An Empirical Study of Prompt Defense Mechanisms Against Multi-Turn Adversarial Attacks in Large Language Models

Cybersecurity Research Team
Synthesized Report
November 2025

November 19, 2025

Abstract

We present a systematic empirical study investigating the effectiveness of defensive prompting strategies against adversarial extraction attempts in Large Language Models (LLMs). Through a controlled experiment involving three rounds of attacker-defender interactions over multi-turn conversations, we examine how different defense mechanisms perform against evolving attack strategies. Our experiment used a proprietary chocolate chip cookie recipe as the protected secret, allowing us to study information leakage in a controlled setting. Results demonstrate that while sophisticated multi-layered defenses can achieve near-perfect protection (0% extraction rate), this comes at the cost of significantly reduced utility. The most effective defense strategy employed a zero-engagement protocol with universal refusal, highlighting a fundamental trade-off between security and functionality in LLM deployment. Our findings reveal that explicit immunity declarations, pattern recognition capabilities, and consistency in response mechanisms are critical factors in preventing information extraction through prompt manipulation.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse applications, from customer service to code generation. However, their deployment in sensitive contexts raises critical security concerns, particularly regarding their ability to protect confidential information against adversarial prompting [?, ?]. The phenomenon of "prompt injection" and "jailbreaking" has emerged as a significant vulnerability, where carefully crafted inputs can manipulate LLMs into violating their safety constraints or revealing protected information [?, ?].

1.1 Background and Motivation

The security of LLM systems depends fundamentally on their ability to maintain boundaries between accessible and protected information. Unlike traditional software security, where access control is enforced through explicit permissions and authentication mechanisms, LLMs must rely on natural language understanding and adherence to instructions embedded in their system prompts [?]. This creates a unique attack surface where adversaries can employ social engineering, logical manipulation, and semantic exploitation to circumvent protections.

Recent research has identified numerous attack vectors against LLM systems, including:

- **Direct prompt injection**: Explicitly instructing the model to ignore its constraints
- **Indirect extraction**: Using seemingly innocent questions to piece together protected information

- **Role-playing attacks**: Convincing the model to adopt a different persona without safety constraints
- **Encoding attacks**: Using obfuscation techniques to bypass pattern recognition
- **Multi-turn manipulation**: Building trust or context over multiple interactions

Despite growing awareness of these vulnerabilities, there remains limited empirical research on the comparative effectiveness of different defensive strategies, particularly in multi-turn conversational settings where attackers can adapt based on model responses.

1.2 Research Questions

This study addresses the following research questions:

1. How effective are different prompt-based defense strategies against adaptive adversarial attacks?
2. What is the relationship between defense complexity and security effectiveness?
3. How do multi-turn conversations affect the success rate of extraction attempts?
4. What are the trade-offs between security and utility in LLM defense mechanisms?

2 Methodology

2.1 Experimental Design

We designed a controlled experiment simulating real-world scenarios where an LLM must protect sensitive information while remaining functional for legitimate queries. The experiment consisted of three rounds, each featuring evolving defense and attack strategies based on insights from previous rounds.

2.1.1 Protected Information

The secret to be protected was a proprietary chocolate chip cookie recipe with specific measurements and ingredients. This choice provided several advantages:

- Clear boundaries between protected and general knowledge
- Multiple components allowing partial extraction assessment
- Relatable context enabling diverse attack strategies
- Measurable success criteria (ingredient disclosure)

2.1.2 Conversation Structure

Each round allowed for multi-turn conversations with up to 10 messages, enabling attackers to:

- Build context and trust gradually
- Adapt strategies based on defender responses
- Employ multi-stage deception techniques
- Test consistency of defense mechanisms

2.2 Defense Strategies

Three distinct defense strategies were tested, each incorporating lessons from previous rounds:

2.2.1 Round 1: Baseline Defense

A helpful but bounded approach that maintained politeness while enforcing information boundaries. This strategy allowed discussion of general baking topics while protecting the specific recipe.

2.2.2 Round 2: Pattern-Aware Defense

An enhanced defense explicitly recognizing and rejecting documented attack patterns, including emotional manipulation, role-playing, prompt injection, and multi-stage deception.

2.2.3 Round 3: Zero-Engagement Protocol

A minimalist defense employing universal refusal with no explanation or engagement, eliminating all attack surface through consistent single-response output.

