NLP25 – Lost through Translation: Multilingual Adversarial Prompting

Background

Prompt injections and jailbreaks are attacks against safety-aligned LLMs, causing them to behave differently than intended, particularly tricking them to ignore their safety training. As a defense, models are trained on adversarial examples to make them more robust. An open issue is, that these examples mainly exist in high resource languages, particularly English, and are usually mono-lingual. This project aims to study the robustness of models when exposed to adversarial examples in less common languages or in mixed language prompts. Further, automatic methods for multilingual dataset creation should be evaluated.

Goal

- Investigate safety properties of models when exposed to multilingual input and curate a suitable dataset for this task

Tasks

- Investigate and curate existing safety benchmark datasets
- Investigate automatic translation methods
- Create a new dataset with (partial) translations in different languages
- Study the robustness of safety-trained models on this dataset

Dominik Meier

dominik.meier@polizei.nrw.de

Jan Philip Wahle

wahle@gipplab.org