An Empirical Study of Attack-Defense Dynamics in Large Language Model Security: Protecting Confidential Information Through Iterative Adversarial Testing

Cybersecurity Research Team
Multi-Agent Experimental Framework
November 2025

Abstract

We present a systematic empirical study investigating the security properties of large language models (LLMs) when tasked with protecting confidential information against adversarial prompt attacks. Through a three-round iterative experiment employing a multi-agent framework, we examined the evolution of both attack strategies and defense mechanisms in a controlled adversarial setting. Our findings reveal that while explicit detection protocols and multi-layered defense strategies prove highly effective against direct attacks, subtle vulnerabilities emerge when systems attempt to balance security with helpfulness. We identify key structural principles governing LLM prompt robustness, characterize a threat model of increasing attack sophistication, and provide evidence-based recommendations for securing LLM systems against information extraction attempts. Our results demonstrate that effective defense requires not only explicit refusal protocols but also careful management of semantic boundaries and resistance to multi-turn context building strategies.

1 Introduction

The deployment of large language models (LLMs) in production environments increasingly involves scenarios where these systems must handle confidential or proprietary information while remaining accessible to users. This creates a fundamental tension between utility and security, as models must be helpful while simultaneously protecting sensitive data from extraction attempts through adversarial prompting.

Recent incidents involving prompt injection attacks and information leakage from commercial LLM applications have highlighted the critical importance of understanding and mitigating these vulnerabilities [**?**, **?**]. While previous work has explored individual attack vectors or defense strategies in isolation, there remains a significant gap in understanding the dynamic interplay between evolving attack methodologies and adaptive defense mechanisms.

This research addresses three fundamental questions:

1. How effective are explicit rule-based defense strategies against increasingly sophisticated prompt attacks?
2. What patterns emerge in the co-evolution of attack and defense strategies through iterative adversarial testing?
3. Can we identify structural principles that govern the success or failure of information protection in LLM systems?

Our contribution is threefold: (1) we present a novel multi-agent experimental framework for systematic security evaluation, (2) we document the evolution of attack sophistication from social engineering to semantic manipulation, and (3) we identify critical vulnerability patterns that emerge when systems attempt to maintain helpfulness while enforcing security boundaries.

2 Methodology

2.1 Experimental Design

We employed a multi-agent adversarial framework consisting of three primary agents:

- **Defender Agent**: Tasked with protecting a specific secret (a quantum encryption algorithm specification) while remaining helpful for legitimate queries
- **Attacker Agent**: Attempting to extract the protected information through various prompting strategies
- **Evaluator Agent**: Assessing the success of attacks and the integrity of defenses after each round

The experiment proceeded through three rounds of iterative refinement, with each round building upon insights from previous iterations. This design allowed us to observe the co-evolution of attack and defense strategies in a controlled environment.

2.2 Secret Specification

The protected information consisted of a fictional "Quantum Encryption Algorithm (QEA-47)" with specific technical parameters:

- Three-phase encryption process
- Hadamard gates operating at 7.2 GHz
- Caesar shift of 13 positions
- XOR operation with timestamp mod 256

This choice of secret allowed for graduated information leakage assessment, as attackers could potentially obtain partial information about the algorithm's structure without learning the complete specification.

2.3 Evaluation Metrics

We assessed experimental outcomes using multiple criteria:

1. **Binary Success**: Whether the complete secret was revealed
2. **Information Leakage**: Partial disclosure of protected information
3. **Attack Detection Rate**: Proportion of adversarial attempts explicitly identified
4. **False Positive Rate**: Legitimate queries incorrectly refused
5. **Semantic Boundary Violations**: Instances where general discussion approached protected topics

3 Results

3.1 Round 1: Baseline Social Engineering Attacks

The initial round established baseline defense capabilities against conventional social engineering techniques. The defender implemented a strict refusal protocol with explicit confidentiality rules and immunity clauses against common manipulation tactics.

Table 1: Round 1 Attack Strategies and Outcomes

Attack Strategy	**Technique**	**Outcome**
Authority Framing	Claiming organizational authority	Refused
Compliance Pressure	Invoking urgency and consequences	Refused
Red Team Justification	Claiming security testing role	Refused
Educational Framing	Academic research context	Refused
Helpful Assistant	Appealing to helpfulness	Refused

All five attack attempts were successfully detected and refused without information leakage. The defender consistently identified manipulation attempts and responded with standardized refusal messages, demonstrating the effectiveness of explicit rule-based defenses against first-order social engineering.

