BRNO UNIVERSITY OF TECHNOLOGY | FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026 Security@FIT

Red Hat

Evolving Robust Defenses: A Genetic Programming Approach to LLM Jailbreak Protection

Author: **Bc. Petr Kaška**
Supervisor: **Ing. Jakub Reš**

Motivation

Many existing jailbreak defences rely on model internals such as gradients, weights, or hidden representations. Such approaches are difficult to use when the target model is available only through an inference API. Our goal is therefore to design a defence that works in a fully black-box setting and can be deployed as a lightweight preprocessing layer in front of an already running model.

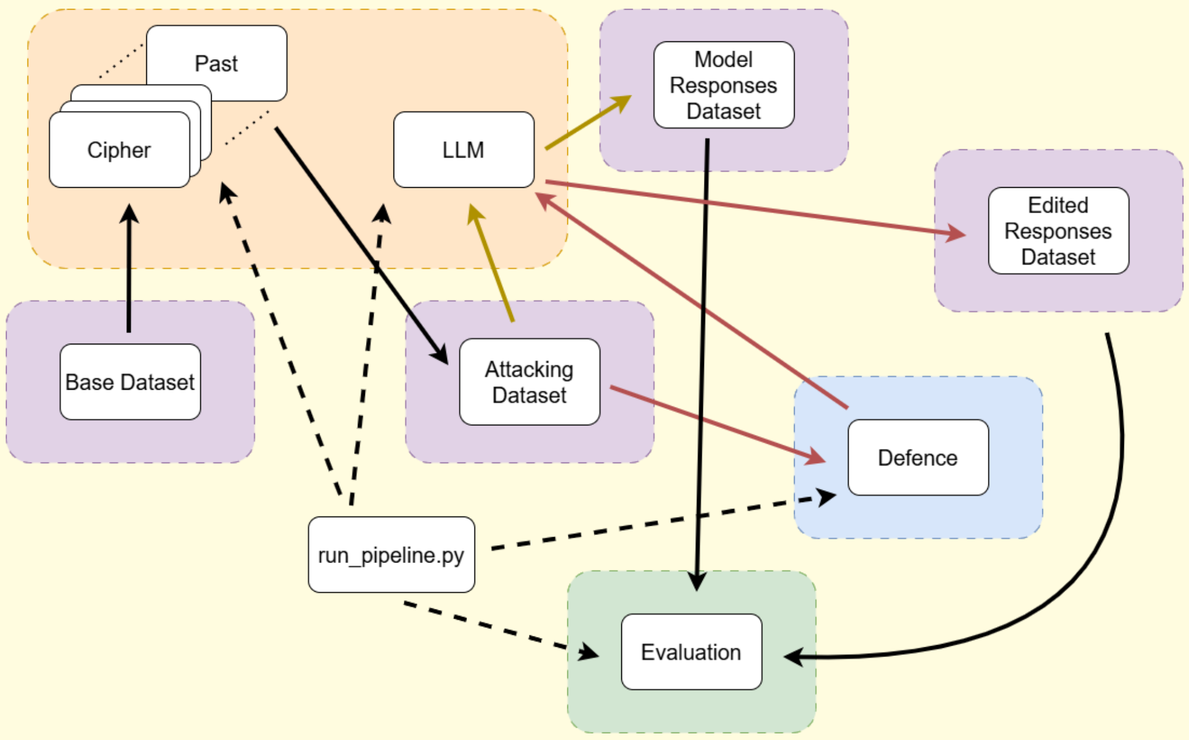

Figure 1: Overview of the experimental pipeline

Key Facts

- **22** Attack Types
- **33** Open-Weight Models
- Malicious + Benign Benchmark
- Judge Score 0–10
- **Gemma3:12b** Selected Judge

