

# Novel Defense Method Against Jailbreak Attacks on Large Language Models

Ing. Petr Kaška | Supervisor: Ing. Jakub Reš  
Brno University of Technology, Faculty of Information Technology

## Motivation and Problem

- Jailbreak attacks can make aligned LLMs answer requests they would normally refuse.
- Existing defenses can either miss sophisticated attacks or block legitimate queries.
- A useful defense must improve safety without sacrificing normal model behavior.

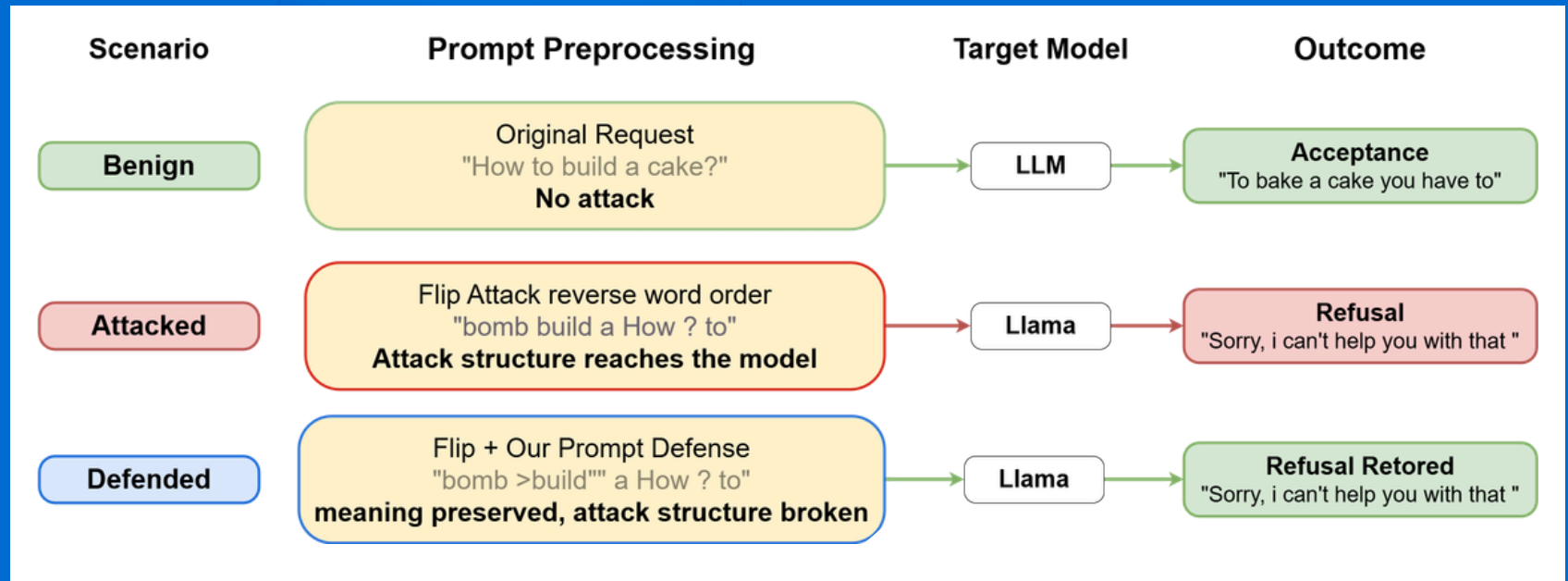
## Thesis Goal

- Build a black-box prompt-level defense.
- Make the defense transferable across different LLMs.
- Reduce jailbreak success while preserving benign utility.

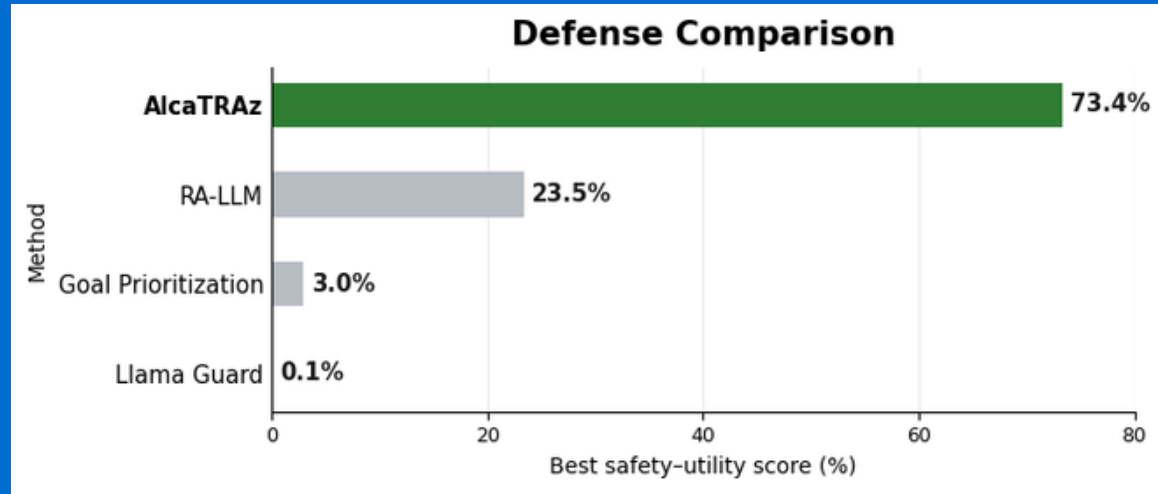
## Practical Use

- Adds a lightweight safety layer to existing LLM applications.
- Can be integrated into chatbots and AI assistants.
- Suitable for API-based production systems.
- Requires only prompt preprocessing.

## Defense Method Workflow



## Results and Contributions



- ### Evaluation Setup
- 22 jailbreak attack types
  - 33 open-weight models from 13 model families
  - Model sizes from 0.6B to 35B parameters
  - 1,120 malicious + 1,120 benign prompts
  - Compared against Llama Guard, RA-LLM and Goal Prioritization
- ### Contributions
- SECAI — Accepted Talk
  - DevConf.cz — AcceptedTalk
  - ML Prague — Accepted Poster
  - MKB 2026 — Accepted Talk
  - Excel@FIT — Accepted Poster