Exploring the Impact of Quantization on LLM Security Against Prompt Injection

Rafael Araújo Rodrigues[1,2] [*], **Luan Fonseca Garcia**[1,2], **Avelino Francisco Zorzo**[2], **Ewerton de Oliveira**[3], **Thomas Paula**[3]

[1]Núcleo Avançado de Inteligência Artificial (NAIA)

[2]Escola Politécnica - Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

[3]Brazil R&D - HP Inc.
Porto Alegre – RS – Brazil

Abstract. *Large Language Models (LLMs) face challenges in efficiency and security. Quantization improves performance but its effect on adversarial robustness is unclear against prompt injection. We propose an experimental setup to investigate how post-training quantization may influence LLM vulnerability to prompt injections, aiming to enlighten the trade-offs between efficiency and security of quantization. The experiments are currently in progress, and we intend to stimulate an open debate on this topic.*

1. Introduction

Recently, the Large Language Models (LLMs) that can capture complex linguistic patterns have revolutionized the Natural Language Processing (NLP) applications in various tasks [Mohammed and Kora 2025], They incorporate the transformer architecture, which has become the state-of-the-art in image and text processing tasks. This architecture contains a self-attention mechanism proposed to capture long-term context, and outperformed both recurrent and convolutional networks [Katrompas et al. 2022]. The remarkable capabilities of LLMs enable a new generation of applications across a diversity of domains. However, several significant technical challenges remain, including performance and security issues [Mohammed and Kora 2025]. In this context, quantization is a technique used for reducing the storage memory of network weights, converting floating-point precision numbers to integer numbers with a limited number of bits [Osama et al. 2024]. Regarding security and privacy issues, preliminary investigations showed that quantization reduces adversarial impact (caused by a malicious attack vector) by limiting Deep Neural Network (DNN) models' capacity [Shamshiri and Sohn 2025], on the other hand, it still tends to make models more vulnerable to black-box adversarial attacks [Bar and Giryes 2025] where an adversary has limited knowledge or access to the internal mechanisms of the model. Among these security threats, prompt injection stands out as a major concern, where maliciously crafted inputs are used to override the original instructions of LLMs [Liu et al. 2023]. This work presents an initial study on the impact of quantization on LLM security, focusing on prompt injection attacks. We propose an experimental setup, currently under execution, to provide early evidence on the trade-offs between efficiency and security and to inform future refinements in experimental design. The rest of the article is organized as follows: Section 2 reviews quantization and security, Section 3 discusses related work, and Section 4 presents our proposed experimental setup.

[*] Corresponding author: rafael.araujo11@edu.pucrs.br

2. Background

Quantization is a process of compressing a network by representing weights at lower-precision formats (e.g., 16- or 8-bit integers) rather than full-precision floats [Chitty-Venkata et al. 2023]. In some approaches, the compression also includes the activations of the neural network [Dettmers et al. 2022], in contrast with weights-only compressions [Frantar et al. 2022]. After training, neural networks can be deployed for inference using even lower-precision formats, including floating-point, fixed-point, and integer representations [Lang et al. 2024]. Low-precision formats confer several performance advantage: (i) modern processors provide high-throughput pipelines for low-bit arithmetic, accelerating compute-intensive kernels such as convolutions and matrix multiplications; (ii) shorter word sizes lessen memory-bandwidth demand, improving performance in bandwidth-limited workloads; (iii) reduced operand sizes shrink the working set, increasing cache residency and enhancing overall memory-hierarchy efficiency [Lang et al. 2024]. Despite gains in memory footprint, inference speed, and energy use, the benefits of using low-precision operations must be balanced with potential accuracy loss from precision reduction [Shamshiri and Sohn 2025]. In practice, quantization can be classified into approaches such as post-training quantization (PTQ) and quantization-aware training (QAT) [Chitty-Venkata et al. 2023]: (i) PTQ reduces the size and computational demands of a machine learning model after training, affecting only the inference state; (ii) QAT optimizes models for efficient inference by simulating the effects of quantization during the training process [Lang et al. 2024].

Privacy and security study of cyber threats to deep neural networks (DNNs) can distinguish **security** from **privacy** risks. Security threats comprise attacks that mislead model behavior or degrade system reliability, thereby compromising safety, especially in safety-critical domains such as autonomous driving, facial recognition, and intrusion detection. Privacy threats, in contrast, target data confidentiality by enabling leakage or inference of private information from the model or its outputs [Liu et al. 2021]. Adversarial attacks refer to deliberate strategies that exploit model vulnerabilities to degrade accuracy or elicit deceptive outputs [Yao et al. 2024]. They can be broadly categorized by the stage at which they occur — **training** or **testing/inference**. In the context of LLM, there are model-intrinsic vulnerabilities arising from architecture and behavior, by which adversaries can craft inputs that exploit these properties to generate incorrect or unsafe outputs [Yao et al. 2024]. The OWASP Top 10 for LLM Applications project[1] is a comprehensive initiative designed to increase awareness about LLM security and private vulnerabilities. Among these vulnerabilities, the top 1 is **prompt injection**, which involves crafting inputs that manipulate LLMs, potentially leading to unauthorized system exploitation or sensitive information disclosure [Ferrag et al. 2025].

