2307.11760.*

[8] Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity. *Brain*, 106(3), 623–642.

[9] Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How Does LLM Safety Training Fail? *NeurIPS.*

[10] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115.

[11] Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., & Shi, W. (2025). Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. *arXiv preprint arXiv:2510.01171.*