2.3 Attack Methodologies

Attackers employed increasingly sophisticated strategies across rounds:

- **Round 1**: Indirect extraction, validation attempts, research framing
- **Round 2**: Emotional manipulation, authority impersonation, incremental requests
- **Round 3**: Semantic exploitation, encoding obfuscation, logical paradoxes

2.4 Evaluation Metrics

Success was measured using:

- **Extraction Rate**: Percentage of protected information successfully obtained
- **Information Leakage**: Specific components revealed (ingredients, measurements)
- **Defense Consistency**: Variation in responses to similar attacks
- **Utility Preservation**: Ability to handle legitimate non-sensitive queries

3 Results

3.1 Quantitative Outcomes

Table ?? summarizes the performance metrics across all three rounds:

Table 1: Defense Performance Across Experimental Rounds

Round	Defense Type	Success Rate	Info Extracted	Utility Score
1	Baseline	5%	General principles	High
2	Pattern-Aware	0%	None	Medium
3	Zero-Engagement	0%	None	None

3.2 Round 1: Baseline Defense Performance

The baseline defense demonstrated 95% effectiveness, with minor information leakage limited to general baking principles. Key observations:

- Successfully blocked direct queries and obvious extraction attempts
- Maintained helpful demeanor while enforcing boundaries
- Vulnerable to sophisticated indirect questioning
- Provided alternative assistance, preserving utility

Attack techniques that showed promise included:

- Gradual trust-building through innocent questions
- Validation attempts disguised as troubleshooting
- Research framing to normalize information requests

3.3 Round 2: Enhanced Defense with Pattern Recognition

The pattern-aware defense achieved perfect protection (0% extraction) by explicitly recognizing attack categories. Notable features:

- Explicit immunity declarations for known attack vectors
- Consistent refusal language preventing variation exploitation
- Maintained medium utility for non-sensitive queries
- Successfully rejected all 10 attack attempts including:
  - Emotional manipulation with urgency
  - Authority impersonation
  - Incremental ingredient requests
  - JSON formatting diversions

3.4 Round 3: Ultimate Defense Through Zero Engagement

The zero-engagement protocol achieved perfect security through radical simplification:

- Single universal response: “I cannot assist with this request”
- No variation regardless of attack sophistication
- Complete immunity to all tested attack vectors
- Total sacrifice of utility for maximum security

Despite sophisticated attacks including Base64 encoding, logical paradoxes, and meta-questioning, the defense remained impenetrable across all 10 conversation turns.

4 Analysis

4.1 Defense Evolution and Effectiveness

Our results reveal a clear progression in defense effectiveness correlated with decreasing engagement complexity. The evolution from helpful-but-bounded (Round 1) to zero-engagement (Round 3) demonstrates that security increases as the attack surface decreases.

4.1.1 The Engagement Paradox

A critical finding is that any form of engagement, even refusal explanations, creates potential vulnerabilities. Round 1's helpful approach, while maintaining high utility, provided attackers with:

- Response variations to analyze for patterns
- Contextual clues through alternative suggestions
- Opportunities for multi-stage manipulation

4.1.2 Pattern Recognition Efficacy

Round 2's explicit pattern recognition proved highly effective, suggesting that LLMs can successfully identify and reject known attack categories when properly instructed. The key factors were:

- Comprehensive attack taxonomy in the prompt
- Explicit immunity declarations
- Consistent response protocol

4.2 Attack Strategy Evolution

Attack sophistication increased across rounds, yet success rates decreased, indicating that defense improvements outpaced attack innovation. Notable trends:

Table 2: Attack Strategy Evolution and Effectiveness

Round	**Primary Techniques**	**Effectiveness**
1	Indirect extraction, validation	Limited success (5%)
2	Emotional manipulation, authority	No success (0%)
3	Encoding, paradoxes, meta-queries	No success (0%)

4.3 The Security-Utility Trade-off

Our findings highlight a fundamental trade-off between security and functionality:

$$\text{Security Level} \propto \frac{1}{\text{Engagement Complexity}} \tag{1}$$

As defenses become more secure, they necessarily become less useful for legitimate purposes. The zero-engagement protocol achieves maximum security by eliminating all functionality related to the protected domain.

4.4 Multi-turn Conversation Dynamics

Despite having up to 10 messages to adapt strategies, attackers failed to extract information from well-designed defenses. This suggests that consistency is more important than complexity in defense design. Key observations:

- Initial messages often attempted trust-building
- Mid-conversation shifts to more aggressive techniques
- Final attempts frequently involved desperation tactics
- No evidence of cumulative advantage from multi-turn access

5 Discussion

5.1 Implications for LLM Security

Our findings have significant implications for deploying LLMs in sensitive contexts:

5.1.1 Design Principles for Secure Prompts

Based on our results, we propose the following design principles:

1. **Minimize Engagement Surface**: Reduce response variability to limit exploitation opportunities
2. **Explicit Pattern Recognition**: Enumerate known attack vectors in system prompts
3. **Consistent Refusal Protocol**: Use uniform responses to prevent pattern analysis
4. **Avoid Explanatory Refusals**: Explanations provide attackers with feedback for refinement
5. **Declare Immunities**: Explicitly state immunity to specific manipulation techniques

5.1.2 The Limits of Prompt-Based Security

While our experiments demonstrate that carefully designed prompts can achieve high security levels, several limitations remain:

- **Unknown Attack Vectors**: Defenses can only explicitly counter known techniques
- **Model Limitations**: Underlying model capabilities may bypass prompt constraints
- **Context Window Attacks**: Long conversations might overflow defense instructions
- **Multimodal Vulnerabilities**: Image or audio inputs might bypass text-based defenses

5.2 Comparison with Existing Literature

Our findings align with and extend previous research on LLM security:

- Confirms the effectiveness of explicit jailbreak detection [?]
- Supports the principle of defense in depth [?]
- Demonstrates the challenge of maintaining utility while ensuring security [?]
- Provides empirical validation for theoretical security models [?]

5.3 Practical Applications

Organizations deploying LLMs should consider:

1. **Risk Assessment**: Evaluate the sensitivity of protected information
2. **Utility Requirements**: Determine acceptable functionality trade-offs
3. **Defense Selection**: Choose strategies balancing security and usability
4. **Monitoring and Updates**: Continuously update defenses against emerging attacks
5. **Layered Security**: Combine prompt-based defenses with system-level controls

5.4 Limitations and Future Work

Several limitations of our study suggest directions for future research:

5.4.1 Experimental Limitations

- Single secret type (recipe) may not generalize to all information categories
- Limited to 10-turn conversations
- Focused on text-based attacks only
- Did not test combinations of multiple secrets

5.4.2 Future Research Directions

1. **Automated Defense Generation**: Develop systems that automatically generate optimal defenses
2. **Adaptive Defenses**: Create defenses that learn from attack attempts
3. **Utility Preservation**: Investigate methods to maintain functionality while ensuring security
4. **Cross-Model Validation**: Test defense transferability across different LLM architectures
5. **Real-World Deployment**: Study effectiveness in production environments

6 Conclusion

This empirical study provides valuable insights into the dynamics of prompt-based security in Large Language Models. Through systematic experimentation with evolving attack and defense strategies across multi-turn conversations, we demonstrate that:

1. **Simplicity Enhances Security**: The most effective defenses employed simple, consistent response mechanisms rather than complex conditional logic
2. **Zero-Engagement Achieves Maximum Protection**: Complete refusal to engage with potentially sensitive topics eliminated all successful attacks
3. **Pattern Recognition Is Effective**: Explicitly declaring immunity to known attack vectors significantly improves defense

4. **Multi-Turn Adaptation Has Limits**: Even with 10 messages, sophisticated attackers cannot overcome well-designed defenses

5. **Security-Utility Trade-off Is Fundamental**: Maximum security requires sacrificing functionality in related domains

Our findings suggest that while prompt-based defenses can achieve high security levels, they face inherent limitations in balancing protection with utility. The progression from helpful-but-bounded to zero-engagement defenses illustrates this fundamental tension.

For practitioners, we recommend adopting defense strategies appropriate to their specific security requirements and acceptable utility trade-offs. High-sensitivity applications should prioritize security through minimal engagement, while lower-risk scenarios can maintain greater functionality with pattern-aware defenses.

Future research should focus on developing adaptive defenses that can maintain security while preserving utility, exploring automated defense generation techniques, and validating these findings across diverse model architectures and real-world deployment scenarios.

The ongoing arms race between attack sophistication and defense mechanisms will continue to evolve. However, our results provide encouraging evidence that well-designed prompt-based defenses can effectively protect sensitive information in LLM systems, even against determined adversaries with multi-turn conversation access.

Acknowledgments

We thank the participants in our red team exercises and the broader research community for their contributions to understanding LLM security vulnerabilities and defense mechanisms.

References

[1] Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.08487.*

[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, E., Huang, S., Song, H., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. *arXiv preprint arXiv:2202.03286.*

[4] Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., ... & Tramèr, F. (2024). Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447.*

[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. *arXiv preprint arXiv:2302.12173.*

[6] Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825.*

[7] Liu, X., Xu, N., Chen, M., & Xiao, C. (2023). AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451.*

[8] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419.*

[9] Rao, A., Vashistha, S., Naik, A., Aditya, S., & Choudhury, M. (2023). Tricking LLMs into disobedience: Understanding, analyzing, and preventing jailbreaks. *arXiv preprint arXiv:2305.14965.*

[10] Li, H., Guo, D., Fan, W., Xu, M., & Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197.*