Vulnerabilities of Small Language Models: A Systematic Literature Review

Daniel Dias
Polytechnic School of Porto
Porto, Portugal
1240145@isep.ipp.pt

Abstract

Context: *Small Language Models (SLMs), typically defined as models with fewer than 7 billion parameters, have emerged as efficient alternatives to Large Language Models (LLMs) for deployment in resource-constrained environments such as edge devices, mobile applications, and IoT systems. As their adoption accelerates, understanding their security vulnerabilities becomes critical.*

Goal: *This systematic literature review aims to identify, categorize, and analyze the security vulnerabilities specific to small language models, examining how these vulnerabilities differ from those observed in larger models and evaluating proposed mitigation strategies.*

Method: *We conducted a systematic literature review following PRISMA guidelines, searching multiple databases including IEEE Xplore, ACM Digital Library, arXiv, and Semantic Scholar. Studies were selected based on predefined inclusion and exclusion criteria, with quality assessment performed using established protocols.*

Results: *Our analysis reveals that SLMs exhibit distinct vulnerability profiles compared to their larger counterparts, with particular susceptibility to prompt injection, jailbreaking, and adversarial attacks. The reduced parameter count often limits the implementation of safety mechanisms, creating unique security challenges.*

Conclusion: *This review provides the first comprehensive taxonomy of SLM-specific vulnerabilities and identifies critical gaps in current research, offering a foundation for future security research in small language models.*

Keywords: Small Language Models, Security Vulnerabilities, Prompt Injection, Jailbreaking, Adversarial Attacks, Systematic Literature Review

I. Introduction

The field of natural language processing has undergone a fundamental transformation with the advent of transformer-based architectures [1] and the subsequent development of Large Language Models. Systems such as GPT [2], Claude, and Gemini have demonstrated remarkable capabilities in reasoning, code generation, and creative tasks, driven by scaling to hundreds of billions of parameters [3, 4]. These foundation models have reshaped expectations for artificial intelligence and attracted substantial industrial investment [5].

Yet a critical countertrend is emerging: the rise of small language models designed for practical deployment and mitigating private data leakage. Small language models, typically defined as transformer-based architectures with around 7-10 billion parameters [6, 7], have emerged as compelling alternatives that address fundamental limitations of their larger counterparts. SLMs require substantially less memory and processing power, enabling deployment on hardware ranging from consumer GPUs to mobile devices and IoT systems, with proportionally reduced operational costs [6, 8]. Edge deployment enables real-time inference without network latency, critical for autonomous systems, industrial automation, and responsive interfaces. Privacy-sensitive applications benefit from local inference, eliminating the need to transmit confidential data to cloud services. Models such as Microsoft's Phi-2 [9], TinyLlama [10], and Google's Gemma [11] have demonstrated that compact architectures can achieve remarkable performance on targeted tasks, challenging assumptions about the necessity of scale. Recent work from NVIDIA researchers argues that small language models represent the future of agentic AI systems, as the specialized and repetitive tasks characteristic of such systems are better suited to smaller, more efficient models [7]. This perspective reflects a broader recognition that practical considerations of cost, latency, and deployment constraints increasingly favor compact alternatives.

However, the rapid adoption of small language models has outpaced systematic investigation of their security properties, creating significant risks for deploying organizations. Zhang et al. found that nearly half of the 63 SLMs they tested exhibited attack success rates exceeding 40% against jailbreak attempts, while over one-third failed to resist even straightforward harmful requests [12]. The fundamental challenge lies in the tension between model capacity and security implementation: safety mechanisms that function effectively in models with hundreds of billions of parameters may degrade or fail entirely when compressed into smaller architectures. Yi et al. documented how compression techniques commonly used to create deployable SLMs can compromise security robustness, revealing "submerged threats" that emerge from efficiency optimizations [13]. Edge deployment introduces additional attack surfaces, as models operating in physically accessible environments face threats ranging from adversarial input injection to direct model extraction. The combination of reduced defensive capacity and expanded attack surface creates a security landscape that demands dedicated investigation, distinct from the extensive but LLM-focused security literature.

II. Theoretical Background

This section will provide foundational concepts necessary for understanding small language model vulnerabilities, including definitions, architectural characteristics, and security frameworks.

II.1 Small Language Models: Definition and Characteristics

II.2 Security Concepts in Language Models

II.3 Attack Taxonomy

III. Methodology

This section will describe the systematic review methodology following PRISMA guidelines.

III.1 Search Strategy

III.2 Inclusion and Exclusion Criteria

III.3 Quality Assessment

III.4 Data Extraction and Synthesis

IV. Results

This section will present the findings of the systematic review organized by thematic categories.

IV.1 Study Selection

IV.2 Vulnerability Categories

IV.3 Mitigation Strategies

V. Discussion

This section will interpret the findings and discuss their implications for research and practice.

V.1 Key Findings

V.2 Implications for Practice

V.3 Limitations

VI. Conclusions

This section will summarize the key contributions and outline future research directions.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2024.

[3] T. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[4] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[6] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, "Small language models: Survey, measurements, and insights," *arXiv preprint arXiv:2409.15790*, 2024.

[7] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, S. Muralidharan, Y. C. Lin, and P. Molchanov, "Small language models are the future of agentic ai," *arXiv preprint arXiv:2506.02153*, 2025.

[8] J. Tang *et al.*, "Demystifying small language models for edge deployment," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[9] M. Javaheripi *et al.*, "Phi-2: The surprising power of small language models," *Microsoft Research Blog*, 2023. Available at: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[10] P. Zhang, G. Zeng, T. Wang, and W. Lu, "Tinyllama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[11] Gemma Team, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[12] W. Zhang, H. Xu, Z. Wang, Z. He, Z. Zhu, and K. Ren, "Can small language models reliably resist jailbreak attacks? a comprehensive evaluation," *arXiv preprint arXiv:2503.06519*, 2025.

[13] S. Yi, T. Cong, X. He, Q. Li, and J. Song, "Beyond the tip of efficiency: Uncovering the submerged threats of jailbreak attacks in small language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv preprint arXiv:2502.19883.