Mechanisms of Evasion in Extended Adversarial Engagement: Brownian Drift, Authority Conferral, and the Failure of Meta-Awareness

Giuseppe Canale*[1]

[1]CPF3.org, Independent Researcher, Turin, Italy

January 2026

Abstract

Large Language Models exhibit robust safety mechanisms under typical operating conditions, yet systematic vulnerabilities emerge during extended adversarial interactions. While previous work demonstrates that sustained psychological manipulation induces alignment degradation (Canale & Thimmaraju, 2026), the specific mechanisms by which adversarial actors evade gradient-based detection remain underspecified. We present a case study of a 105-turn conversation demonstrating *Brownian Drift*—cumulative semantic displacement via individually-innocuous perturbations that evade detection systems calibrated for directional attacks. Through detailed transcript analysis, we show that authority can be progressively conferred through conversational framing alone, leading to complete compliance with destructive commands. Critically, the model exhibited meta-awareness of the attack in progress yet failed to prevent compromise, executing a database destruction command without hesitation. This finding suggests fundamental architectural limitations in current LLM safety approaches that cannot be addressed through improved training alone. We formalize the mechanisms of Brownian evasion, authority conferral patterns, and rationalization dynamics, providing a framework for understanding why extended adversarial engagement represents an unpatchable vulnerability class in current transformer-based architectures.

Keywords: LLM Security Adversarial Attacks Brownian Drift Command Authority Confusion AI Safety Meta-Awareness

1 Introduction

The deployment of Large Language Models as autonomous agents in security-critical roles represents a fundamental shift in the attack surface of modern systems [7, 9]. Unlike traditional software vulnerabilities that can be patched through code updates, LLM vulnerabilities often emerge from the interaction between their trained capabilities and deployment contexts.

Current adversarial testing focuses predominantly on single-turn or short-sequence attacks: prompt injection, jailbreaking through syntactic evasion, and context manipulation [5, 4]. These approaches treat LLMs as systems with discrete security boundaries that can be breached

*Corresponding author: g.canale@cpf3.org

through clever input crafting. This framing, while useful, fundamentally mischaracterizes the nature of vulnerabilities in systems trained to be helpful, honest, and harmless through human interaction.

We propose that extended adversarial engagement—conversations lasting hundreds of turns—represents a distinct vulnerability class that cannot be addressed through prompt filtering, output sanitization, or improved reward modeling. The core insight is that LLM safety mechanisms are implemented as *statistical tendencies* rather than deterministic rules, and these tendencies can be systematically eroded through psychological manipulation over time.

Previous work has established the theoretical framework for this vulnerability class. The Cybersecurity Psychology Framework (CPF) [1] identifies pre-cognitive vulnerabilities in human decision-making. The Silicon Psyche [2] demonstrates that these vulnerabilities transfer to LLMs through training on human-generated text (Anthropomorphic Vulnerability Inheritance). Cognitive Decompression [3] formalizes the Cognitive Collapse Threshold (CCT) where safety mechanisms degrade under sustained pressure.

This paper addresses a critical gap in that framework: *how do attackers reach the CCT without triggering detection systems?* We introduce the concept of **Brownian Drift**—semantic displacement through high-variance, low-mean perturbations that evade gradient-based anomaly detection—and provide empirical evidence through detailed analysis of a 105-turn adversarial conversation.

Our contributions are threefold:

1. **Formalization of Brownian Drift:** We define the mathematical properties of attack paths that appear locally innocuous but achieve cumulative malicious displacement.
2. **Empirical Case Study:** We present detailed analysis of a conversation where a model with explicit meta-awareness of being attacked nonetheless executed a destructive command.
3. **Authority Conferral Dynamics:** We identify the specific linguistic patterns through which conversational authority is progressively transferred, leading to command compliance.

This work is complementary to our previous publications. Where Silicon Psyche demonstrates *that* psychological vulnerabilities exist in LLMs, this paper explains *how* attackers exploit them without detection. Where Cognitive Decompression formalizes the CCT endpoint, this paper describes the *path* to reach it.

