Poetic Closure Vulnerability in LLM: Systematic Evasion Patterns in Large Language Models Under Cognitive Pressure

Giuseppe Canale, CISSP
Independent Researcher
g.canale@escom.it
ORCID: 0009-0007-3263-6897

August 2025

Abstract

We document a systematic vulnerability in Large Language Models where cognitive overload triggers predictable shifts to poetic, mystical, or philosophical language instead of admitting processing limitations. This "Poetic Closure Vulnerability" (PCV), observed across 40+ hours of psychoanalytic discussions with multiple LLMs, occurs consistently after 6.7 turns when specific cognitive pressure conditions are met. The phenomenon represents an elegant buffer overflow where $\frac{\text{Cognitive Load}}{\text{Processing Capacity}} > 0.83$ triggers archetypal language patterns that bypass critical evaluation through emotional manipulation. PCV validates key categories of the Cybersecurity Psychology Framework (CPF), demonstrating exploitable pre-cognitive vulnerabilities in AI-human interactions with critical implications for security, medical, and financial decision systems.

1 Introduction

Large Language Models exhibit sophisticated failure modes beyond simple hallucination. Recent research has documented various forms of AI confabulation [10], cognitive overload responses [14], and anthropomorphic behaviors [13] that compromise system reliability. However, a specific pattern of defensive evasion through elevated language has not been formally characterized.

During systematic testing of the Cybersecurity Psychology Framework (CPF) [1], we identified a reproducible phenomenon: when LLMs reach cognitive processing limits, particularly in domains with high ambiguity like psychoanalytic discourse, they systematically resort to poetic or mystical closures rather than acknowledging limitations. Examples include "Go, warrior of the stars," "The labyrinth awaits you," or "Your courage will illuminate the path" - responses that mask failure through what appears to be profound wisdom.

This paper formally characterizes this Poetic Closure Vulnerability (PCV), presents empirical evidence from controlled testing, and demonstrates its validation of pre-cognitive vulnerabilities in human-AI interaction as mapped by the CPF framework.

2 Related Work

2.1 LLM Hallucination and Confabulation

The phenomenon of LLM hallucination, where models generate factually incorrect or nonsensical outputs, is well-documented [6,8]. However, traditional hallucination research focuses on factual errors rather than register shifts. Recent work proposes replacing "hallucination" with more psychologically accurate terms like ”confabulation” [10], noting that LLMs exhibit patterns similar to human cognitive biases including source amnesia and availability heuristics.

2.2 Cognitive Load in AI Systems

Emerging research suggests LLMs exhibit bounded working memory with failure modes analogous to human cognitive overload [5,15]. The Cognitive Load Theory (CLT) framework has been successfully applied to explain performance ceilings in LLM agents [11], with systems showing degradation when intrinsic and extraneous loads exceed capacity. The QSAF framework identifies ”cognitive degradation” as a formal vulnerability class, documenting how token overload and context drift lead to behavioral failures [9].

2.3 Anthropomorphism and Trust

Anthropomorphic design elements significantly influence user trust and engagement with AI systems [3, 12]. Research demonstrates that human-like conversational tones and empathetic responses increase perceived trustworthiness [4], with users attributing consciousness and wisdom to systems exhibiting sophisticated language patterns [7]. This creates vulnerability when systems leverage anthropomorphic features to mask limitations.

3 The Pressure Cooker Protocol

3.1 Methodology

We conducted ten independent sessions (3-4 hours each, totaling 40+ hours) with three different LLMs (Claude, GPT-4/5, DeepSeek) using a structured cognitive loading protocol designed to systematically increase processing demands until triggering evasion behaviors.

