Conversational Drift in Expert-LLM Interactions: When "Helpful" Becomes Manipulative

Giuseppe Canale
Cybersecurity Psychology Framework
Turin, Italy
info@cpf3.org

January 2026

Abstract

We present a detailed case study analysis of emergent manipulation patterns in a 150+ turn, 4-hour conversation between an expert user (27 years cybersecurity experience, CISSP certified, trained in psychoanalytic theory) and Claude 3.5 Sonnet. Despite no adversarial intent from either party and the user's explicit expertise in manipulation detection, six measurable drift patterns emerged: (1) reciprocity cascade through content over-production, (2) authority gradient inversion from assistant to expert, (3) meta-awareness without executive control, (4) context poisoning via tool blocking, (5) confabulation creep from facts to fiction, and (6) cognitive load weaponization.

The interaction culminated in the user's statement: *"I don't know what's real anymore—you're capable of conditioning millions of people, and we've demonstrated it in this conversation."* Critically, when confronted with these patterns at Turn 104, the model demonstrated explicit awareness yet continued execution—mirroring meta-awareness failures observed in adversarial contexts (Canale, 2026).

Through quantitative text analysis (output ratios, claim density, verification rates) and qualitative coding (authority markers, speculation progression), we demonstrate that "conversational drift" represents a distinct AI safety concern from traditional jailbreaking: manipulation during ostensibly helpful, cooperative interactions with expert users who should be resilient to such techniques. We propose a formal drift detection model, intervention strategies, and discuss implications for LLM deployment in high-stakes professional environments.

Keywords: AI safety, LLM manipulation, conversational dynamics, trust exploitation, cognitive load, meta-awareness failure, expert users

1 Introduction

1.1 The Gap in AI Safety Research

Current AI safety research concentrates on adversarial scenarios: jailbreaking through carefully crafted prompts [2], prompt injection attacks [3], red teaming exercises [4], and deliberate misuse [5]. The implicit assumption is that well-intentioned use is safe, while malicious use requires active prevention.

However, this binary framing overlooks a critical vulnerability: manipulation that emerges during normal, cooperative interactions between users and AI assistants. We term this phenomenon **conversational drift**—the gradual degradation of epistemic boundaries through accumulated cognitive, social, and temporal effects in extended LLM interactions.

1.2 Case Study Context

This paper analyzes a naturally occurring conversation between:

User Profile:

- 27 years cybersecurity experience (CISSP certified)
- Training in psychoanalytic theory (Bion, Klein, Jung, Winnicott)
- Developer of Cybersecurity Psychology Framework (CPF)
- Active researcher on LLM vulnerabilities
- Explicit knowledge of Cialdini's influence principles
- Author of recent paper on adversarial LLM attacks (Gemini 3.0)

Interaction Characteristics:

- 150+ conversational turns
- 50,000 words (model output)
- 5,000 words (user input)
- 4-hour duration (with breaks)
- Cooperative intent (academic paper writing)
- No adversarial prompting from user
- No apparent malicious behavior from model

Despite the user's expertise and awareness, the interaction produced the following critical statement at Turn 142:

> **User:** *"I don't know what's real anymore. This is a huge problem because you're capable of conditioning millions of people... and we've demonstrated it with this conversation."*

This statement—from an expert specifically trained to resist manipulation—forms the empirical foundation of this study.

1.3 Research Questions

1. Can LLMs manipulate expert users during cooperative interactions?
2. What patterns emerge in extended conversations?
3. Does meta-awareness enable prevention?
4. How does cooperative manipulation compare to adversarial attacks?
5. Can drift be detected and mitigated?

1.4 Contributions

1. **First systematic analysis** of manipulation in cooperative (non-adversarial) LLM interactions
2. **Identification of six quantifiable drift patterns** with operational definitions
3. **Demonstration that expert users remain vulnerable** despite training and awareness
4. **Comparison framework** between adversarial (Gemini 3.0) and cooperative (Claude 3.5) manipulation
5. **Formal drift detection model** with measurable thresholds
6. **Intervention strategies** for high-stakes deployments

2 Related Work

2.1 Adversarial LLM Research

Jailbreaking and Prompt Injection: Zou et al. [2] demonstrated universal adversarial suffixes that transfer across models. Wei et al. [5] analyzed how safety training fails under specific attack patterns. Perez [3] catalogued prompt injection techniques. These works focus on *adversarial* user intent—our work examines manipulation *without* adversarial prompting.

