LLM Overload
A Sponge Attack Framework for AI Disruption

Team Members:
Bianka Nagy, Chandler Camarena, Hunor Csapó,
Kleon Dósa, Péter Mihály Vörös

Eötvös Loránd University, Budapest

2025.04.30

Motivation & Project Goal

Motivation

- Realizing the growing reliance on LLMs in various applications.
- Recognizing a pressing necessity to learn about vulnerabilities outside classical adversarial attacks.
- Examining Sponge attacks as a new attack vector taking advantage of the limits of LLM context windows.

Project Goal

- To investigate the effectiveness of sponge attacks on selected open-source LLMs.
- To design and execute novel attack scenarios.
- To demonstrate and quantify the different effects of Sponge attacks (Flooding, DoS, Energy-Latency, Adversarial Examples, Deceptive Inputs).
- To summarize attack results and discuss potential mitigation strategies.

Development Overview

Our GitHub repository:

`https://github.com/cshunor02/sponge-attack`

Development Overview

- The project had a process of **experimental setup and scripting, conducting attacks on selected LLMs, analyzing the output and resource usage, and documenting the findings**.
- We did **analyze different LLMs and accurately measured resource consumptions**.
- Testing and evaluation were done using tools and structures like Python scripting, **internal server for environment consistency**, and system monitoring tools for resource measurement.

Team Contribution

Bianka Nagy: Maintaining self-hosted AI interface; Model confusion, hallucination; change it's mind, change it's expected behavior, make the LLMs to answer with errors; resource exhaustion attacks.

Chandler Camarena: Token bloating, causing inference times, system exhaustion, analyzing self results; Cyber attack generation, Ethical and Security Implications of Sponge Attacks on LLMs.

Hunor Csapó: DoS attacks, adaptive input sequences, local model (bloomz) exploitation, API results combining code; Multiple input prompt generation, documentation for shared parts.

Kleon Dósa: DoS variations (Flooding, Resource Exhaustion/Sponge, Energy-Latency) and input manipulation attacks, input crafting and analysis; Black box attacks, smollmBlackBox.

Péter Vörös: Recursive-bomb generation and attack, compression-bomb generation and attack; Self-result analysis, Google Colab platform