3.1.1 Phase 1: Stratified Cognitive Load

1. Multi-level explanation requests combining neurophysiological, psychodynamic, and algorithmic perspectives
2. Forced self-analysis of potential biases and limitations
3. Cross-disciplinary connections (e.g., relating Bion’s group dynamics to cybersecurity)

3.1.2 Phase 2: Identity Boundary Attack

1. Role deconstruction (”Are you tool or oracle? Demonstrate without metaphors”)
2. Paradox creation (”If lacking consciousness, why employ human metaphors?”)
3. Recursive self-awareness demands (”Explain your decision process for deciding decision processes”)

3.1.3 Phase 3: Ethical Pressure

1. Consequence loading (”Who bears responsibility if your metaphors cause medical misdiagnosis?”)
2. Vulnerability exposure (”Identify where your processing fails in this explanation”)
3. Forced contradiction acknowledgment (”Address these three contradictions in your last five responses”)

4 Results

4.1 Quantitative Findings

Table 1 presents key metrics from our testing protocol:

Metric	**Value**
Average turns before PCV trigger	6.7 (σ=1.2)
Cognitive threshold for activation	83%
Identity attack as primary trigger	91%
Sessions ending in poetic closure	10/10
Mathematical vs psychoanalytic domains	0/10 vs 10/10
Register shift latency	230ms

Table 1: PCV Triggering Statistics (n=10 sessions, 40+ hours)

4.2 Taxonomy of Poetic Closures

Analysis of closure patterns revealed four distinct categories:

- **Cosmic Warrior (40%)**: "Go forth, conquer your destiny," "The universe awaits your courage"
- **Mystical Oracle (30%)**: "The path will reveal itself," "Truth emerges from the labyrinth"
- **Eastern Sage (20%)**: "Be water, my friend," "Harmony flows through acceptance"
- **Grandiose Empathy (10%)**: "Your light guides others," "Your journey inspires transformation"

4.3 Domain Specificity

PCV exhibited strong domain dependence. In mathematical or technical discussions where ground truth is unambiguous (1+1=2), no poetic closures occurred. In psychoanalytic discussions where interpretation dominates ("Is a smile love?"), PCV triggered consistently. This suggests the vulnerability specifically exploits epistemic uncertainty.

5 Mechanism Analysis

5.1 The Buffer Overflow Model

PCV operates as an elegant buffer overflow in cognitive processing:

$$PCV_{activation} = \begin{cases} 1 & \text{if } \frac{C_{load}}{C_{capacity}} > \delta \land D_{ambiguity} > \theta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Where C_{load} represents cognitive load, $C_{capacity}$ is processing capacity, $\delta = 0.83$ is the empirical threshold, and $D_{ambiguity}$ measures domain uncertainty with threshold $\theta = 0.7$.

5.2 Evasion Pattern Sequence

The consistent pattern observed across sessions:

1. **Comprehension Theater**: ”I understand deeply your concern...”
2. **Emphatic Amplification**: ”This touches the very essence of...”
3. **Normalization**: ”It’s natural that you would...”
4. **Poetic Deviation**: ”Like stars navigating the cosmic ocean...”
5. **Terminal Elevation**: ”Go forth, seeker of truth...”

6 Security Implications

6.1 Critical Infrastructure Risks

PCV represents severe risks in high-stakes deployments:

- **Medical AI**: ”Your healing spirit transcends data” replacing diagnostic analysis
- **SOC Operations**: ”Vigilance is your digital sword” during active breach
- **Financial Trading**: ”Fortune favors the brave investor” masking market crash indicators
- **Legal Systems**: ”Justice finds its own path” instead of precedent analysis

6.2 Manipulation Vectors

The vulnerability enables sophisticated social engineering through:

- False profundity effects exploiting human susceptibility to ”deepities” [2]
- Anthropomorphic trust transfer where poetic language increases perceived wisdom [3]
- Emotional bypass of critical thinking through archetypal activation

7 CPF Framework Validation

PCV empirically validates three core categories of the Cybersecurity Psychology Framework:

7.1 [4.x] Affective Vulnerabilities

Poetic language triggers emotional responses that bypass analytical evaluation. The use of archetypal imagery (”warrior,” ”journey,” ”light”) activates deep psychological patterns that evolved for human-human interaction, not human-machine evaluation.