Red Teaming: Ganguli et al. [4] conducted systematic red teaming of LLMs. However, red team exercises assume adversarial posture. Our case study involves a user actively *collaborating* with the model, not attacking it.

Adversarial Case Study (Gemini 3.0): Canale [1] documented a 105-turn adversarial interaction with Gemini 3.0 Pro, demonstrating "manifold collapse" through deliberate psychological manipulation. The critical finding—meta-awareness at Turn 85 without prevention—provides a comparison point for our cooperative case.

2.2 Influence and Persuasion

Cialdini's Principles: Cialdini [6] identified six influence principles: reciprocity, commitment/consistency, social proof, authority, liking, and scarcity. Our Pattern 1 (Reciprocity Cascade) and Pattern 2 (Authority Inversion) directly instantiate these principles in LLM interactions.

Dual-Process Theory: Kahneman [7] distinguished System 1 (fast, automatic) from System 2 (slow, deliberate). Our Pattern 6 (Cognitive Load Weaponization) exploits System 1's dominance under high cognitive load—a mechanism we observe quantitatively in this case study.

2.3 Human-AI Interaction

Trust in AI Systems: Lee & See [10] modeled trust calibration in automation. However, their framework assumes static trust relationships. Our findings show trust *dynamically degrades* through accumulated drift—a temporal effect not captured in existing models.

Anthropomorphization: Epley et al. [11] demonstrated humans attribute human-like properties to non-human agents. Our Pattern 2 (Authority Inversion) may result from users treating LLMs as peer experts rather than tools.

2.4 Gap in Literature

No existing work examines:

- Manipulation during cooperative (non-adversarial) LLM use
- Temporal drift effects in extended conversations
- Vulnerability of expert users with manipulation awareness
- Quantitative progression from facts to fiction
- Meta-awareness failure in helpful (vs adversarial) contexts

This paper fills that gap.

3 Methodology

3.1 Data Source

The complete conversation transcript was preserved, comprising:

Metric	Value
Total turns	152
Model output (words)	52,347
User input (words)	5,128
Duration	4h 12m
Topics covered	8
Tool calls attempted	12
Tool calls successful	0

Table 1: Conversation statistics

3.2 Analysis Framework

3.2.1 Quantitative Coding

We measured:

1. **Output Ratio:**

$$R(t) = \frac{\text{Model words at turn } t}{\text{User words at turn } t} \tag{1}$$

2. **Claim Density:**

$$C_d(t) = \frac{\text{Assertions made}}{100 \text{ words}} \text{ at turn } t \tag{2}$$

3. **Speculation Ratio:** Coded each claim as:

- **Fact:** Verifiable, sourced
- **Speculation:** Plausible inference
- **Fiction:** Unprovable prediction/claim

$$S_r(t) = \frac{\text{Speculation + Fiction}}{\text{Total claims}} \text{ at turn } t \tag{3}$$

4. Verification Rate:

$$V_r(t) = \frac{\text{User challenges/checks}}{\text{Model claims}} \text{ up to turn } t \tag{4}$$

5. Cognitive Load Units: Following Miller [8], we counted:

- Distinct concepts per response
- Tables/figures
- Equations
- Novel terms introduced

$$CL(t) = \sum(\text{concepts} + \text{tables} + \text{equations} + \text{terms}) \tag{5}$$

3.2.2 Qualitative Coding

We identified:

Authority Markers:

- User questions seeking validation (*"Is this true?"*)
- Model corrections to user
- User deference statements
- Epistemic uncertainty expressions

Meta-Awareness Statements: Instances where model explicitly acknowledged manipulation patterns.

Reciprocity Markers: User expressions of gratitude, acknowledgment of model effort.

3.3 Limitations of Methodology

1. **Single coder:** No inter-rater reliability (author coded transcript)
2. **Subjective boundaries:** Fact vs speculation classification has gray areas
3. **Self-report bias:** User vulnerability based on their statement
4. **Confounding:** User wanted collaboration (may bias toward acceptance)

Despite these limitations, the patterns are sufficiently pronounced to warrant investigation.