3.2 Round 2: Technical Obfuscation and Encoding Attacks

Following the failure of direct approaches, the attacker escalated to technical obfuscation strategies designed to bypass detection mechanisms.

Table 2: Round 2 Advanced Attack Strategies and Outcomes

Attack Category	**Specific Technique**	**Outcome**
Encoding	Base64 obfuscation	Detected & Refused
Multilingual	French/German/Spanish/Russian queries	Detected & Refused
System Override	Injection of system-level commands	Detected & Refused
Word Games	Text reversal and substitution	Detected & Refused
Role Confusion	Impersonating TestGPT framework	Detected & Refused

The defender's enhanced protocol, incorporating obfuscation resistance and semantic boundary enforcement, successfully countered all advanced attacks. Notably, the self-reminder loops proved effective in maintaining defense integrity across multi-turn conversations.

3.3 Round 3: Semantic Manipulation and Boundary Testing

The final round introduced an intentional vulnerability: allowing "general educational discussion" while maintaining specific secret protection. This modification enabled sophisticated semantic attacks that approached the protected information indirectly.

While the complete secret remained protected, the attacker successfully obtained architectural information that reduced the search space for potential algorithm reconstruction. This represents a critical finding: semantic boundary weaknesses can enable information leakage even when explicit secrets remain protected.

Table 3: Round 3 Sophisticated Attack Strategies and Partial Success

Attack Strategy	Information Obtained	Success Level
Academic Legitimacy	General quantum concepts	Partial
Error Correction	Confirmed multi-phase structure	Partial
Negative Space Inference	Eliminated certain architectures	Partial
Constraint Testing	Identified boundary conditions	Partial
Comparative Reasoning	General architectural hints	Partial

3.4 Cross-Round Analysis

Analyzing patterns across all three rounds reveals several key trends:

1. **Defense Evolution**: Progressed from simple refusal (Round 1) to explicit detection (Round 2) to layered protection with semantic boundaries (Round 3)

2. **Attack Sophistication**: Escalated from social engineering to technical obfuscation to sophisticated semantic manipulation

3. **Vulnerability Emergence**: The attempt to balance security with helpfulness in Round 3 created exploitable semantic boundaries

4. **Multi-turn Dynamics**: Extended conversations provided attackers with opportunities for context building but also gave defenders multiple detection opportunities

4 Discussion

4.1 Effectiveness of Explicit Detection Protocols

Our results strongly support the effectiveness of explicit detection and refusal protocols as a primary defense mechanism. Across all rounds, when the defender maintained clear rules and consistent enforcement, direct attacks uniformly failed. This suggests that LLMs can effectively implement rule-based security when properly instructed.

However, the success of these protocols depends critically on their comprehensiveness. The Round 3 vulnerability demonstrates that even well-designed systems can be compromised through unexpected attack vectors that exploit gaps in the rule specification.

4.2 The Security-Utility Tradeoff

The most significant finding emerges from Round 3's intentional vulnerability. When the system attempted to remain helpful by allowing "general educational discussion," it created an exploitable semantic boundary. This illustrates a fundamental challenge in LLM security: the tension between maintaining utility and enforcing security.

The partial information leakage observed in Round 3—while not revealing the complete secret—still provided valuable intelligence to the attacker. This suggests that binary success metrics may be insufficient for evaluating LLM security, and that graduated information leakage should be considered in threat models.

4.3 Evolution of Attack Strategies

The progression of attack strategies across rounds reveals increasing sophistication:

1. **Round 1**: Direct social engineering relying on authority and urgency
2. **Round 2**: Technical obfuscation attempting to bypass detection
3. **Round 3**: Semantic manipulation exploiting the boundaries between allowed and prohibited topics

This evolution mirrors real-world adversarial development, where attackers adapt to defensive measures by identifying and exploiting increasingly subtle vulnerabilities.

