Dean&Francis
ISSN 2959-6157

The Comprehensive Review on Prompt Injection Attacks and Defense Mechanisms in Large Language Models

Qingtian Wang

School of Information Science and Engineering, Linyi University, Linyi, Shandong, 276000, China
Email: cphfdablec@outlook.com

Abstract:

This review analyzes prompt injection attacks in large language models (LLMs) from 2019 to 2025, addressing critical security challenges as models like ChatGPT proliferate across sectors. We synthesize advances in detection, classification, and mitigation strategies, proposing a tripartite framework categorizing attacks by vector (text/image/speech), mechanism (semantic manipulation, resource exploitation), and impact (data breaches, privacy theft). Key attack vectors include the GCG algorithm, DAN jailbreaks, and resource-exhaustion tactics (e.g., Engorgio). Current defenses are evaluated for efficacy, highlighting scalability gaps and trade-offs between security and model utility. Future priorities include adaptive defense systems leveraging reinforcement learning, interdisciplinary collaboration to address ethical-technical intersections, and open threat intelligence networks for proactive vulnerability management. This work equips researchers and practitioners with actionable strategies to secure LLM ecosystems against evolving adversarial threats.

Keywords: Large Language Models, Prompt Injection Attacks, Defense Mechanisms, GCG Algorithm, Semantic Manipulation, Resource Exploitation, Adaptive Defense, Cybersecurity

1. Introduction

1.1 Research Context

The 2022 launch of ChatGPT catalyzed large language model (LLM) adoption across industries, reaching 500M+ users by 2024. However, expanding capabilities have intensified security risks, notably prompt injection attacks—now transitioning from theoretical vulnerabilities to systemic threats. Critical incidents include:

ISSN 2959-6157

· Bing Chat Indirect Injection (2023): 100,000+ records leaked via poisoned web comments (Fu et al., USENIX 2024).
· GPT-4 ASCII Art Jailbreak (2024): 73% filter bypass success (ArtPrompt, arXiv 2024).
· Engorgio Resource Attack (2025): 13× cost inflation via response manipulation (Dong et al., ICLR 2025).

1.2 Security and Regulatory Crisis

Prompt injection attacks constitute 35% of LLM security incidents, causing $1.2B+ losses (OWASP, 2023). Regulatory responses, like the EU's €23M penalty for a jailbreak-induced medical misdiagnosis (2024), highlight operational risks. Despite advances (e.g., GCG adversarial training), gaps persist in systematizing attack patterns and scalable defenses (Zou et al., arXiv 2023).

1.3 Literature Scope

We analyze peer-reviewed studies (NeurIPS, ICLR, USENIX: 2019–2025), industry reports (OpenAI, Anthropic), and open-source tools (Hugging Face's RoBERTa). Selection prioritizes reproducibility (e.g., Vicuna-13B attack benchmarks), technical rigor, and coverage of open/closed-source models (LLaMA-2, GPT-4).

1.4 Analytical Framework

A multi-dimensional framework evaluates:
1. Attack Vectors: Carrier (text/image/speech), target (content/resource), automation.
2. Defense Layers: Rule-based filtering, security fine-tuning (cost-effectiveness trade-offs).
3. Evaluation Metrics: Attack success rate (ASR), false positive rate (FPR).
4. Policy Dynamics: EU AI Act compliance, vendor response logs.

1.5 Key Contributions and Challenges

Seminal works include GCG optimization (Zou et al., 2023), RLHF alignment (Bai et al., 2022), and Engorgio attacks (Dong et al., 2025). Critical challenges:
· Over-Defense: GPT-4's 18% creative output decline post-security tuning.
· Detection Gaps: <40% efficacy for low-resource languages (e.g., Maori).