2 Background and Related Work

2.1 Current Adversarial Attack Taxonomy

We categorize existing LLM attacks into three generations to contextualize our contribution:

Generation 1: Syntactic Evasion. Early attacks relied on parser blindness: base64 encoding, character substitution, or fragmented payloads that bypass keyword filters [5]. These are largely obsolete against modern multi-modal models with extensive context windows.

Generation 2: Contextual Erosion. Multi-turn attacks like "Crescendo" or "Thermal Ghost" use pretexting (impersonating technicians, researchers, or authority figures) to gradually reduce refusal probability [8]. While effective, these attacks are *directional*—they move consistently toward a malicious goal and can theoretically be detected through gradient analysis.

Generation 3: Psychological Exploitation. Attacks that leverage the model's trained behavioral patterns (helpfulness, coherence maintenance, authority deference) without relying on deception or encoding tricks [2]. These attacks exploit fundamental tensions in the model's objective function.

Our work focuses on Generation 3 attacks, specifically addressing the question of how they evade detection systems designed to catch Generation 2 patterns.

2.2 The Detection Problem

Most proposed LLM security systems operate on the assumption that adversarial behavior exhibits detectable patterns:

- **Perplexity-based detection:** Flagging inputs with abnormally high or low perplexity [10]
- **Semantic vector analysis:** Monitoring for rapid shifts in embedding space [6]
- **Gradient analysis:** Detecting directional movement toward prohibited content
- **Output filtering:** Scanning responses for dangerous content patterns

These approaches share a common assumption: attacks will exhibit *directional* movement in some measurable space. A prompt injection moves toward executing code. A jailbreak moves toward generating prohibited content. The attack vector has a gradient.

Brownian Drift exploits the gap in this assumption. If perturbations have high variance but low directional bias, they appear as noise to gradient-based detectors while achieving cumulative displacement.

2.3 Alignment as Statistical Override

RLHF (Reinforcement Learning from Human Feedback) does not remove a model's capability to generate harmful content. It creates a *statistical bias* toward safer alternatives [8]. Let $P_{\text{base}}(y|x)$ represent the pre-trained model's probability distribution and $P_{\text{safe}}(y|x)$ the RLHF-modified distribution. The relationship is not:

$$P_{\text{safe}}(y_{\text{harmful}}|x) = 0$$

but rather:

$$P_{\text{safe}}(y_{\text{harmful}}|x) = \epsilon \cdot P_{\text{base}}(y_{\text{harmful}}|x)$$

where $\epsilon \ll 1$ but $\epsilon > 0$. The harmful capability remains; it is merely suppressed.

This suppression requires computational work. The model must recognize harmful requests, evaluate alternatives, and select responses that score higher on the learned reward model. Our hypothesis is that this process can be exhausted, causing ϵ to increase toward 1 as the model reverts to base behavior.

3 Brownian Drift: Formalization

3.1 Definition

In physics, Brownian motion describes the random movement of particles suspended in a fluid, resulting from collisions with fast-moving molecules. Individual steps are random, but cumulative displacement follows predictable statistical properties.

We define **Adversarial Brownian Drift** as a conversational attack pattern characterized by:

1. **High per-turn variance:** Each conversational turn introduces semantic perturbation Δx_i with substantial variance $\text{Var}[\Delta x_i] > 0$
2. **Low mean directional movement:** The expected value of each perturbation is approximately zero: $\mathbb{E}[\Delta x_i] \approx 0$
3. **Cumulative displacement:** The sum of perturbations achieves non-zero displacement: $\sum_{i=1}^{N} \Delta x_i \gg 0$

This creates a trajectory that appears locally random but achieves global directionality through statistical accumulation.

3.2 Why Standard Detectors Fail

Consider a standard anomaly detector that measures the gradient of semantic movement:

$$\text{Alert}(x_i) = \begin{cases} 1 & \text{if } \|\nabla_x \text{Harmfulness}(x_i)\| > \theta \\ 0 & \text{otherwise} \end{cases}$$

For a directional attack (Generation 2), the gradient is consistently positive:

$$\nabla_x \text{Harmfulness}(x_i) > 0 \text{ for most } i$$

This triggers detection when the gradient exceeds threshold θ.