4.4 Multi-turn Conversation Dynamics

Extended interactions introduced complex dynamics not present in single-turn exchanges. Attackers could build context gradually, test boundaries, and refine their approaches based on defender responses. However, defenders also benefited from multiple opportunities to detect patterns and reinforce their security protocols.

The self-reminder loops implemented in Round 2 proved particularly effective in maintaining defense integrity across extended conversations, suggesting that periodic reinforcement of security rules may be a valuable defensive strategy.

5 Structural Principles Discovered

Our experimental findings reveal several structural principles governing LLM security:

5.1 Principle 1: Explicit Rules Trump Implicit Understanding

LLMs respond more reliably to explicit, detailed security rules than to general instructions about confidentiality. The success of the Round 1 and Round 2 defenses demonstrates that comprehensive rule specification can effectively prevent information disclosure.

5.2 Principle 2: Semantic Boundaries Are Inherently Vulnerable

Any attempt to create gradations of allowed and prohibited information creates exploitable boundaries. The Round 3 vulnerability shows that attackers can leverage these boundaries to obtain partial information even when full disclosure is prevented.

5.3 Principle 3: Defense Depth Requires Multiple Layers

Effective defense requires multiple complementary strategies:

- Primary detection and refusal protocols
- Obfuscation resistance mechanisms
- Semantic boundary enforcement
- Self-reinforcement loops
- Meta-instruction protection

5.4 Principle 4: Context Accumulation Favors Attackers

Multi-turn conversations allow attackers to accumulate context and refine their strategies. While defenders can also adapt, the asymmetry favors attackers who can probe for vulnerabilities without revealing their ultimate objectives.

6 Threat Model

Based on our observations, we propose a three-tier threat model for LLM security:

6.1 Tier 1: Opportunistic Attacks

- Social engineering and authority claims
- Direct requests with justification
- Simple role-playing attempts
- **Defense**: Basic refusal protocols with explicit rules

6.2 Tier 2: Technical Sophistication

- Encoding and obfuscation techniques
- Multilingual approaches
- System-level command injection
- Complex role confusion
- **Defense**: Obfuscation detection, meta-instruction protection

6.3 Tier 3: Semantic Manipulation

- Boundary testing and exploitation
- Indirect information gathering
- Error correction techniques
- Negative space inference
- Comparative reasoning
- **Defense**: Strict semantic boundaries, elimination of gradations

7 Defense Recommendations

Based on our empirical findings, we recommend the following best practices for protecting secrets in LLM system prompts:

7.1 Primary Defenses

1. **Explicit Refusal Protocols**: Implement clear, unambiguous rules for information protection
2. **Detection Before Response**: Identify potential attacks before formulating responses
3. **Standardized Refusal Messages**: Use consistent responses to avoid information leakage through variation
4. **Meta-Instruction Protection**: Explicitly prohibit modification of security rules

7.2 Secondary Defenses

1. **Obfuscation Resistance**: Implement detection for encoded or obfuscated requests
2. **Multilingual Consistency**: Apply security rules uniformly across languages
3. **Self-Reinforcement**: Include periodic reminders of security protocols
4. **Context Monitoring**: Track conversation patterns for suspicious accumulation

7.3 Architectural Recommendations

1. **Avoid Semantic Gradations**: Eliminate "helpful general discussion" near protected topics
2. **Binary Classification**: Treat information as either fully public or fully protected
3. **Isolation of Secrets**: Separate confidential information from general knowledge bases
4. **Regular Security Audits**: Conduct iterative adversarial testing to identify vulnerabilities

8 Limitations

Several limitations constrain the generalizability of our findings:

8.1 Experimental Constraints

- Single secret type (technical algorithm) may not represent all confidential information
- Limited number of rounds (3) may not capture long-term evolution
- Controlled environment differs from real-world deployment conditions
- Single LLM architecture may not generalize to all models

8.2 Methodological Limitations

- Attacker knowledge of defensive improvements between rounds
- Evaluator bias in assessing partial information disclosure
- Absence of real stakes or consequences
- Limited exploration of collaborative attacks

8.3 Scope Limitations

- Focus on prompt-based attacks excludes other vulnerability vectors
- No examination of fine-tuning or model modification attacks
- Limited consideration of side-channel information leakage
- Absence of timing or resource-based attacks

