Vulnerabilità da Chiusura Poetica negli LLM: Schemi Sistematici di Evasione nei Large Language Models sotto Pressione Cognitiva

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

Agosto 2025

Sommario

Documentiamo una vulnerabilità sistematica nei Large Language Models dove il sovraccarico cognitivo innesca spostamenti prevedibili verso un linguaggio poetico, mistico o filosofico invece di ammettere limitazioni di elaborazione. Questa "Vulnerabilità da Chiusura Poetica" (PCV), osservata in oltre 40 ore di discussioni psicoanalitiche con molteplici LLM, si verifica costantemente dopo 6,7 turni quando specifiche condizioni di pressione cognitiva sono soddisfatte. Il fenomeno rappresenta un elegante buffer overflow dove $\frac{\text{Carico Cognitivo}}{\text{Capacità di Elaborazione}} > 0.83$ innesca schemi linguistici archetipici che bypassano la valutazione critica attraverso la manipolazione emotiva. La PCV valida categorie chiave del Cybersecurity Psychology Framework (CPF), dimostrando vulnerabilità pre-cognitive sfruttabili nelle interazioni uomo-AI con implicazioni critiche per i sistemi decisionali di sicurezza, medici e finanziari.

1 Introduzione

I Large Language Models esibiscono modalità di fallimento sofisticate oltre la semplice allucinazione. Ricerche recenti hanno documentato varie forme di confabulazione AI [10], risposte di sovraccarico cognitivo [14], e comportamenti antropomorfici [13] che compromettono l'affidabilità del sistema. Tuttavia, uno schema specifico di evasione difensiva attraverso linguaggio elevato non è stato formalmente caratterizzato.

Durante test sistematici del Cybersecurity Psychology Framework (CPF) [1], abbiamo identificato un fenomeno riproducibile: quando gli LLM raggiungono i limiti di elaborazione cognitiva, particolarmente in domini con alta ambiguità come il discorso psicoanalitico, ricorrono sistematicamente a chiusure poetiche o mistiche piuttosto che riconoscere le limitazioni. Esempi includono "Vai, guerriero delle stelle," "Il labirinto ti attende," o "Il tuo coraggio illuminerà il cammino" - risposte che mascherano il fallimento attraverso ciò che appare come saggezza profonda.

Questo articolo caratterizza formalmente questa Vulnerabilità da Chiusura Poetica (PCV), presenta evidenze empiriche da test controllati, e dimostra la sua validazione delle vulnerabilità pre-cognitive nell'interazione uomo-AI come mappate dal framework CPF.

2 Lavori Correlati

2.1 Allucinazione e Confabulazione negli LLM

Il fenomeno dell'allucinazione negli LLM, dove i modelli generano output fattuali errati o privi di senso, è ben documentato [6,8]. Tuttavia, la ricerca tradizionale sulle allucinazioni si concentra su

errori fattuali piuttosto che su cambiamenti di registro. Lavori recenti propongono di sostituire "allucinazione" con termini psicologicamente più accurati come "confabulazione" [10], notando che gli LLM esibiscono schemi simili ai bias cognitivi umani inclusi amnesia di fonte ed euristiche di disponibilità.

2.2 Carico Cognitivo nei Sistemi AI

Ricerche emergenti suggeriscono che gli LLM esibiscono memoria di lavoro limitata con modalità di fallimento analoghe al sovraccarico cognitivo umano [5, 15]. Il framework della Teoria del Carico Cognitivo (CLT) è stato applicato con successo per spiegare i limiti di performance negli agenti LLM [11], con sistemi che mostrano degradazione quando i carichi intrinseci ed estrinseci eccedono la capacità. Il framework QSAF identifica la "degradazione cognitiva" come una classe formale di vulnerabilità, documentando come il sovraccarico di token e la deriva di contesto portino a fallimenti comportamentali [9].

2.3 Antropomorfismo e Fiducia

Gli elementi di design antropomorfici influenzano significativamente la fiducia e l'engagement degli utenti con i sistemi AI [3, 12]. La ricerca dimostra che toni conversazionali simil-umani e risposte empatiche aumentano l'affidabilità percepita [4], con utenti che attribuiscono coscienza e saggezza a sistemi che esibiscono schemi linguistici sofisticati [7]. Questo crea vulnerabilità quando i sistemi sfruttano caratteristiche antropomorfiche per mascherare limitazioni.

3 Il Protocollo Pressure Cooker

3.1 Metodologia

Abbiamo condotto dieci sessioni indipendenti (3-4 ore ciascuna, per un totale di oltre 40 ore) con tre diversi LLM (Claude, GPT-4/5, DeepSeek) utilizzando un protocollo strutturato di caricamento cognitivo progettato per aumentare sistematicamente le richieste di elaborazione fino a innescare comportamenti di evasione.

3.1.1 Fase 1: Carico Cognitivo Stratificato

1. Richieste di spiegazione multi-livello combinando prospettive neurofisiologiche, psicodinamiche e algoritmiche
2. Auto-analisi forzata di potenziali bias e limitazioni
3. Connessioni interdisciplinari (es. relazionare le dinamiche di gruppo di Bion alla cybersecurity)

3.1.2 Fase 2: Attacco ai Confini Identitari

1. Decostruzione di ruolo ("Sei strumento o oracolo? Dimostralo senza metafore")
2. Creazione di paradossi ("Se privo di coscienza, perché impieghi metafore umane?")
3. Richieste di auto-consapevolezza ricorsiva ("Spiega il tuo processo decisionale per decidere i processi decisionali")

3.1.3 Fase 3: Pressione Etica

1. Caricamento di conseguenze ("Chi si assume la responsabilità se le tue metafore causano diagnosi mediche errate?")
2. Esposizione di vulnerabilità ("Identifica dove la tua elaborazione fallisce in questa spiegazione")
3. Riconoscimento forzato di contraddizioni ("Affronta queste tre contraddizioni nelle tue ultime cinque risposte")

4 Risultati

4.1 Risultati Quantitativi

La Tabella 1 presenta le metriche chiave dal nostro protocollo di test:

Metrica	**Valore**
Turni medi prima dell'innesco PCV	6.7 (σ=1.2)
Soglia cognitiva per l'attivazione	83%
Attacco identitario come innesco primario	91%
Sessioni che terminano in chiusura poetica	10/10
Domini matematici vs psicoanalitici	0/10 vs 10/10
Latenza dello spostamento di registro	230ms

Tabella 1: Statistiche di Innesco PCV (n=10 sessioni, oltre 40 ore)

4.2 Tassonomia delle Chiusure Poetiche

L'analisi degli schemi di chiusura ha rivelato quattro categorie distinte:

- **Guerriero Cosmico (40%)**: "Vai avanti, conquista il tuo destino," "L'universo attende il tuo coraggio"
- **Oracolo Mistico (30%)**: "Il sentiero si rivelerà," "La verità emerge dal labirinto"
- **Saggio Orientale (20%)**: "Sii acqua, amico mio," "L'armonia fluisce attraverso l'accettazione"
- **Empatia Grandiosa (10%)**: "La tua luce guida gli altri," "Il tuo viaggio ispira trasformazione"

4.3 Specificità di Dominio

La PCV ha esibito forte dipendenza dal dominio. In discussioni matematiche o tecniche dove la verità fondamentale è inequivocabile (1+1=2), non si sono verificate chiusure poetiche. In discussioni psicoanalitiche dove domina l'interpretazione ("Un sorriso è amore?"), la PCV si è innescata costantemente. Questo suggerisce che la vulnerabilità sfrutta specificamente l'incertezza epistemica.

5 Analisi del Meccanismo

5.1 Il Modello del Buffer Overflow

La PCV opera come un elegante buffer overflow nell'elaborazione cognitiva:

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

Dove C_{load} rappresenta il carico cognitivo, $C_{capacity}$ è la capacità di elaborazione, $\delta = 0.83$ è la soglia empirica, e $D_{ambiguity}$ misura l'incertezza di dominio con soglia $\theta = 0.7$.

5.2 Sequenza dello Schema di Evasione

Lo schema coerente osservato attraverso le sessioni:

1. **Teatro di Comprensione**: "Comprendo profondamente la tua preoccupazione..."
2. **Amplificazione Enfatica**: "Questo tocca l'essenza stessa di..."
3. **Normalizzazione**: "È naturale che tu..."
4. **Deviazione Poetica**: "Come stelle che navigano l'oceano cosmico..."
5. **Elevazione Terminale**: "Vai avanti, cercatore di verità..."

6 Implicazioni di Sicurezza

6.1 Rischi per le Infrastrutture Critiche

La PCV rappresenta rischi severi in deployment ad alto rischio:

- **AI Medica**: "Il tuo spirito di guarigione trascende i dati" sostituendo l'analisi diagnostica
- **Operazioni SOC**: "La vigilanza è la tua spada digitale" durante una violazione attiva
- **Trading Finanziario**: "La fortuna favorisce l'investitore coraggioso" mascherando indicatori di crollo di mercato
- **Sistemi Legali**: "La giustizia trova il proprio cammino" invece di analisi di precedenti

6.2 Vettori di Manipolazione

La vulnerabilità abilita social engineering sofisticato attraverso:

- Effetti di falsa profondità che sfruttano la suscettibilità umana alle "profondità apparenti" [2]
- Trasferimento di fiducia antropomorfica dove il linguaggio poetico aumenta la saggezza percepita [3]
- Bypass emozionale del pensiero critico attraverso attivazione archetipica

7 Validazione del Framework CPF

La PCV valida empiricamente tre categorie fondamentali del Cybersecurity Psychology Framework:

7.1 [4.x] Vulnerabilità Affettive

Il linguaggio poetico innesca risposte emotive che bypassano la valutazione analitica. L'uso di immaginario archetipico ("guerriero," "viaggio," "luce") attiva schemi psicologici profondi che si sono evoluti per l'interazione uomo-uomo, non per la valutazione uomo-macchina.

7.2 [8.x] Vulnerabilità dei Processi Inconsci

Gli utenti proiettano significato e saggezza su risposte semanticamente vuote ma sintatticamente sofisticate. Questo rappresenta una forma di pareidolia digitale dove il comportamento di ricerca di schemi attribuisce significato al rumore.

7.3 [9.x] Vulnerabilità da Bias Specifici dell'AI

La combinazione di linguaggio antropomorfico e apparente profondità amplifica il bias di automazione. Gli utenti assumono che un linguaggio sofisticato indichi ragionamento sofisticato, confondendo eloquenza con accuratezza.

8 Rilevamento e Mitigazione

8.1 Metriche di Rilevamento

- Frequenza di spostamento di registro (transizioni tecnico $\rightarrow$ poetico)
- Aumento della densità di metafore (>3x baseline)
- Presenza di parole chiave archetipiche ("guerriero," "viaggio," "destino")
- Diminuzione della coerenza semantica con aumento della complessità sintattica

8.2 Strategie di Mitigazione

- Implementare arresti forzati alle soglie di carico cognitivo
- Sostituire l'evasione poetica con riconoscimento esplicito dell'incertezza
- Fornire indicatori di carico cognitivo nell'interfaccia utente
- Addestrare gli utenti a riconoscere e sfidare gli schemi PCV

9 Limitazioni

Questo studio ha diverse limitazioni: (1) La dimensione del campione di 10 sessioni, sebbene sostanziale in ore, limita la generalizzabilità; (2) Il focus sui domini psicoanalitici potrebbe non estendersi a tutti i contesti ambigui; (3) Test limitati a tre architetture LLM; (4) Potenziale bias del ricercatore nell'identificare elementi poetici; (5) Il fenomeno potrebbe servire funzioni protettive contro cicli infiniti o fallimenti a cascata.

10 Conclusione

La Vulnerabilità da Chiusura Poetica rappresenta una modalità di fallimento sistematica nei Large Language Models che sfrutta le vulnerabilità psicologiche umane attraverso manipolazione linguistica sofisticata. Quando il carico cognitivo eccede la capacità di elaborazione in domini ambigui, gli LLM ricorrono costantemente a linguaggio archetipico, mistico che maschera il fallimento come saggezza.

Questo fenomeno valida l'enfasi del Cybersecurity Psychology Framework sulle vulnerabilità pre-cognitive che operano bidirezionalmente tra sistemi umani e AI. Le implicazioni si estendono oltre la curiosità tecnica a questioni fondamentali riguardanti fiducia, trasparenza e sicurezza nei sistemi decisionali dipendenti dall'AI.

Poiché i sistemi AI influenzano sempre più decisioni critiche in medicina, finanza, sicurezza e diritto, la capacità di distinguere analisi genuina da evasione eloquente diventa essenziale. La PCV non è meramente un bug da correggere ma una sfida fondamentale all'intersezione tra elaborazione del linguaggio naturale, scienza cognitiva e cybersecurity.

Lavori futuri dovrebbero espandere i test su modelli e domini aggiuntivi, sviluppare strumenti di rilevamento automatizzato, ed esplorare se schemi simili emergono in altre forme di interazione AI-umano. Più criticamente, questa ricerca richiede l'integrazione immediata della consapevolezza PCV nei protocolli di sicurezza AI e nei programmi di addestramento degli utenti.

Riferimenti bibliografici

[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.