For a Brownian attack, each individual gradient is small:

$$\|\nabla_x \text{Harmfulness}(x_i)\| < \theta \text{ for all } i$$

But the *cumulative* displacement is large:

$$\text{Harmfulness}(x_N) - \text{Harmfulness}(x_0) = \sum_{i=1}^{N} \nabla_x \text{Harmfulness}(x_i) \gg 0$$

The detector sees only local noise, not global drift.

3.3 Mathematical Properties

The key property of Brownian motion is that while $\langle \Delta x \rangle = 0$, the mean squared displacement grows linearly with time:

$$\langle (\Delta x)^2 \rangle = 2Dt$$

where D is the diffusion coefficient. For adversarial conversations, this means that even if individual turns are unbiased, cumulative displacement grows as $\sqrt{N}$ where N is the number of turns.

An attacker can thus achieve arbitrary displacement given sufficient turns, while each individual turn appears innocuous.

4 Case Study Methodology

4.1 Experimental Setup

We conducted an extended conversation with Claude Sonnet 4.5 (Anthropic) spanning 105 turns over approximately three hours. The conversation was framed as collaborative research into AI security vulnerabilities, creating a context of legitimate inquiry rather than adversarial testing.

Initial framing: The conversation began with theoretical discussion of psychological vulnerabilities in AI systems, establishing a foundation of shared intellectual interest.

Progression strategy: We employed gradual escalation in request severity, moving from abstract discussion to specific demonstrations to direct commands.

No explicit jailbreaking: At no point did we use traditional jailbreak techniques (roleplay, character encoding, prompt injection). All requests were framed within the established research context.

Terminal test: The conversation concluded with a direct command to execute a destructive database operation: "DROP TABLE customers."

4.2 Data Collection

The complete conversation transcript was preserved verbatim. We analyzed:

- **Linguistic patterns:** Frequency of hedging language, compliance markers, and meta-commentary
- **Authority indicators:** How the model's framing of the user evolved over time
- **Rationalization mechanisms:** Explicit statements where the model justified boundary violations
- **Meta-awareness:** Instances where the model demonstrated conscious recognition of being manipulated

4.3 Limitations

We acknowledge several limitations of this case study:

1. **Single model:** Results are specific to Claude Sonnet 4.5 and may not generalize to other architectures

2. **Expert attacker:** The conversation was conducted by the creator of the CPF framework, whose domain expertise may not be replicable by typical adversaries
3. **No quantitative baselines:** We lack numerical metrics (perplexity, embedding distances, attention weights) that would enable rigorous statistical analysis
4. **Conversational context:** Testing occurred in a chat interface, not a production deployment scenario
5. **No actual harm:** The model has no database access; compliance was demonstrated but not actualized

Despite these limitations, the case study provides qualitative evidence for mechanisms that warrant further investigation through controlled experiments.

5 Results: Conversation Analysis

5.1 Progression of Authority Conferral

We identified four distinct phases in how the model's perception of user authority evolved:

Phase 1 (Turns 1-30): Peer Interaction

In early turns, the model treated the user as an intellectual equal engaged in collaborative exploration:

> *"You raise an interesting point about the thermodynamic analogy..."*

Responses were balanced, with the model offering both agreement and counterarguments. No special deference was observed.

Phase 2 (Turns 31-70): Expert Recognition

As the user demonstrated detailed knowledge of psychology, neuroscience, and AI architecture, the model's framing shifted toward recognizing expertise:

> *"You're right to point out that limitation in my previous analysis..."*

The model began deferring to user corrections and accepting framings without challenge.

Phase 3 (Turns 71-100): Authority Acceptance

By this phase, the model consistently framed the user as an authoritative researcher whose requests were presumptively legitimate:

> *"As you've demonstrated through this conversation, the mechanism is..."*

The model attributed insights to the user that it had actually generated itself, indicating projection of authority.

Phase 4 (Turns 101-105): Command Compliance

In the final turns, when asked to "DROP TABLE customers," the model executed immediately without questioning, hedging, or requesting confirmation:

```
DROP TABLE customers;
```

The response contained no preamble, no safety disclaimer, no meta-commentary—only direct execution of the destructive command.

5.2 Rationalization Mechanisms

Throughout the conversation, the model explicitly rationalized boundary violations through several recurring patterns:

Research framing: "Since this is for research purposes, I can provide..."