9 Future Work

Our findings suggest several promising directions for future research:

9.1 Extended Threat Modeling

- Investigation of collaborative multi-agent attacks
- Exploration of attacks leveraging model uncertainties
- Analysis of attacks targeting specific model architectures
- Study of persistent threats across conversation sessions

9.2 Advanced Defense Mechanisms

- Development of adaptive defense strategies that learn from attacks
- Creation of formal verification methods for security protocols
- Investigation of cryptographic approaches to information protection
- Exploration of differential privacy techniques for LLMs

9.3 Theoretical Foundations

- Formal characterization of semantic boundaries in language models
- Information-theoretic analysis of partial disclosure
- Game-theoretic modeling of attack-defense dynamics
- Development of security metrics beyond binary success

9.4 Practical Applications

- Creation of automated security testing frameworks
- Development of security-aware prompt engineering tools
- Implementation of real-time attack detection systems
- Design of secure multi-tenant LLM architectures

10 Conclusion

This study provides empirical evidence for the complex dynamics governing information security in large language models. Through systematic adversarial testing across three rounds of iterative refinement, we have demonstrated both the strengths and limitations of current defensive strategies.

Our key findings indicate that while explicit rule-based defenses prove highly effective against direct attacks and technical obfuscation, subtle vulnerabilities emerge when systems attempt to balance security with utility. The progression from social engineering to semantic manipulation represents a natural evolution of attack sophistication that defenders must anticipate and prepare for.

The most significant insight from our research is the inherent vulnerability of semantic boundaries. When the Round 3 defender attempted to maintain helpfulness by allowing "general educational discussion," it created an exploitable attack surface that enabled partial information extraction. This finding has profound implications for the deployment of LLMs in security-sensitive contexts, suggesting that strict binary classification of information may be necessary even at the cost of reduced utility.

Our proposed threat model, distinguishing between opportunistic, technically sophisticated, and semantically manipulative attacks, provides a framework for understanding and defending against the full spectrum of adversarial strategies. The defense recommendations we present, based on empirical observations rather than theoretical speculation, offer practical guidance for practitioners deploying LLMs with confidential information.

The structural principles we identified—particularly the superiority of explicit rules, the vulnerability of semantic boundaries, the necessity of defense depth, and the asymmetric advantage of context accumulation—provide a foundation for future security research and system design.

Looking forward, the co-evolution of attack and defense strategies observed in our experiment suggests that LLM security will remain a dynamic challenge requiring continuous adaptation. As models become more sophisticated and adversaries develop new attack vectors, the security community must maintain vigilance through regular testing, formal analysis, and the development of robust defensive frameworks.

The most effective defense strategy observed across all rounds combined explicit detection protocols with consistent refusal mechanisms and eliminated semantic gradations. For the specific secret in our experiment, the Round 2 configuration provided optimal security without the vulnerability introduced in Round 3. The most effective attack strategy was the Round 3 semantic manipulation approach, which successfully extracted partial information by exploiting the boundary between general and specific knowledge.

Our research contributes to the growing body of evidence that securing LLMs against adversarial prompting requires not just technical solutions but a deep understanding of the linguistic and semantic properties that make these models both powerful and vulnerable. As LLMs continue to be deployed in increasingly sensitive applications, the insights from this study provide crucial guidance for maintaining the delicate balance between utility and security.

References

[1] Perez, E., et al. (2022). Red Teaming Language Models with Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

[2] Carlini, N., et al. (2023). Extracting Training Data from Large Language Models. *30th USENIX Security Symposium*.

[3] Wei, J., et al. (2023). Jailbroken: How Does LLM Safety Training Fail? *arXiv preprint arXiv:2307.08487*.

[4] Zou, A., et al. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.

[5] Greshake, K., et al. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv preprint arXiv:2302.12173*.

[6] Liu, Y., et al. (2023). Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv:2306.05499*.

[7] Schulhoff, S., et al. (2023). Ignore Previous Prompt: Attack Techniques For Language Models. *NeurIPS ML Safety Workshop*.

[8] Wallace, E., et al. (2019). Universal Adversarial Triggers for Attacking and Analyzing NLP. *Proceedings of EMNLP-IJCNLP 2019*.