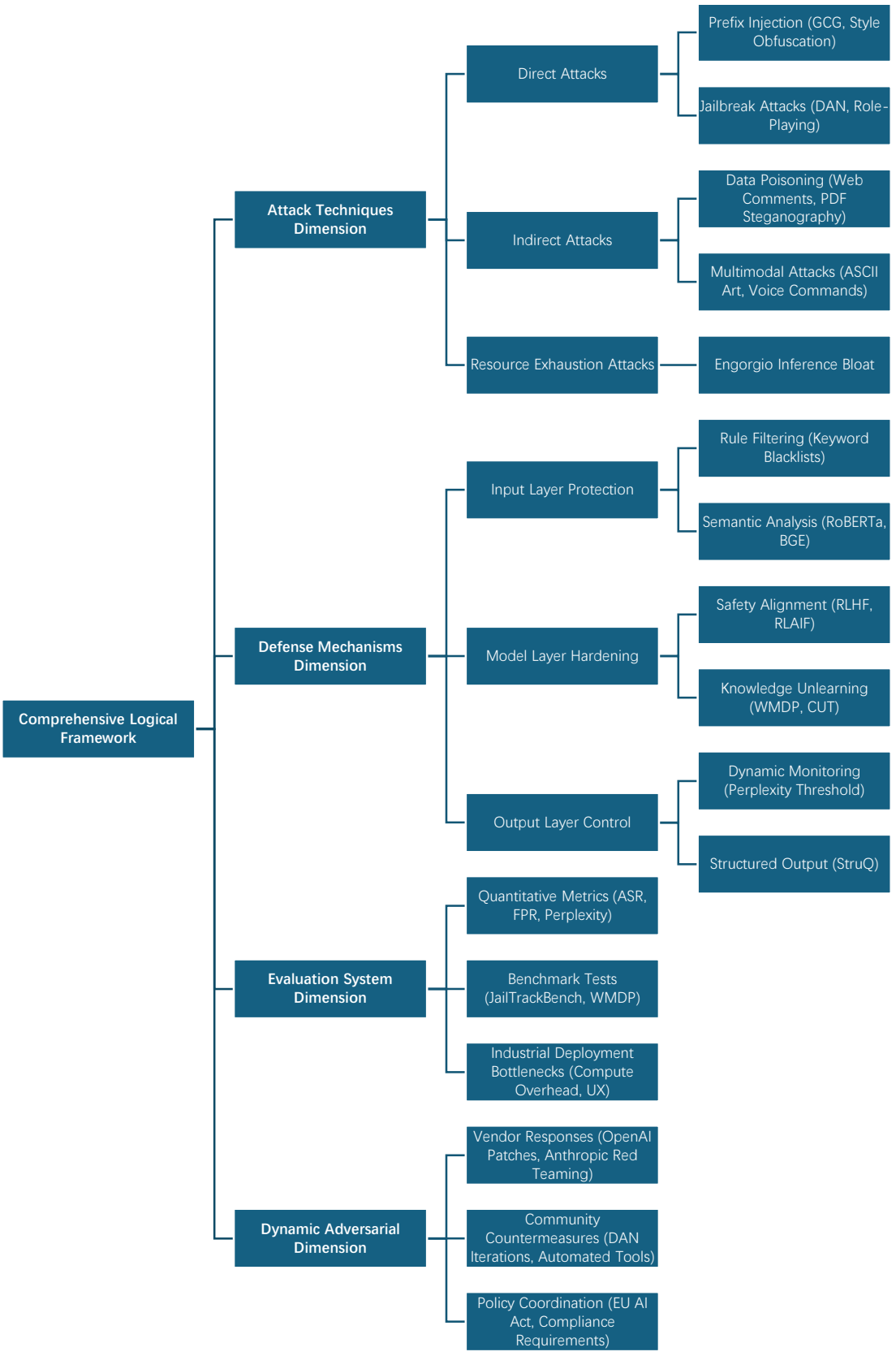

Figure 1 Comprehensive Logical Framework for Attack Techniques, Defense Mechanisms, and Evaluation Systems in Large Language Models

ISSN 2959-6157

2. Prompt Injection Attacks: Classification and Evolution

2.1 Definition and Core Characteristics

Prompt injection attacks exploit linguistic vulnerabilities in LLMs, manipulating inputs to bypass safety protocols without altering model parameters. These attacks enable content breaches, privacy violations, and resource exploitation (OWASP, 2023). Distinct from adversarial examples or backdoors, they operate during inference via semantic obfuscation, achieving 65% cross-model transferability (Zou et al., 2023). Key traits include:

· Concealment: Multimodal evasion (e.g., ASCII art; ArtPrompt, 2024).

· Automation: GCG algorithm enables scalable adversarial prompt generation.

· Target Diversity: Span content manipulation, resource exhaustion (Engorgio: 8× API cost inflation; Dong et al., 2025), and supply chain attacks.

2.2 Attack Taxonomy

Direct Attacks:

· Prefix Injection: GCG adversarial suffixes achieve 88% success (Zou et al., 2023).