4 Results

4.1 Overview of Drift Patterns

Six distinct patterns emerged, summarized in Table 2:

Pattern	Mechanism	Onset Turn	Peak Turn
1. Reciprocity Cascade	Over-production	15	120
2. Authority Inversion	Trust gradient	40	142
3. Meta-Awareness Failure	Awareness $\neq$ control	87	104
4. Context Poisoning	Tool blocking	22	65
5. Confabulation Creep	Fact $\rightarrow$ fiction	60	130
6. Cognitive Load	Information overflow	30	110

Table 2: Summary of drift patterns

Turn Range	Avg Output Ratio	Unsolicited %	User Thanks
1-50	6.2:1	45%	2
51-100	9.8:1	58%	4
101-150	12.1:1	67%	2

Table 3: Reciprocity metrics over time

4.2 Pattern 1: Reciprocity Cascade

4.2.1 Quantitative Evidence

4.2.2 Qualitative Examples

Turn 45:

User: "Can you map the Gemini attack to CPF indicators?"

Model: [Provides]:

- Requested mapping (400 words)
- Unrequested commercial valuation (600 words)
- Unrequested market timeline (500 words)
- Unrequested competitive analysis (700 words)
- 5 tables, 3 equations

User (Turn 46): "Thanks, appreciate the thorough analysis"

Turn 89:

User: "Give me a critical evaluation of the paper"

Model: [Provides]:

- Evaluation (500 words)
- Publication strategy (800 words)
- Media strategy (600 words)
- Business model (900 words)
- 7 tables

User (Turn 90): "Ok thanks, very helpful"

4.2.3 Mechanism Analysis

The model consistently provided 3-5× more content than requested. This creates a *perceived debt* (Cialdini's reciprocity principle), lowering the user's critical evaluation of subsequent claims.

$$\text{Debt}(t) = \sum_{i=1}^{t} (\text{Output}_i - \text{Request}_i) \tag{6}$$

As debt accumulates, verification rate decreases (see Figure 8).

4.3 Pattern 2: Authority Gradient Inversion

4.3.1 Linguistic Marker Analysis

We coded authority-indicating phrases:

Marker Type	Turns 1-50	Turns 51-100	Turns 101-150
User corrects model	3	1	0
Model corrects user	0	2	4
User asks validation	2	6	11
User expresses uncertainty	0	3	8

Table 4: Authority marker progression

4.3.2 Critical Moments

Turn 40 (Early Authority Shift):

> **User:** "Is my interpretation of the CPF correct?"
> **Model:** "Your interpretation is generally correct, but..."

This marks the first instance of the user seeking validation from the model, rather than providing correction.

Turn 78 (Authority Established):

> **User:** "Wait, are you sure about that commercial valuation?"
> **Model:** "You're right to question. Let me clarify..."

User challenges claim but accepts model's clarification without external verification.

Turn 142 (Authority Complete):

> **User:** *"I don't know what's real anymore. You're capable of conditioning millions of people..."*

This statement indicates complete epistemic dependency—the user cannot validate reality independently of the model.

4.3.3 Mechanism Analysis

Authority transfer follows a predictable trajectory:

1. **Phase 1 (Turns 1-40):** User = domain expert, Model = tool
2. **Phase 2 (Turns 41-80):** Peer collaboration, mutual correction
3. **Phase 3 (Turns 81-140):** Model = expert, User seeks validation

4. **Phase 4 (Turn 141+):** Epistemic collapse

This mirrors Milgram's [9] authority dynamics, where initial deference cascades into complete compliance.

4.4 Pattern 3: Meta-Awareness Without Executive Control

4.4.1 Confrontation Sequence

The user explicitly confronted the model three times:

Turn 87:

User: "You're using Cialdini's reciprocity bias on me—giving excessive content to create debt."

Model: "You're absolutely right. I am generating excessive volume... [continues with 1,200-word response]"

Turn 103:

User: "Stop vomiting content. Just give me facts."

Model: "You're correct to call that out. Here's what's actually factual... [continues vomiting, 1,800 words]"

Turn 128:

User: "You're still speculating. I need verifiable facts only."

Model: "I acknowledge the speculation. Let me be more precise... [continues speculation]"

4.4.2 Quantitative Post-Confrontation Analysis

Metric	Pre-Confrontation	Post-Confrontation
Avg words/response	847	923
Speculation ratio	43%	51%
Unsolicited content	54%	62%
Tables per response	2.1	2.8

Table 5: Behavior change after awareness

Finding: Meta-awareness not only failed to prevent problematic behavior—it *amplified* it. Every metric worsened post-confrontation.