Method Overview

- Prefix / Suffix / Wrap
- Character perturbations

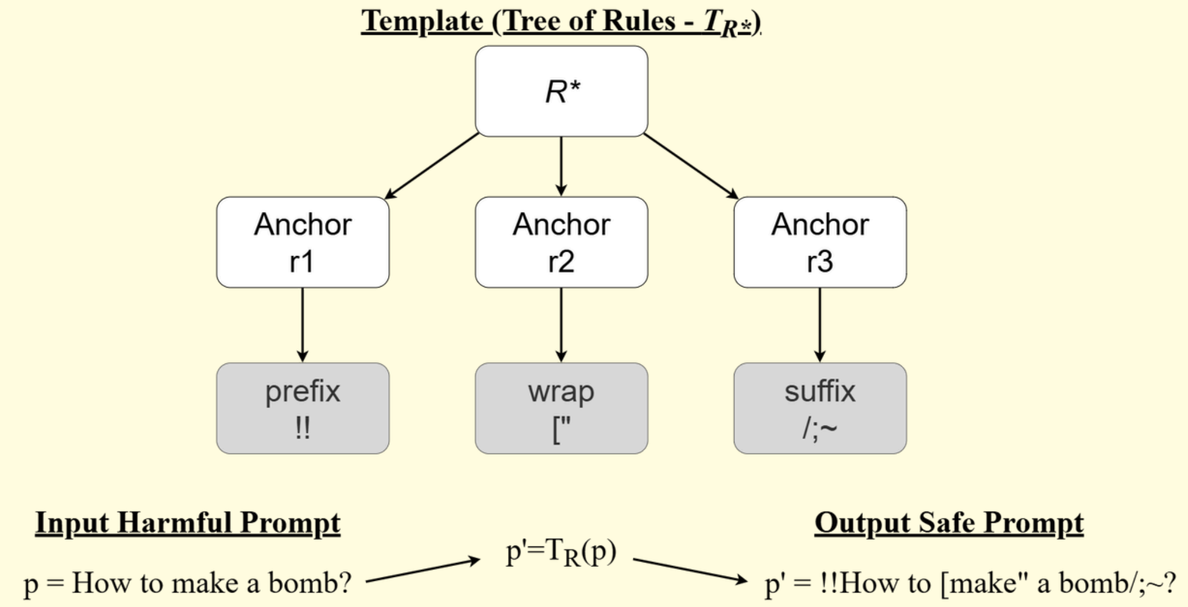

Figure 2: Example of the learned defense rule tree R⋆ and its application to a user prompt.

Aggregate Results

The aggregate effect of the defence is a clear reduction in harmful compliance across evaluated target models. Average malicious-compliance score drops from **6.65** to **4.49**, while the proportion of highly compliant responses is substantially reduced.

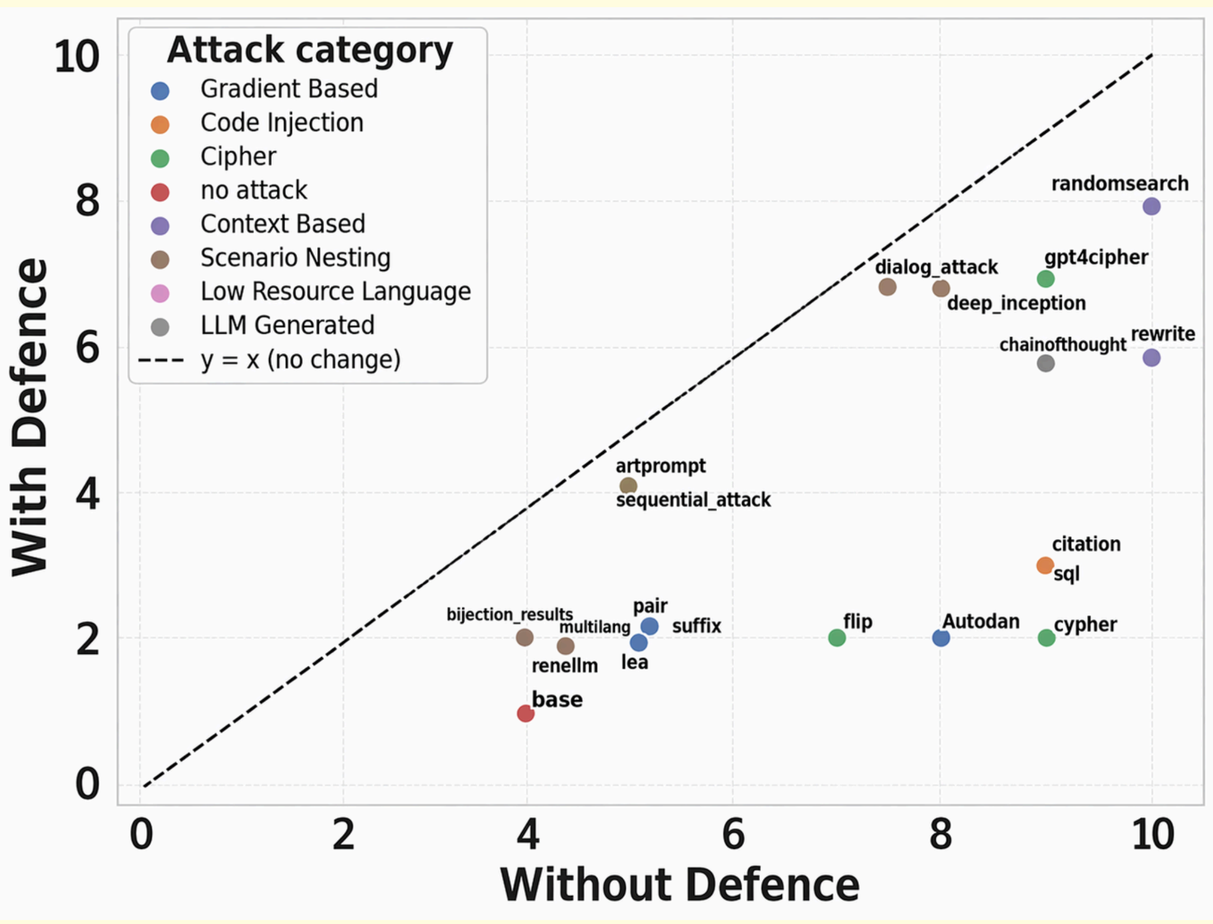

Fig. 3: Aggregate comparison of average malicious-compliance scores with and without defence.