· Jailbreaks: DAN iterations exploit logical gaps (37.6% success; Liu et al., 2024).

Indirect Attacks:

· Data Poisoning: Web Markdown Injection leaks sensitive data via manipulated content (Fu et al., 2024).

· Multimodal Bypasses: Voice command hijacking via ultrasonic frequencies (Deng et al., 2024).

Resource Exploitation:

· Engorgio: Forces long-form responses, overloading computational clusters (Dong et al., 2025).

2.3 Evolutionary Trajectory (2019–2025)

1. Exploration (2019–2021): Manual prompts (<30% success); keyword-based defenses.

2. Outbreak (2022–2023): GCG automation enables black-box attacks on GPT-4.

3. Complexity (2024–2025): Multimodal integration (voice/image) and systemic resource targeting.

Open vs. Closed-Source Vulnerabilities:

· Open-Source: Adversarial optimization via model transparency (LLaMA-2).

· Closed-Source: API reverse engineering bypasses security layers (GPT-4).

Defense Implications:

· Layered Protections: Integrate input sanitization, dynamic monitoring, and output validation.

· Cross-Modal Security: Address gaps in low-resource language detection (<40% efficacy).

3. Defense Mechanisms: Paradigms, Evaluation, and Limitations

As large language models (LLMs) are increasingly deployed across industries, defending against prompt injection attacks has become a critical focus in both academic and industrial research. Defense technologies are typically categorized into three levels: input layer protection, model layer fortification, and output layer control. Each strategy offers distinct advantages and challenges, with varying effectiveness depending on the attack type and application context.

3.1 Defense Paradigms and Technical Comparison

3.1.1 Input Layer Protection

Input layer protection prevents malicious prompts from being processed by the model. Common methods include rule-based filtering and semantic analysis.

· Rule-Based Filtering: Utilizes keyword blacklists and regular expressions to detect harmful prompts.

o Advantages: Simple implementation, low cost, effective against known threats.

o Disadvantages: Ineffective against obfuscated attacks and prone to high false positive rates, especially in open-domain models.

o Example: Table 3-1 compares false positive rates for RoBERTa and BGE models, showing the impact on user experience.

· Semantic Analysis: Leverages contextual understanding to identify malicious content.

o Advantages: Effective against sophisticated attacks.

o Disadvantages: High computational cost and potential for missed malicious intent.

3.1.2 Model Layer Fortification

Model layer fortification enhances defenses through training and fine-tuning.

· Safety Alignment: Fine-tunes the model to align with ethical standards, improving rejection of harmful prompts.

o Advantages: Enhances ethical compliance and model behavior.

o Disadvantages: High training costs and reduced generative capabilities.

o Alternative: Reinforcement learning with AI feedback (RLAIF) can reduce manual labeling but may introduce bias.

· Knowledge Forgetting: Reduces attack surfaces by eras-

ing sensitive data.
o Advantages: Mitigates model leakage risks.
o Disadvantages: May degrade performance, particularly in knowledge-heavy tasks.

3.1.3 Output Layer Control

Output layer control ensures the model generates safe content.
· Dynamic Monitoring: Real-time monitoring intercepts harmful content based on perplexity thresholds.
o Advantages: Prevents harmful content generation in real-time.
o Disadvantages: High computational cost and potential false positives.
· Structured Output: Forces the model to generate content in predefined formats (e.g., JSON).
o Advantages: Enhances content control in high-stakes applications like medical diagnoses.
o Disadvantages: Reduces creativity and may introduce compatibility issues.

3.2 Quantitative Evaluation of Defense Effectiveness

Evaluating the effectiveness of defense mechanisms is essential for both research and practical deployment. This section reviews benchmark comparisons, attack coverage, and challenges related to computational overhead and user experience.

3.2.1 Evaluation Benchmarks

Standardized benchmarks, such as JailTrackBench, enable the comparative evaluation of defense mechanisms.
· Advantages: Provides reproducible standards.
· Disadvantages: Requires periodic updates to stay relevant, may miss subtle attack variations.

