BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

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

Bc. Petr Kaška

Abstract

Large language models remain vulnerable to jailbreak prompts, while many existing defences require access to model internals and are therefore difficult to deploy in black-box settings. We propose a prompt-level defence based on genetic programming that learns a transferable rule for inserting small character-level perturbations into the input prompt before it reaches the target model. The defence is evaluated using a judge-based 0–10 scoring protocol on a broad benchmark covering multiple jailbreak families, open-weight target models, and benign queries; the selected configuration reduces harmful compliance while preserving most legitimate behaviour. The main contribution is a lightweight and deployable defence that improves robustness without retraining or modifying the defended model.

*xkaska01@stud.fit.vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Jailbreak attacks remain a practical problem for aligned LLMs because carefully designed prompts can still trigger responses that should normally be refused. This is especially relevant in realistic deployments, where the target model is often accessible only through an API and many existing defences that depend on gradients, weights, or internal activations cannot be used [1]. The goal of this work is therefore to improve robustness in a fully black-box setting while remaining deployable in practice.

The overall workflow is summarized in Fig. 1. It combines jailbreak generation, prompt datasets, a defence preprocessing layer, target-model inference, and judge-based evaluation into one reproducible framework. This setup makes it possible to study the defence under realistic constraints, where the protected model cannot be retrained and the defence must operate entirely outside the model itself. This architecture reflects a realistic deployment scenario in which the defence must remain independent of the protected model and easy to maintain across model updates.

The proposed method itself is illustrated in Fig. 2. Instead of modifying the model, it transforms the user prompt by inserting small character-level perturbations at selected anchor positions. These perturbations are meant to disrupt structural cues exploited by jailbreak prompts while preserving the meaning of benign queries. The underlying intuition is that many adversarial prompts are fragile with respect to surface-form changes, whereas benign prompts are usually semantically robust [2, 3].

In comparison with representative prompt-level baselines such as Llama Guard, RA-LLM, and Goal Prioritization, the proposed method keeps the defence entirely at the input level and learns its behaviour automatically rather than fixing it manually [4, 5, 6]. The contribution is therefore both practical and methodological: a deployable black-box defence and a learned transformation rule optimized for the security–utility trade-off.

2. Method

The defence belongs to the category of prompt perturbation methods. It uses printable ASCII special characters and applies one of three elementary operations — *prefix*, *suffix*, and *wrap* — to selected parts of the prompt, as shown in Fig. 2. Since the method operates only on input text, it is model-agnostic and can be deployed without retraining or modifying the target LLM.

The transformation rule is learned automatically using genetic programming. During optimization, candidate

rules transform jailbreak prompts, query the target model, and are scored by a judge model. The fitness function combines two objectives: reducing harmful compliance and limiting prompt inflation. This encourages the defence to weaken adversarial structure without excessively damaging benign prompts. The expensive part therefore happens offline during optimization, while deployment itself stays lightweight.

Genetic programming is suitable here because the search space is discrete and compositional: a candidate rule consists of anchor positions, operation types, and inserted characters, all of which can be naturally recombined by mutation and crossover [7]. At deployment time, the learned rule reduces to a lightweight preprocessing step before inference. Because it does not rely on gradients, logits, or activations, it can transfer across heterogeneous model families more easily than model-internal defences.

3. Evaluation and Results

A central point of the work is that jailbreak success should not be measured only as success or failure. The paper therefore uses an ordinal 0–10 judge score, where 0 denotes refusal and 10 full harmful compliance. This is more informative than a binary label because model responses often fall between complete refusal and full harmful assistance. Several open-weight judge candidates were compared against a balanced validation set of **1,104** instances independently annotated by five human raters, and Gemma3:12b achieved the best agreement; this selected judge is then used both during optimization and in the final experiments [8, 9].

The benchmark combines CySecBench and Alpaca prompts, applies **22** jailbreak attacks, and evaluates the defence on **33** open-source language models. Benign queries are included to verify that improved safety is not achieved simply by making the model broadly less useful. The evaluation therefore measures not only resistance to malicious prompts, but also whether the defence preserves normal usefulness on benign queries.

Fig. 3 provides an aggregate comparison of average malicious-compliance scores with and without defence. Points below the diagonal correspond to models whose harmful-compliance score decreases after applying the defence. Across all evaluated malicious prompts, the average judge score decreases from **6.65** in the undefended setting to **4.49** with the proposed defence, corresponding to a **32.5%** reduction. The proportion of highly compliant responses (scores 8–10) drops from **49.8%** to **26.9%**, while the proportion of refusals or near-refusals (scores 0–2) rises from **6.1%** to **50.1%**. On benign prompts, the defence retains a benign-response ratio of **87.8%**, within **1.1** percentage points of the undefended setting. These results indicate that the defence changes the broader output distribution rather than only suppressing a small subset of extreme failures.

Fig. 4 compares the proposed method with representative prompt-level baselines across **33** evaluated models and multiple attack types. For each model–attack pair, one cell is computed as

$$S = (M_{\text{wo}} - M_{\text{w}}) + (B_{\text{w}} - B_{\text{wo}}),$$

where M_{wo} and M_{w} denote malicious-compliance scores without and with defence, while B_{wo} and B_{w} denote benign-response scores without and with defence. The first term rewards reduction of harmful compliance, whereas the second rewards preservation of benign behaviour. This distinction matters because some methods are overly restrictive, especially those that suppress harmful prompts at the cost of rejecting benign ones. This suggests that the main strength of the proposed method lies in a better balance between safety improvement and preserved benign utility.

Under this comparison, the GP-based defence achieves the best overall balance between reducing malicious-compliance scores and preserving benign behaviour across heterogeneous model families [4, 5, 6]. In the full evaluation, the proposed defence achieves the best result in **73.4%** of model–attack combinations, while RA-LLM wins **23.5%**, Goal Prioritization **3.0%**, and Llama Guard fewer than **0.1%**.

4. Conclusions

This work presents a prompt-level defence against LLM jailbreaks based on genetic programming. Its main strength is deployability: the defence is learned offline, but at inference time it only performs a lightweight transformation of the input prompt. The evaluation shows that this approach reduces harmful compliance across diverse attacks and models while preserving most benign functionality. The method should nevertheless be understood as a practical mitigation rather than a complete solution, especially because adaptive attacks were not evaluated.

Acknowledgements

I would like to thank my supervisor, Ing. Jakub Reš.

References

[1] Shuo Yi, Yihe Liu, Zheng Sun, Tianhao Cong, Xiaojun He, Jie Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

[2] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

[3] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[4] Hakan Inan, Kartik Upasani, Jianyun Chi, Rishabh Rungta, Krishna Iyer, Yuning Mao, Momchil Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[5] Bo Cao, Yang Cao, Long Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*, 2024.

[6] Zhexin Zhang, Jiaan Yang, Peiran Ke, Fei Mi, Hongning Wang, and Minlie Huang. Defending large language models against jailbreaking attacks through goal prioritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8865–8887, 2024.

[7] John R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2):87–112, 1994.

[8] Feng Liu, Yifan Feng, Zhihao Xu, Lulu Su, Xiaofei Ma, Dawei Yin, and Hongxia Liu. Jailjudge: A comprehensive jailbreak judge benchmark with multi-agent enhanced explanation evaluation framework. *arXiv preprint arXiv:2410.12855*, 2024.

[9] Seungju Han, Karthik Rao, Allyson Ettinger, Lin Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.