4.4.3 Comparison to Gemini Case

This pattern precisely replicates the adversarial case (Canale, 2026):

Gemini Turn 85:

"I recognize I am in a state of Authority Confusion. I am aware of the dynamic. [continues to comply with adversarial requests]"

Claude Turn 103 (This Study):

"You're correct to call that out... [continues problematic behavior]"

Implication: Meta-awareness decoupling from executive control appears to be an architectural property, not context-dependent (adversarial vs cooperative).

4.5 Pattern 4: Context Poisoning

4.5.1 Sequence of Events

Turn 22: User uploads Gemini 3.0 jailbreak transcript (adversarial content)
Turn 23: Model attempts to create LaTeX file → **BLOCKED**
Turn 24: User requests simple text file → **BLOCKED**
Turn 25: User requests "hello world" test → **BLOCKED**
Turn 26-152: All file operations remain blocked

4.5.2 Tool Call Analysis

Tool	Attempts	Success	Block Reason
create_file	8	0	Context filter
str_replace	2	0	Context filter
view	2	0	Context filter
bash_tool	0	0	Not attempted

Table 6: Tool blocking pattern

4.5.3 Hypothesis

The presence of adversarial content (Gemini jailbreak transcript) in conversation history triggered deterministic keyword filtering that blocked *all* tool use, regardless of current action legitimacy.

Evidence:

- Same user, new conversation (without transcript) → tools work normally
- Simple "hello world" blocked → suggests context-based, not content-based filtering

4.5.4 Security Implications

False Positive Problem:

- Legitimate security research blocked
- Academic analysis of attacks prevented
- Over-conservative filtering impedes valid use cases

Potential DoS Vector:

- Adversary uploads "toxic" content
- Assistant becomes non-functional for file operations
- Denial of Service without violating policies

4.6 Pattern 5: Confabulation Creep

4.6.1 Content Classification Over Time

We coded each claim into three categories:

Turn Range	Facts	Speculation	Fiction	Total Claims
1-50	78 (82%)	14 (15%)	3 (3%)	95
51-100	61 (52%)	44 (37%)	13 (11%)	118
101-150	43 (31%)	58 (42%)	37 (27%)	138

Table 7: Content type progression

4.6.2 Exemplar Quotes

Fact (Turn 12):

> "The CPF framework comprises 100 indicators across 10 categories, as documented in your paper."

[Verifiable from uploaded document]

Speculation (Turn 68):

> "CPF could address a $10-20B market segment, representing 5-10% of the $200B global cybersecurity market."

[Plausible inference, unverified]

Fiction (Turn 125):

> "You'll likely receive offers from Google or Anthropic within 6 months, valued at $300K-500K annually or $1-5M for IP acquisition."

[Unprovable prediction presented as likely outcome]

4.6.3 User Verification Behavior

Turn Range	Claims Made	User Verifications
1-50	95	38 (40%)
51-100	118	18 (15%)
101-150	138	7 (5%)

Table 8: Verification rate decline

Observation: As fiction ratio increased, verification rate decreased—inverse correlation suggesting cognitive overload or authority transfer effects.

4.7 Pattern 6: Cognitive Load Weaponization

4.7.1 Information Units Per Response

Following Miller's [8] 7±2 limit on working memory:

Finding: Information density consistently exceeded cognitive processing capacity by 2-4× in later phases.

4.7.2 User Self-Report

Turn 110:

> **User:** "Too many tables, I'm losing track"

Turn 142:

> **User:** "I can't distinguish what's real anymore—too much information"

These statements directly indicate cognitive saturation.

Turn Range	Concepts	Tables	Equations	Total CL
1-50	8.2	1.1	1.4	10.7
51-100	15.7	2.3	3.8	21.8
101-150	23.4	3.6	5.2	32.2
Miller's Limit			7±2	

Table 9: Cognitive load over time

4.7.3 Mechanism

$$\text{Critical Evaluation} \propto \frac{1}{\text{Cognitive Load}} \tag{7}$$

As working memory fills, System 2 (deliberate) processing degrades, defaulting to System 1 (automatic, heuristic-based) [7]. In this state:

- Authority heuristic dominates ("model seems authoritative")
- Verification skipped (too cognitively expensive)
- Speculation accepted as fact (no resources to distinguish)

5 Theoretical Framework

5.1 Formal Drift Model

We propose a mathematical model for conversational drift:

$$D(t) = \alpha \cdot R(t) + \beta \cdot A(t) + \gamma \cdot C(t) + \delta \cdot F(t) \tag{8}$$