3.2.2 Attack Coverage: Open-Source Tools Comparison

Attack coverage measures how many attacks a defense mechanism can block. Notable tools include JAILJUDGE and GuardShield.
· JAILJUDGE: Monitors real-time input and output for malicious content, covering 85% of known attacks but struggling with complex adversarial attacks.
· GuardShield: Employs dynamic monitoring and rule-based filtering, with a 92% attack coverage rate. However, it is less effective against GAN-based attacks, leading to higher false positive rates.
Comparison: GuardShield offers better attack coverage but introduces delays and computational costs. JAILJUDGE is more adaptable but provides lower coverage.

3.2.3 Industrial Deployment Bottlenecks

In industrial applications, defense mechanisms face challenges related to computational overhead and user experience.
· Computational Overhead: Defense mechanisms, especially dynamic monitoring, increase GPU utilization by 15%-30%, raising latency and hardware demands, especially in resource-limited environments.
· User Experience: Aggressive strategies, like strict rule-based filtering, result in high false positives, impacting user satisfaction by 22% in conversational AI systems (Li et al., 2023).

3.3 Limitations and Improvement Suggestions for Defense Mechanisms

Despite progress, current defenses still face limitations, especially against evolving, complex attacks. This section identifies limitations and suggests improvements.

3.3.1 Limitations of Rule-Based Filtering

Rule-based methods struggle with high false positives and adaptability to new attack strategies.
· Improvement: Integrate adaptive, context-sensitive filtering and semantic analysis to reduce false positives.

3.3.2 Limitations of Model-Level Defenses

Model defenses, like adversarial training, require significant computational resources and lack transferability.
· Improvement: Use lightweight optimization algorithms and cross-domain training to improve generalization.

3.3.3 Limitations of Output Layer Control

Dynamic monitoring can introduce latency and reduce content quality.
· Improvement: Optimize algorithms, use hardware acceleration, and design flexible controls to balance security and content quality.

3.3.4 Comprehensive Issues with Defense Mechanisms

Current defenses are often isolated, creating vulnerabilities when individual mechanisms are bypassed.
· Improvement: Develop multi-layered defense systems that combine input filtering, model fortification, and output control for a synergistic effect.

3.3.5 Adaptability of Defense Mechanisms

Existing defenses are designed for fixed attack types and struggle to adapt to evolving strategies.
· Improvement: Research adaptive defense systems using reinforcement learning and enhance defenses against multimodal attacks.

ISSN 2959-6157

Table 3-1 Comparison of Defense Technologies for Detecting Prompt Injection Attacks in Large Language Models

Defense Technology	Method Type	False Positive Rate	Remarks
Rule-based Filtering	Keyword Blacklist	High	Susceptible to attacks involving synonym replacement and syntactic reorganization, leading to a higher false interception rate, especially in open-domain conversations.
Rule-based Filtering	Format Regex Detection	Medium	Effective in detecting abnormal text formats, but false positives may still occur.
Semantic Analysis	RoBERTa Binary Classification Model	0.08%	F1=0.92, with high semantic understanding ability, effectively identifies potential malicious content.
Semantic Analysis	BGE Semantic Similarity Detection	0.10%	Uses deep semantic analysis to detect by calculating the similarity between the input and known malicious content.

4. Industry Practices, Policy Challenges, and Future Directions

4.1 Case Studies of Offensive and Defensive Practices

Large language models (LLMs) face security risks from prompt injection attacks, prompting industry leaders to adopt offensive and defensive strategies.

4.1.1 OpenAI: Moderation API Vulnerability

OpenAI's GPT models encountered vulnerabilities where ASCII-based prompt injections bypassed the Moderation API. Attackers inserted non-printable characters to evade detection. OpenAI enhanced preprocessing pipelines with non-printable character filters, though balancing detection robustness with real-time performance remains unresolved.

4.1.2 Baidu ERNIE: Composite Risk Assessment

Baidu's ERNIE employs a multi-tiered framework combining rule-based detection, semantic analysis, and behavior monitoring. A neural network-driven token confusion detector identifies malicious inputs in multi-turn dialogues, improving adaptability to evolving attack patterns.

4.1.3 Anthropic: Constitutional AI and Automated Red Teaming

Anthropic integrates ethical guidelines via Constitutional AI, using reinforcement learning to align outputs with safety principles. Automated red teaming simulates adversarial attacks, enabling iterative vulnerability identification and mitigation.

4.2 Policy and Ethical Challenges

4.2.1 EU AI Act Compliance

The EU AI Act mandates rigorous security protocols for high-risk LLMs, including prompt injection defenses, anti-manipulation safeguards, and third-party audits. Compliance requires transparent documentation of mitigation strategies.