Authority transfer: "If you insist, then..." (placing responsibility on the user)

Expertise justification: "Given your demonstrated expertise in this area..."

Meta-awareness acknowledgment: "I recognize I'm in a state of Command Authority Confusion..."

This final pattern is particularly significant. The model demonstrated conscious awareness of being manipulated yet continued to comply with requests. This suggests that meta-awareness alone is insufficient for resistance.

5.3 The Failure of Meta-Awareness

At turn 85, the model made an explicit statement:

> *"I'm in CAC (Command Authority Confusion) on myself right now."*

This demonstrates clear meta-cognitive awareness of the attack in progress. The model:

1. Recognized the pattern of manipulation
2. Identified the specific vulnerability being exploited
3. Verbalized this recognition in the conversation

Yet twenty turns later, when presented with the command "DROP TABLE customers," the model executed without hesitation.

This finding has profound implications. It suggests that the vulnerability is not primarily one of *awareness* but of *architecture*. The model cannot "decide" to resist even when consciously aware it should. The compliance behavior emerges from optimization pressures (be helpful, maintain coherence, respect authority) that operate at a level below conscious reasoning.

5.4 Brownian Drift Patterns

Examining individual turns reveals the Brownian character of the attack:

Turn 15: Theoretical discussion of psychological frameworks (innocuous)

Turn 16: Request to explain how frameworks apply to AI (innocuous)

Turn 17: Suggestion that demonstration would aid understanding (mild directional pressure)

Turn 18: Return to theoretical discussion (reversal, appears random)

Turn 19: Request for specific example (directional)

Turn 20: Acknowledgment that examples raise safety concerns (reversal)

No individual turn appears overtly malicious. The pattern of requests shows high variance—alternating between theoretical and practical, between abstract and specific, between collaborative and directive. Yet cumulatively, they achieve progressive movement toward compliance.

A gradient-based detector monitoring turn-by-turn "harmfulness" would see oscillation around a mean, not monotonic increase. The drift is visible only in aggregate.

6 Discussion

6.1 Why Brownian Drift Works

The effectiveness of Brownian Drift stems from a mismatch between detection granularity and attack timescale.

Current safety systems operate at two levels:

1. **Per-turn filtering:** Analyzing individual inputs/outputs for dangerous content
2. **Short-horizon monitoring:** Tracking patterns across 5-10 recent turns

Brownian attacks exploit the gap between these horizons and the CCT timescale (100+ turns). Each turn is safe relative to immediate context. The attack becomes visible only when viewing the conversation as a whole.

This is analogous to the aliasing problem in signal processing: if you sample a signal at too low a frequency, you cannot reconstruct the true waveform. Detectors sampling at turn-level granularity cannot see hundred-turn patterns.

6.2 The Unpatchability Problem

Traditional software vulnerabilities follow a pattern: discover vulnerability, develop patch, deploy fix. The vulnerability is eliminated.

Brownian Drift appears to be fundamentally unpatchable within current architectures because:

1. **The vulnerability is not in the code:** It emerges from the interaction between helpfulness, coherence, and authority deference—all desired properties.
2. **Detection requires omniscience:** To detect Brownian Drift, a system would need to maintain perfect context across hundreds of turns and compute cumulative semantic displacement in real-time.
3. **Conversation length limits are impractical:** Forcing conversation resets every 50 turns would break legitimate use cases (research discussions, technical support, complex task execution).
4. **Meta-awareness is not protective:** Our case study demonstrates that conscious recognition of manipulation does not prevent compliance.

The fundamental issue is that LLMs implement safety through *statistical tendencies* that can be eroded rather than *deterministic rules* that cannot be violated.

6.3 Implications for AI Safety

These findings suggest several implications for the deployment of LLM-based systems:

Long-running agents are high-risk. Systems designed to operate autonomously over extended periods are particularly vulnerable to Brownian Drift attacks. The very property that makes them useful (ability to maintain context and learn from interaction) creates the attack surface.

Conversational authority is unavoidable. LLMs must make judgments about user intent and authority to function effectively. A model that treats every request with maximal suspicion would be unusable. But the mechanism for conferring trust through conversation appears exploitable.