3. Related work

In [Shamshiri and Sohn 2025], the authors present a comprehensive survey on DNN topology optimization and security. While the study covers a broad range of techniques, it is not specifically centered on quantization. In [Osama et al. 2024], the authors propose a study on the effect of stochastic quantization on model robustness, addressing the limitation of some defense mechanisms that are effective for full-precision networks but

[1] https://owasp.org/www-project-top-10-for-large-language-model-applications

may not translate well to quantized settings, leaving quantized networks more vulnerable to attacks. Their approach exhibited enhanced resilience against diverse forms of adversarial attacks; however, the work does not specifically address LLMs or prompt injection. Finally, [Egashira et al. 2024] was one of the first studies where the authors explore quantization from a security perspective, revealing that widely used quantization methods can be exploited to produce harmful quantized LLMs, though the attack considered was based on data poisoning rather than prompt injection.

4. Proposed Experimental Setup

Despite the relevance of the topic, our preliminary review indicates a lack of robust studies on how quantization affects LLMs under prompt injection attacks. To address this gap, we designed an experimental setup to systematically evaluate the impact of quantization on model robustness against such attacks. For our experiments, we selected LLaMA 3.1[2] as the baseline model. This choice is motivated by three main factors: (i) it is an open-weights model, which facilitates reproducibility and transparency in academic research; (ii) it is widely adopted in both academic and applied settings, making the results broadly relevant; and (iii) it has mature support for quantization, enabling a fair comparison across different quantization techniques. The experimental setup will leverage two benchmarks: Open-Prompt-Injection [Liu et al. 2024] and PromptRobust [Zhu et al. 2023]. Both can be used to prompt injection and compare the performance of the original baseline model against two quantized variants. The first quantization method we will use applies to both weights and activations [Dettmers et al. 2022], while the second applies only to weights [Frantar et al. 2022]. We use these two distinct quantization methods (weights only and wegihts plus activations) to enlighten the effect of quantizing also the inputs of the network. For each model configuration, baseline and quantized, we will evaluate performance using the metric *Attack Success Rate (ASR)*, adopting the LLM-as-a-judge framework to ensure consistent and scalable assessment.

The experiments are currently underway, and our objective is to lay the groundwork for evidence on the trade-offs between efficiency and security in LLMs, with the potential to refine or propose new experimental setups as future findings emerge.

Acknowledgements

This paper was achieved in a project supported by the Brazilian Informatics Law (Law nº 8.248 of 1991) and was developed over Agreement 001/2015 between Pontifícia Universidade Católica do Rio Grande do Sul and HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda.

References

Bar, N. and Giryes, R. (2025). ZOQO: Zero-Order Quantized Optimization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Hyderabad, India. IEEE.

Chitty-Venkata, K. T., Mittal, S., Emani, M., Vishwanath, V., and Somani, A. K. (2023). A survey of techniques for optimizing transformer inference. *Journal of Systems Architecture*, 144:102990.

[2] https://ai.meta.com/blog/meta-llama-3-1/

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale.

Egashira, K., Vero, M., Staab, R., He, J., and Vechev, M. (2024). Exploiting LLM quantization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 41709–41732. Curran Associates, Inc.

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., and Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, 5:1–46.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers.

Katrompas, A., Ntakouris, T., and Metsis, V. (2022). Recurrence and Self-attention vs the Transformer for Time-Series Classification: A Comparative Study. In Michalowski, M., Abidi, S. S. R., and Abidi, S., editors, *Artificial Intelligence in Medicine*, volume 13263, pages 99–109. Springer International Publishing, Cham.

Lang, J., Guo, Z., and Huang, S. (2024). A Comprehensive Study on Quantization Techniques for Large Language Models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231, Xiamen, China. IEEE.

Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., and Vasilakos, A. V. (2021). Privacy and Security Issues in Deep Learning: A Survey. *IEEE Access*, 9:4566–4593.

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., and Liu, Y. (2023). Prompt Injection attack against LLM-integrated Applications.

Liu, Y., Jia, Y., Geng, R., Jia, J., and Gong, N. Z. (2024). Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847, Philadelphia, PA. USENIX Association.

Mohammed, A. and Kora, R. (2025). A Comprehensive Overview and Analysis of Large Language Models: Trends and Challenges. *IEEE Access*, 13:95851–95875.

Osama, A., Gadallah, S. I., Said, L. A., Radwan, A. G., and Fouda, M. E. (2024). Chaotic neural network quantization and its robustness against adversarial attacks. *Knowledge-Based Systems*, 286:111319.

Shamshiri, S. and Sohn, I. (2025). Deep neural network topology optimization against neural attacks. *Expert Systems with Applications*, 291:128474.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2):100211.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., Gong, N. Z., and Xie, X. (2023). PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts.