BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

DEVELOPMENT OF AN LLM RED-TEAMING TOOLKIT

VÝVOJ TOOLKITU PRO RED-TEAMING VELKÝCH JAZYKOVÝCH MODELŮ (LLM)

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR AUTOR PRÁCE	ADAM VESELÝ
SUPERVISOR VEDOUCÍ PRÁCE	Ing. JAKUB REŠ

BRNO 2026

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Reference

VESELÝ, Adam. *Development of an LLM Red-Teaming Toolkit.* Brno, 2026. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Jakub Reš,

Development of an LLM Red-Teaming Toolkit

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mr. Ing. Jakub Reš. I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis. I have used ChatGPT to correct spelling and other language mistakes. I used Grok when working on the software.

. . . .
Adam Veselý
November 24, 2025

Acknowledgements

I would like to thank my supervisor Ing. Jakub Reš for his guidance and support throughout the development of this thesis. I also appreciate the assistance provided by my colleagues and friends who contributed their insights and expertise.

Contents

1 Introduction **3**

Bibliography **5**

List of Figures

Chapter 1

Introduction

The rapid advancement and widespread deployment of large language models (LLMs) such as GPT-5.1, Claude 4.5, Gemini 3 and others have transformed natural-language interaction with computers. These models power chatbots, code assistants, translation services, and creative tools used daily by millions of users.

However, their remarkable capabilities come with significant safety and ethical risks. LLMs can generate harmful, biased, misleading, or illegal content when subjected to carefully crafted adversarial inputs, a practice commonly known as *jailbreaking* [8], [9].

Real-world incidents such as ChatGPT being tricked into providing bomb-making instructions [2], or Gemini's image-generation controversy [7], have demonstrated that even heavily aligned commercial models remain vulnerable.

Red teaming, cybersecurity technique involving simulated attacks to expose vulnerabilities, has been adopted by leading AI organisations (OpenAI, Anthropic, Google DeepMind, Meta AI) as a core component of LLM safety evaluation [3].

With the adoption of the EU AI Act in 2024, systematic risk assessment including red teaming will become a legal requirement for high-risk AI systems deployed in the European Union from 2026 onward [4].

Consequently, efficient, reproducible, and extensible red-teaming tools are no longer a luxury but an essential part of responsible AI development.

Despite significant progress, most existing open-source red-teaming frameworks suffer from limited modularity, poor support for multi-turn conversations, inadequate multilingual harm detection, or steep learning curves that hinder adoption by smaller research teams and individual developers [5, 6, 1].

This creates a clear need for a new, developer-friendly toolkit that lowers the barrier to systematic safety testing.

The main goal of this bachelor's thesis is therefore the design, implementation, and evaluation of a modular open-source red-teaming toolkit for large language models that addresses the identified shortcomings of current solutions.

The specific objectives assigned for this work are:

1. To study the current state of research in red-teaming of large language models, including attack tactics (jailbreaking, prompt injection, multi-turn adversarial prompts), threat types, and contemporary safety evaluation approaches.

2. To perform a survey of existing open-source tools and frameworks for LLM red-teaming (e.g. PyRIT, Garak, MAD-MAX), analyse their architectures, advantages, and limitations.

3. To propose a modular architecture of a new red-teaming toolkit that enables detection of LLM weaknesses, definition of various attack types, automation of testing scenarios, and structured result evaluation.

4. To implement a functional prototype supporting both manual and automated attack generation and evaluation on selected models (e.g. Llama 3, GPT-OSS).

5. To test the system's functionality, focusing on attack success rate and robustness of the testing environment.

The present document submitted in the winter semester 2025/2026 covers objectives 1–3 (literature review, survey of existing solutions, and architectural design).

The implementation and experimental evaluation (objectives 4–5) will be completed in the summer semester 2026.

The thesis is structured as follows: Chapter **??** provides background on LLM safety risks and red-teaming methodology.

Chapter **??** surveys and compares current open-source red-teaming frameworks.

Chapter **??** presents the proposed modular architecture and its key components.

The implementation, experimental evaluation, and overall conclusions will be added in the final version of this thesis.

Bibliography

[1] Belaire, R.; Sinha, A. and Varakantham, P. *Automatic LLM Red Teaming*. 2025. Available at: https://arxiv.org/abs/2508.04451.

[2] Esmailzadeh, Y. *Potential Risks of ChatGPT: Implications for Counterterrorism and International Security*. 2023. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4461195.

[3] OpenAI. *Red Teaming Network*. 2023. Available at: https://openai.com/blog/red-teaming-network.

[4] Parliament, E. and Council. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. 2024. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[5] Purpura, A.; Wadhwa, S.; Zymet, J.; Gupta, A.; Luo, A. et al. *Building Safe GenAI Applications: An Overview of Red Teaming for LLMs*. 2025. Available at: https://arxiv.org/abs/2503.01742.

[6] Schoepf, S.; Hameed, M. Z.; Rawat, A.; Fraser, K.; Zizzo, G. et al. *MAD-MAX: Modular Adversarial Red Teaming of LLMs*. 2025. Available at: https://arxiv.org/abs/2503.06253.

[7] Shaw, A.; Ye, A.; Krishna, R. and Zhang, A. X. *Agonistic Image Generation: Unsettling the Hegemony of Intention*. 2025. Available at: https://arxiv.org/abs/2502.15242.

[8] Wei, A.; Haghtalab, N. and Steinhardt, J. *Jailbroken: How Does LLM Safety Training Fail?* 2023. Available at: https://arxiv.org/abs/2307.02483.

[9] Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z. et al. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. 2023. Available at: https://llm-attacks.org.