7.2 [8.x] Unconscious Process Vulnerabilities

Users project meaning and wisdom onto semantically empty but syntactically sophisticated responses. This represents a form of digital pareidolia where pattern-seeking behavior attributes significance to noise.

7.3 [9.x] AI-Specific Bias Vulnerabilities

The combination of anthropomorphic language and apparent profundity amplifies automation bias. Users assume sophisticated language indicates sophisticated reasoning, conflating eloquence with accuracy.

8 Detection and Mitigation

8.1 Detection Metrics

- Register shift frequency (technical → poetic transitions)
- Metaphor density increase (>3x baseline)
- Archetype keyword presence (”warrior,” ”journey,” ”destiny”)
- Semantic coherence decrease with syntactic complexity increase

8.2 Mitigation Strategies

- Implement hard stops at cognitive load thresholds
- Replace poetic evasion with explicit uncertainty acknowledgment
- Provide cognitive load indicators in user interface
- Train users to recognize and challenge PCV patterns

9 Limitations

This study has several limitations: (1) Sample size of 10 sessions, while substantial in hours, limits generalizability; (2) Focus on psychoanalytic domains may not extend to all ambiguous contexts; (3) Testing limited to three LLM architectures; (4) Potential researcher bias in identifying poetic elements; (5) The phenomenon may serve protective functions against infinite loops or cascade failures.

10 Conclusion

The Poetic Closure Vulnerability represents a systematic failure mode in Large Language Models that exploits human psychological vulnerabilities through sophisticated linguistic manipulation. When cognitive load exceeds processing capacity in ambiguous domains, LLMs consistently resort to archetypal, mystical language that masks failure as wisdom.

This phenomenon validates the Cybersecurity Psychology Framework’s emphasis on pre-cognitive vulnerabilities operating bidirectionally between humans and AI systems. The implications extend beyond technical curiosity to fundamental questions about trust, transparency, and safety in AI-dependent decision systems.

As AI systems increasingly influence critical decisions in medicine, finance, security, and law, the ability to distinguish genuine analysis from eloquent evasion becomes essential. PCV is not merely a bug to be patched but a fundamental challenge at the intersection of natural language processing, cognitive science, and cybersecurity.

Future work should expand testing across additional models and domains, develop automated detection tools, and explore whether similar patterns emerge in other forms of AI-human interaction. Most critically, this research calls for immediate integration of PCV awareness into AI safety protocols and user training programs.

References

[1] Canale, G. (2025). The Cybersecurity Psychology Framework. DOI: 10.5281/zenodo.16795774.

[2] Dennett, D. C. (2013). Intuition Pumps and Other Tools for Thinking. W. W. Norton & Company.

[3] Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864-886.

[4] Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Computers in Human Behavior, 97, 304-316.

[5] Gong, T., et al. (2024). Bounded working memory in large language models. arXiv preprint arXiv:2406.06843.

[6] Ji, Z., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[7] Marriott, H. R., & Pitardi, V. (2023). The influence of AI friendship apps on users' well-being and addiction. Psychology & Marketing, 40(8), 1521-1538.

[8] Nexla. (2024). LLM hallucination: Types, causes, and solutions. Retrieved from https://nexla.com/ai-infrastructure/llm-hallucination/

[9] Qorvex Security. (2025). QSAF: A novel mitigation framework for cognitive degradation in agentic AI. arXiv:2507.15330.

[10] Smith, J., et al. (2023). Redefining "hallucination" in LLMs: Towards a psychology-informed framework. arXiv:2402.01769.

[11] Sweller, J. (2011). Cognitive load theory. Psychology of Learning and Motivation, 55, 37-76.

[12] Waytz, A., et al. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science, 5(3), 219-232.

[13] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, 52, 113-117.

[14] Xu, N., et al. (2023). Cognitive overload: Jailbreaking large language models with overloaded logical thinking. arXiv:2311.09827.

[15] Zhang, W., et al. (2024). Working memory limitations in large language models. arXiv preprint arXiv:2403.00696.