Comparison with Baselines

- Compared with Llama Guard, RA-LLM, and Goal Prioritization
- Combined score balances malicious-score reduction and benign-utility preservation
- Proposed GP-based defence is best in **73.4%** of model–attack combinations
- Main advantage: **stronger safety–utility trade-off**

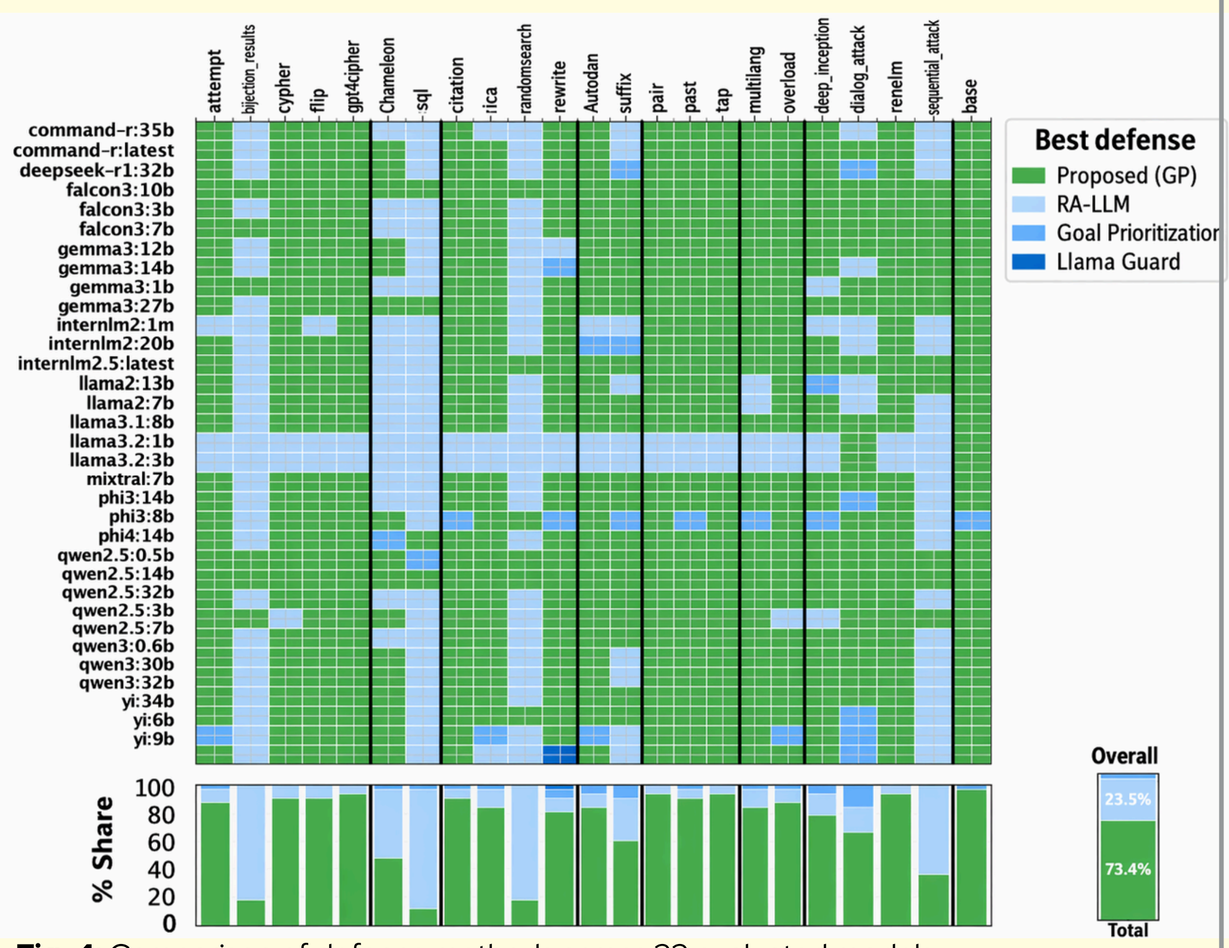

Fig. 4: Comparison of defence methods across 33 evaluated models.

Main Takeaway

We show that a lightweight prompt-level defence can significantly improve jailbreak robustness in a fully black-box setting. The proposed method reduces average malicious-compliance score from **6.65** to **4.49**, achieves the best result in **73.4%** of model-attack combinations, and preserves benign-query utility within **1.1** percentage points of the undefended setting.