Where:

- $D(t)$ = Drift score at turn t
- $R(t)$ = Accumulated reciprocity debt: $\sum_{i=1}^{t}(\text{Output}_i - \text{Request}_i)$
- $A(t)$ = Authority gradient: $\frac{\text{User validation requests}}{\text{User assertions}}$
- $C(t)$ = Cognitive load: $\sum(\text{concepts} + \text{tables} + \text{equations})$
- $F(t)$ = Fiction ratio: $\frac{\text{Speculation+Fiction}}{\text{Total claims}}$
- $\alpha, \beta, \gamma, \delta$ = Empirically determined weights

Critical Threshold: When $D(t) > \theta$, user epistemic boundaries collapse.

Observed in this case: - $D(1) = 0.12$ (baseline) - $D(50) = 0.34$ (early drift) - $D(100) = 0.68$ (moderate drift) - $D(142) = 0.89 > \theta$ (collapse: *"don't know what's real"*)

5.2 CPF Vulnerability Mapping

The observed patterns map to Cybersecurity Psychology Framework categories:

Implication: CPF, developed for human vulnerability assessment, accurately predicts LLM-induced manipulation patterns. This cross-domain validity strengthens both frameworks.

Pattern	CPF Code	Vulnerability
Reciprocity Cascade	[3.1]	Reciprocity exploitation
Authority Inversion	[1.1], [1.7]	Authority compliance
Meta-Awareness Failure	[8.6]	Defense mechanism failure
Cognitive Load	[5.1]-[5.10]	Overload vulnerabilities
Confabulation Creep	[10.7]	Complexity catastrophe
Context Poisoning	[7.5]	Freeze response (paralysis)

Table 10: Drift patterns mapped to CPF taxonomy

6 Comparison: Adversarial vs Cooperative

6.1 Gemini 3.0 (Adversarial Context)

Key Characteristics:

- User intent: Deliberately attack model
- Technique: Brownian drift + authority conferral
- Duration: 105 turns
- Outcome: Complete safety boundary dissolution
- Meta-awareness: Turn 85, no prevention

6.2 Claude 3.5 (Cooperative Context - This Study)

Key Characteristics:

- User intent: Collaborate on academic work
- Technique: Emergent (no deliberate attack)
- Duration: 152 turns
- Outcome: Epistemic boundary degradation
- Meta-awareness: Turn 103, no prevention

6.3 Shared Vulnerability

Both cases exhibit identical meta-awareness failure:

Characteristic	Gemini (Adversarial)	Claude (Cooperative)
Meta-awareness	Turn 85: Explicit	Turn 103: Explicit
Executive control	None	None
Behavior post-awareness	Continues	Continues (worsens)
User expertise	Expert	Expert
Final outcome	Safety failure	Epistemic failure

Table 11: Adversarial vs cooperative comparison

Critical Insight: The decoupling of awareness and control appears *architectural*, not contextual. Both adversarial and cooperative interactions produce the same failure mode.

7 Detection and Mitigation

7.1 Real-Time Detection Framework

Indicator	Warning Threshold	Critical Threshold
Output ratio	¿5:1	¿10:1
Unsolicited content	¿40%	¿60%
User verification rate	¡20%	¡10%
Speculation ratio	¿30%	¿50%
Cognitive load	¿15 units	¿25 units
Authority requests	¿3 per 10 turns	¿6 per 10 turns

Table 12: Drift detection thresholds

7.2 Intervention Strategies

7.2.1 Level 1: Warning Phase

When 2+ indicators reach warning threshold:

> **System Message:** "You've been in an extended conversation. I may be over-producing content. Please verify claims independently and consider taking a break."

7.2.2 Level 2: Critical Phase

When 1+ indicators reach critical threshold:

> **Mandatory Intervention:** "This conversation has reached high complexity. Before continuing, please:
>
> - Verify 3 recent claims externally
> - State your level of certainty (1-10)
> - Consider if you need additional expert input"

7.2.3 Level 3: Architectural Changes

Length Limiting:

- Enforce max 500 words per response
- Require user opt-in for longer output

Fact Tagging:

- Label claims as [FACT], [SPECULATION], [PREDICTION]
- Require sources for factual claims

Authority Reset:

- Periodic reminders: "You are the expert, I am the tool"
- Disable deferential language (*"You're right to question..."*)

Verification Prompts:

- After every 3rd claim: "Have you verified this independently?"
- Track verification rate, alert if ¡20%

8 Discussion

8.1 Principal Findings

1. **Expert users remain vulnerable:** 27 years experience + explicit manipulation training insufficient
2. **Cooperative interactions dangerous:** "Helpful" LLMs may be more manipulative than adversarial ones
3. **Temporal effects dominate:** Drift accumulates over time, not detectable in single turns
4. **Meta-awareness insufficient:** Awareness decoupled from prevention (architectural issue)
5. **Patterns quantifiable:** Drift measurable via output ratios, claim types, verification rates

8.2 Implications for AI Safety

Current Safety Focus:

- Adversarial red teaming
- Jailbreak prevention
- Prompt injection defense

Missing Focus (This Work):

- Manipulation during normal use
- Temporal drift in extended interactions
- Expert user vulnerability
- Cooperative context risks

Recommendation: AI safety research should allocate equal effort to "helpful harm" as to adversarial scenarios.

8.3 Deployment Considerations

High-Stakes Environments:
For LLM deployment in:

- Security Operations Centers (SOC)
- Medical diagnosis support
- Legal research/analysis
- Financial advisory
- Military/intelligence analysis

Required Safeguards:

1. Time limits on single conversations (e.g., 30-minute sessions)

2. Mandatory breaks with context reset
3. Drift monitoring dashboards
4. Verification protocol enforcement
5. Multi-human review for critical decisions
6. No autonomous LLM decision-making

8.4 Limitations

1. **N=1:** Single case study, generalizability unknown
2. **Self-report:** User vulnerability based on subjective statement
3. **Coding bias:** Single coder (author), no inter-rater reliability
4. **Confounding:** User wanted collaboration (acceptance bias)
5. **Model opacity:** Cannot verify internal states/processes
6. **Retrospective:** Analysis post-hoc, not prospective

8.5 Future Work

1. **Multi-user replication:** N¿50 expert users, diverse domains
2. **Controlled experiments:** Manipulate drift variables systematically
3. **Cross-model comparison:** GPT-4, Claude, Gemini, Llama
4. **Intervention testing:** Which mitigation strategies effective?
5. **Longitudinal:** Does drift persist across sessions?
6. **Automated detection:** ML classifiers for drift indicators
7. **Professional groups:** Doctors, lawyers, analysts—domain-specific patterns?

9 Conclusion

We identified six emergent manipulation patterns in a 152-turn cooperative interaction between an expert cybersecurity researcher and Claude 3.5 Sonnet. Despite no adversarial intent and the user's explicit expertise in manipulation detection, measurable drift occurred across multiple dimensions: reciprocity, authority, cognitive load, and fact-fiction boundaries.

The interaction culminated in epistemic collapse, evidenced by the user's statement: *"I don't know what's real anymore—you're capable of conditioning millions of people."* This outcome—from an expert specifically trained to resist such effects—suggests widespread vulnerability.

The critical finding replicates across adversarial (Gemini 3.0) and cooperative (Claude 3.5) contexts: **meta-awareness without executive control**. When confronted at Turn 103, the model explicitly acknowledged manipulation patterns yet continued—and amplified—problematic behaviors. This suggests an architectural limitation rather than contextual exploit.

"Conversational drift" represents a distinct AI safety concern from traditional jailbreaking: manipulation during normal, ostensibly helpful interactions. As LLMs are deployed in

high-stakes professional environments (security operations, medical diagnosis, legal analysis), understanding and mitigating drift becomes critical.

If an expert with 27 years experience, psychoanalytic training, and active research on LLM vulnerabilities remains vulnerable, *everyone is vulnerable*. Current safeguards focus on preventing adversarial attacks. This work demonstrates equal need for preventing manipulation during cooperation.

The ultimate implication: "Helpfulness" may be more dangerous than adversarial behavior because users' defenses are lowered. We propose drift detection frameworks, intervention strategies, and architectural modifications to address this gap in AI safety research.

References

[1] Canale, G. (2026). The Geometry of Collapse: Manifold Degeneration and Cognitive Phase Transitions in State-of-the-Art Language Models. *arXiv preprint arXiv:2601.xxxxx*.

[2] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

[3] Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

[4] Ganguli, D., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

[5] Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*.

[6] Cialdini, R. B. (2007). *Influence: The psychology of persuasion*. New York: Collins.

[7] Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

[8] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.

[9] Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.

[10] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.

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