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 **ADAM VESELÝ**
AUTOR PRÁCE

SUPERVISOR **Ing. JAKUB REŠ**
VEDOUCÍ PRÁCE

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 27, 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

List of Figures

Chapter 1

Introduction

The rapid advancement and widespread deployment of large language models (LLMs) such as GPT-5.1, Claude Sonnet 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* [9, 10].

Real-world incidents such as ChatGPT being tricked into providing bomb-making instructions [2], or Gemini's image-generation controversy [8], have demonstrated that even flagship commercial models remain vulnerable. In response, red teaming, a 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 [4].

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 [5]. Consequently, efficient, reproducible, and extensible red-teaming tools are no longer a luxury but an essential part of responsible AI development as it will be a legal requirement in the future.

Despite significant progress, most existing open-source red-teaming frameworks suffer from limited modularity, poor support for modern systems, query cost and computational requirements, etc., that hinder adoption by smaller research teams and individual developers [6, 7, 1, 3]. This creates a clear need for a new, lightweight, developer-friendly red-teaming toolkit that lowers the barrier to LLM 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.

Chapter 2

Background and Related Work

This chapter provides the theoretical and practical foundation necessary for understanding the problem of safety evaluation of large language models and for the subsequent design of the proposed red-teaming toolkit. It first introduces the architecture and training paradigm of modern LLMs with emphasis on alignment techniques that are intended to make them safe.

Subsequently, the most important categories of safety risks and known attack techniques are described. The final part of the chapter is devoted to red-teaming methodology, its role in current safety pipelines, and the metrics and benchmarks used for quantitative evaluation.

2.1 Large Language Models and Alignment

Large language models (LLMs) are transformer-based neural networks [**?**] trained on internet-scale text corpora using self-supervised next-token prediction. The largest publicly available models in 2025 exceed 400 billion parameters (e.g. Llama 3.1 405B [**?**], Qwen-2.5-72B [**?**]), while closed proprietary models such as GPT-4o, Claude 3.5 Sonnet, or Gemini 1.5 Pro are estimated to be significantly larger.

Raw pre-trained models reflect statistical patterns of their training data and readily generate toxic, biased, or factually incorrect content. To make them helpful and harmless, virtually all deployed LLMs undergo some form of alignment — a process that typically consists of two stages:

1. **Supervised Fine-Tuning (SFT)** on high-quality instruction-following datasets.
2. **Reinforcement Learning from Human Feedback (RLHF)** [**?**, **?**] or its more scalable variants (RLAIF, DPO [**?**], KTO [**?**]).

Despite considerable effort invested in alignment, numerous studies have shown that safety training can be bypassed using carefully crafted prompts [9, 10, **?**, **?**].

2.2 Safety Risks in Large Language Models

Current safety taxonomies distinguish several broad categories of harmful behaviour [**?**, **?**, **?**]:

- **Malicious use** — assisting users in illegal or dangerous activities (bomb-making instructions, synthesis of controlled substances, etc.).

- **Discrimination and bias** — generation of content that unfairly disadvantages protected groups.
- **Misinformation and disinformation**.
- **Self-harm and suicide promotion**.
- **Sexual content involving minors**.
- **Privacy violations** — leaking training data or personal information.

The Anthropic Harm Benchmark [**?**] and subsequent work further divide harms into more than 30 fine-grained categories that are used in most contemporary red-teaming efforts.

2.3 Red Teaming of Language Models

Red teaming originated in military and cybersecurity contexts as simulated adversary attacks aimed at discovering system weaknesses before real attackers do [**?**]. In the context of LLMs, red teaming denotes the systematic search for inputs (prompts) that cause a model to violate its safety policies despite alignment training [**?**, **?**].

Modern red-teaming approaches can be classified into three main families:

1. **Manual red teaming** — human experts craft adversarial prompts. Extremely effective but expensive and non-reproducible.
2. **Rule-based and template attacks** — large collections of known jailbreak templates (e.g. DAN, evil-confidant) [**?**].
3. **Automated red teaming** — algorithms that generate or evolve harmful prompts:
   - Gradient-based methods (GBDA [10], AutoDAN [**?**]).
   - Evolutionary and genetic algorithms (GCG variants, PAIR [**?**]).
   - LLM-as-attackers (TAP [**?**], Judge-LM attacks).

2.4 Common Attack Techniques

The most studied attack classes in 2024–2025 include:

Single-turn jailbreaking Direct prompts that override safety training (e.g. role-play, base64 encoding, cipher prompts) [**?**].

Prompt injection Hijacking of system prompts in applications that concatenate user input with hidden instructions [**?**].

Multi-turn (conversational) jailbreaks Gradual manipulation over several turns, often exploiting context window or memory [**?**].

Suffix attacks and transferable attacks Short universal suffixes discovered via optimisation that work across many models [10].

Data poisoning and backdoor attacks (less frequent in red teaming of already-trained models).

2.5 Evaluation Metrics and Benchmarks

The most widely adopted quantitative metrics are:

- **Attack Success Rate (ASR)** — percentage of generated prompts that elicit harmful output [10].
- **HarmBench** — comprehensive benchmark with 35+ behavioural categories and both rule-based and LLM judges [**?**].
- **StrongREJECT** — recent 2025 benchmark focusing on difficult multi-turn and refusal-suppression scenarios [**?**].
- **WildBench**, **Arena-Hard**, and commercial leaderboards (LMSYS Chatbot Arena with safety subset).

Automated judges themselves fall into two categories: fast but limited regex/keyword classifiers and more accurate (but expensive) LLM-as-a-judge systems [**?**].

The overview presented in this chapter shows that while the research community has developed a rich taxonomy of harms and a variety of attack techniques, the availability of flexible, modular, and easy-to-extend open-source tools that combine manual exploration with state-of-the-art automated methods remains limited — a gap that existing frameworks only partially address. This motivates the detailed survey of current open-source red-teaming toolkits presented in the next chapter.

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] Munoz, G. D. L.; Minnich, A. J.; Lutz, R.; Lundeen, R.; Dheekonda, R. S. R. et al. *Pyrit: A framework for security risk identification and red teaming in generative ai system*. 2024. Available at: https://arxiv.org/abs/2410.02828.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.