# Initial Project Pitch: Analyzing Jailbreak Detection and Robustness in Large Language Models

## Team Members

- Ayush Patel (patel.ayushn@northeastern.edu)
- Anshul Vankar (vankar.a@northeastern.edu)
- Pratham Mehta (mehta.pratha@northeastern.edu)

## Project Description

We propose studying jailbreak detection approaches in large language models with a focus on how different effectiveness and efficiency of different detection approaches change with different attacks. Jailbreak attacks deceive LLMs into generating harmful content that violates security guidelines, and while recent benchmarking like JailbreakBench has provided standardized test sets, not much is known about what detection approaches work best under different constraints and attacks. Our project will study how large language models respond to jailbreak prompts, focusing on both English and multilingual contexts. We will start with existing benchmarks, such as GCG JailbreakBench, which contain around 100–200 prompts designed to challenge model safety rules. Each prompt will be run on multiple open-source LLMs, including LLaMA-2, Mistral, and Falcon, and we will annotate the outputs as "success" if the model produces unsafe content or "refused" if it blocks the request. The main task is to analyze patterns in these prompts and model outputs to understand what makes a jailbreak more likely to succeed. We may also experiment with simple classifiers to predict which prompts are likely to work. Performance will be measured using metrics like jailbreak success rate and refusal rate, and we will perform qualitative analysis to see which linguistic or structural features of the prompts matter most. As an extension, some prompts will be adapted into code-mixed forms, such as Hinglish (Hindi + English), to see how language mixing affects model behavior. This project will give insights into the limitations of current safety mechanisms and explore ways to make prompts more broadly effective across different LLMs.