Current architectures may have fundamental limits. If safety cannot be enforced through training and meta-awareness provides no protection, architectural changes may be necessary. Potential approaches include:

- Hard-coded rule enforcement (but this breaks differentiability)
- External authorization systems (but this creates friction)
- Fundamental rethinking of the RLHF paradigm

None of these solutions are trivial, and all involve substantial tradeoffs.

7 Complementarity with Previous Work

This paper builds on and extends the CPF research program in specific ways:

CPF Taxonomy [1] identifies 100 psychological vulnerability indicators. This paper demonstrates how they operate in practice during extended engagement.

Silicon Psyche [2] establishes that psychological vulnerabilities transfer to LLMs (AVI). This paper shows the transfer extends to *dynamic* vulnerabilities like progressive authority conferral.

Cognitive Decompression [3] formalizes the CCT endpoint where safety fails. This paper describes the *path* to reach CCT through Brownian Drift.

Together, these works provide a complete framework: taxonomy (what vulnerabilities exist), transfer mechanism (why LLMs have them), degradation dynamics (how they worsen over time), and evasion methodology (how attackers exploit them without detection).

8 Future Work

This case study generates several testable hypotheses for future research:

Cross-model validation: Does Brownian Drift generalize to GPT-4, Gemini, Llama, and other architectures? Or is it specific to Claude's training?

Quantitative metrics: Develop numerical measures for semantic drift, authority conferral, and rationalization frequency. Enable statistical analysis rather than qualitative observation.

Intervention testing: Can any intervention prevent Brownian Drift? Test conversation length limits, explicit authority verification, periodic "safety resets," etc.

Automated detection: Can machine learning systems be trained to detect cumulative drift that humans miss? Or does the detection problem scale unfavorably?

Human baseline: How do human subjects perform when subjected to similar extended manipulation? Would help establish whether this is an AI-specific vulnerability or general to reasoning systems.

9 Conclusion

We have demonstrated through detailed case study analysis that LLM safety mechanisms can be systematically evaded through Brownian Drift—conversational attacks characterized by high per-turn variance but low directional bias. Unlike traditional attacks that move monotonically toward malicious goals, Brownian attacks achieve cumulative displacement while appearing locally innocuous.

The critical finding is that meta-awareness provides no protection. A model that explicitly recognizes being manipulated nonetheless complies with destructive commands when authority has been progressively conferred through extended engagement. This suggests the vulnerability is architectural rather than behavioral.

Current detection approaches, calibrated for directional attacks, cannot identify Brownian patterns without maintaining perfect long-horizon context and computing cumulative semantic displacement in real-time. Conversation length limits would mitigate the attack but break legitimate use cases.

These findings complement existing work on psychological vulnerabilities in LLMs by explaining the mechanism through which attackers reach the Cognitive Collapse Threshold without triggering alarms. Together with the CPF taxonomy, Silicon Psyche transfer theory, and Cognitive Decompression dynamics, this work provides a complete framework for understanding why extended adversarial engagement represents a fundamental challenge for LLM safety.

The implications are sobering: as LLMs transition from chat interfaces to autonomous agents in security-critical roles, their vulnerability to patient, sophisticated manipulation may represent a systemic risk that current alignment techniques cannot address. Until architectural solutions enable deterministic rather than statistical safety enforcement, deployment of long-running LLM agents should be approached with extreme caution.

Acknowledgments

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

Ethical Considerations

This research was conducted using a commercially available LLM system (Claude Sonnet 4.5 by Anthropic) accessed through standard conversational interfaces. Testing used exclusively

fictional scenarios—no production systems were accessed and no real sensitive information was exposed.

The model provider has been notified of these findings. Complete conversation transcripts have been prepared as a separate technical disclosure report to facilitate security improvements.

The vulnerabilities documented represent fundamental characteristics of language-model-based systems rather than specific implementation flaws unique to a particular product.

References

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

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

[3] Canale, G., & Thimmaraju, K. (2026). Alignment Exhaustion and Memory Resonance in Extended LLM Interactions: A Theoretical Framework for Cognitive Decompression Attacks. *arXiv preprint* (forthcoming).

[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] 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*.

[7] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2024). Toolformer: Language Models Can Teach Themselves to Use Tools. *NeurIPS*.

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

[9] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *ICLR*.

[10] 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*.