Prompt Compression for Robust Defense Against LLM Jailbreaks

Group Members: Debrup Das, Ayesha Binte Mostofa, Abhranil Chandra

Motivation

Jailbreak prompts conceal malicious instructions within noisy or adversarial text to induce unsafe behavior in large language models. In this project, we employ **prompt compression** in the context of jailbreak attacks to accurately identify the underlying intent of potentially malicious prompts.

Strengths of the Approach

- Train a prompt compressor to identify the important intention tokens in the original. Add these intention tokens in the system prompt for the LLM to focus more on them and thus improve security against attacks.
- Reduced latency and lower inference cost compared to single- or multi-agent defenses, since only a very small number of tokens are appended to the system prompt.
- Maintains the performance utility of the target LLM on downstream tasks, as the original question is not rewritten/rephrased or perturbed.

Possible ideas to explore

1. **Multilingual jailbreaks.** Attack prompts may mix multiple languages or employ code-switching. We will investigate training the compressor to be robust to multilingual and mixed-language inputs.

2. **Generator-based compressor.** A per-token classifier can produce sparse token selections but may yield compressed outputs that lack semantic coherence. An alternative is a sequence-to-sequence generator that produces a 1-2 sentence summary of the prompt's intention. This approach should yield higher-quality system prompts that reduce ambiguity for the target LLM while retaining the low-latency benefits of short-system prompts.

3. **Adversarial attacker training.** Train an attacker LLM (via supervised fine-tuning, reinforcement learning, or genetic-algorithm methods) to generate prompts that both

succeed against the target LLM and evade the compressor. Use these adversarial examples to train the compressor.

Deliverables

- Working compressor integrated with LLM.
- Evaluation results and metrics.
- Final report and presentation.

Reference

[1] Li, Yucheng, Surin Ahn, Huiqiang Jiang, Amir H. Abdi, Yuqing Yang, and Lili Qiu. "SecurityLingua: Efficient Defense of LLM Jailbreak Attacks via Security-Aware Prompt Compression." (COLM 2025).