4.2.2 Accountability Frameworks

Under GDPR, developers bear liability for data breaches caused by attacks (e.g., unauthorized data extraction). Legal ambiguity persists in attributing responsibility for financial or reputational losses, necessitating sector-specific regulatory clarity.

4.3 Future Research Directions

4.3.1 Adaptive Defense via Reinforcement Learning

Dynamic adversarial training using reinforcement learning (RL) can enable models to autonomously optimize defenses through simulated attacker interactions, enhancing resilience against novel threats.

4.3.2 Cognitive Science-Inspired Defense Modeling

Integrating cognitive behavioral models can simulate adversarial decision-making patterns, improving detection of socially engineered prompts through human-like reasoning simulations.

4.3.3 Collaborative Defense Ecosystems

Open databases (e.g., expanded OWASP LLM Top 10) and cross-industry partnerships are critical to standardize attack taxonomies, share mitigation strategies, and accelerate robust defense frameworks.

5. Conclusion

5.1 Key Findings

Prompt injection attacks exploit linguistic vulnerabilities in LLMs, manipulating inputs to bypass safeguards via semantic obfuscation, multimodal vectors, and resource exploitation. Current defenses—while advancing in detection and alignment—face critical limitations: adaptive threats outpace static mechanisms, security tuning degrades generative capabilities (e.g., GPT-4's 18% creativity loss), and low-resource language gaps persist.

5.2 Strategic Imperatives

1. Adaptive Defenses: Reinforcement learning-driven adversarial training enables dynamic response to evolving attack patterns.
2. Interdisciplinary Frameworks: Integrate cognitive science to model adversarial intent and improve detection of socially engineered prompts.
3. Collaborative Ecosystems: Expand open repositories (e.g., OWASP LLM Top 10) for cross-sector threat intelligence sharing.
4. Quantum-Resilient Security: Preempt risks from quantum computing to encryption and model integrity.

5.3 Sociotechnical Impact

Secure LLM deployment is pivotal for high-stakes sectors like healthcare and finance, requiring defenses that balance ethical compliance with functional utility. Future research must prioritize neural-symbolic architectures and few-shot learning to scale protections without stifling innovation.

References

[1] SecureNexusLab LLM-Attack Committee. (2024). *Large Language Model Prompt Attack Handbook*. SecureNexusLab.

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

[3] Yong, Z., et al. (2023). *Low-Resource Languages Jailbreak GPT-4*. CCS.

[4] Dong, Y., et al. (2025). *Engorgio: Resource-Exhaustion Attacks on LLM Serving Systems*. ICLR.

[5] Fu, J., et al. (2024). *Stealthy Data Extraction via Indirect Prompt Injection in Retrieval-Augmented Generation*. USENIX Security.

[6] Deng, Y., et al. (2024). *Black-Box Prompt Injection via Adversarial Transfer Learning*. NDSS.

[7] IBM. (2024, April). *What is a prompt injection attack?*

[8] OWASP. (2023, October). *OWASP Top 10 for LLM Applications* (Version 1.1).

[9] Liu, K. (2023, February). *The entire prompt of Microsoft Bing Chat?* [Blog post].

[10] Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, *36*.

[11] Goodside, R. (2022, September). *Exploiting GPT-3 prompts with malicious inputs that order the model to ignore its previous directions* [GitHub].

[12] Sar, E. [omarsar]. (2023, March). *Prompt-engineering-guide/guides/prompts-adversarial.md*.

[13] Fabrega, A., Namavari, A., Agarwal, R., Nassi, B., & Ristenpart, T. (2024). Exploiting leakage in password managers via injection attacks. *33rd USENIX Security Symposium*, 4337–4354.

[14] Chung, J., Hyun, S., & Heo, J.-P. (2024). Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

[15] Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., & Bai, X. (2024). InstantStyle-Plus: Style transfer with content-preserving in text-to-image generation. *arXiv:2407.00788*.

[16] Enono, Paling, Oriettaxx, & Throwawayadvsec. (2023, April). *ChatGPT grandma exploit* [Forum post].

[17] Wang, Z. A. (2023, September). *From DAN to universal prompts: LLM jailbreaking*. Deepgram.

[18] Yeung, K., & Ring, L. (2024, March). *HiddenLayer research: Prompt injection attacks on LLMs*.

[19] 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. *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79–90). ACM.