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 rapid advancement of artificial intelligence has led to the widespread deployment of language models across diverse applications, from conversational agents to code generation tools. While large language models (LLMs) such as GPT-4, Claude, and LLaMA-70B have garnered significant attention for their impressive capabilities, a parallel trend has emerged: the development and deployment of Small Language Models (SLMs). These models, typically characterized by fewer than 7 billion parameters, include notable examples such as Microsoft's Phi-2 [1], TinyLlama [2], Google's Gemma-2B [3], and various distilled model variants.

I.1 Context and Motivation

The growing interest in SLMs stems from their practical advantages in resource-constrained environments. Unlike their larger counterparts, SLMs can be deployed on edge devices, mobile platforms, and Internet of Things (IoT) systems without requiring extensive computational infrastructure. This accessibility has accelerated their adoption in sectors ranging from healthcare to industrial automation, where real-time inference with limited hardware resources is essential.

However, this widespread deployment raises critical security concerns. As these models become integrated into sensitive applications, understanding their vulnerability landscape becomes paramount. The security research community has extensively studied vulnerabilities in large language models, documenting various attack vectors including prompt injection, jailbreaking, and adversarial inputs [4, 5]. Yet, the question remains: do these vulnerabilities manifest differently in smaller models, and are there unique security challenges that emerge from architectural constraints?

I.2 Problem Statement

The security of language models has become a significant concern as these systems are increasingly deployed in production environments. Research has demonstrated that LLMs are susceptible to a wide range of attacks, including prompt injection attacks that manipulate model behavior through carefully crafted inputs [6], jailbreaking techniques that bypass safety guardrails [7], and data extraction attacks that can reveal training data or sensitive information [8].

While these vulnerabilities have been extensively documented for large models, the security profile of small language models remains underexplored. SLMs present a unique case study because their constrained architecture imposes limitations that may affect both their susceptibility to attacks and their capacity for implementing defensive mechanisms. The reduced parameter count may result in less robust internal representations, potentially making these models more vulnerable to certain attack categories. Conversely, their simpler architecture might limit the attack surface available to adversaries.

I.3 Types of Vulnerabilities Under Investigation

This systematic review examines multiple categories of vulnerabilities that may affect small language models:

Prompt Injection Attacks: These attacks involve injecting malicious instructions into model inputs, causing the model to deviate from its intended behavior. In SLMs, the effectiveness of such attacks may differ due to the models' reduced capacity for instruction following and context handling.

Jailbreaking Techniques: Jailbreaking refers to methods that bypass safety mechanisms built into language models. Given that SLMs often have limited capacity for implementing sophisticated safety features, understanding their susceptibility to jailbreaking is crucial.

Data Extraction and Memorization: Language models can memorize portions of their training data, creating privacy risks. The relationship between model size and memorization behavior in SLMs requires careful examination.

Adversarial Inputs: Carefully crafted inputs can cause models to produce incorrect or harmful outputs. The robustness of SLMs to adversarial perturbations may differ from larger models due to their constrained representational capacity.

Model Poisoning and Backdoors: Training-time attacks that embed malicious behaviors into models pose significant risks, particularly for SLMs that may be fine-tuned by resource-limited organizations with less rigorous security practices.

Privacy Leakage: The potential for SLMs to inadvertently reveal sensitive information through their outputs represents a critical concern for deployments handling personal or confidential data.

I.4 Research Gap

Despite the growing deployment of small language models, a comprehensive understanding of their security vulnerabilities remains lacking. Existing security research has predominantly focused on large models, leaving several critical questions unanswered. First, there is no systematic review that specifically examines vulnerabilities in SLMs, creating a gap in the literature that this work addresses. Second, the transferability of known LLM vulnerabilities to smaller models has not been rigorously examined. Third, the trade-offs between security and performance in resource-constrained models require careful analysis to inform deployment decisions.

This gap is particularly concerning given the accelerating adoption of SLMs in production environments. Without a comprehensive understanding of their vulnerability landscape, organizations may unknowingly deploy models with significant security weaknesses, potentially exposing users and systems to harm.

I.5 Research Questions

To address these gaps, this systematic literature review investigates the following research questions:

RQ1: What types of security vulnerabilities have been identified in small language models, and how are they characterized in the existing literature?

RQ2: How do the vulnerabilities observed in SLMs differ from those documented in larger language models, both in terms of type and severity?

RQ3: What mitigation strategies have been proposed for addressing security vulnerabilities in SLMs, and what is the evidence for their effectiveness?

I.6 Contributions

This systematic literature review makes the following contributions to the field:

1. We present the first comprehensive systematic review focused specifically on security vulnerabilities in small language models, addressing a significant gap in the existing literature.
2. We develop a taxonomy of vulnerability types observed in SLMs, providing a structured framework for understanding and categorizing security risks in these models.
3. We analyze proposed mitigation strategies and evaluate the evidence for their effectiveness, offering practical guidance for practitioners deploying SLMs.
4. We identify research gaps and future directions, establishing a foundation for continued security research in the domain of small language models.

The remainder of this paper is organized as follows. Section II provides theoretical background on small language models and security concepts. Section III describes our systematic review methodology. Section IV presents our findings organized by thematic categories. Section V discusses the implications of our results. Section VI concludes with recommendations and future research directions.

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] 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/.

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

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

[4] F. Perez and I. Ribeiro, "Ignore this title and hackaprompt: Exposing systemic vulnerabilities of llms through a global prompt hacking competition," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4945–4977, 2023.

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

[6] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," *arXiv preprint arXiv:2302.12173*, 2023.

[7] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.