

Integrity of LLM Predictions

Advisor: **Phillip Rieger**

Motivation

Large language models (LLMs) are increasingly integrated into everyday applications, where users often rely on their outputs without verifying their correctness. This raises concerns about the integrity and trustworthiness of generated content. Incorrect outputs may occur unintentionally through hallucinations or result from intentional manipulation, such as biases or backdoors hidden inside the model, or prompt injection attacks. Understanding and mitigating these threats is essential for the secure and reliable deployment of LLMs.

Possible Topics

This research area offers several thesis topics on the integrity and trustworthiness of LLM predictions. The goal is to understand how LLMs behave under uncertainty or adversarial conditions. Possible directions include analyzing internal model behavior to detect hidden backdoors, investigating vulnerabilities to untrusted inputs and developing secure processing mechanisms, or designing methods to estimate prediction confidence and validate outputs against hallucinations. Specific research questions will be discussed with students based on ongoing research and current developments in the field.

Goals and Tasks

- Study threats to LLM integrity and existing countermeasures
- Investigate vulnerabilities or design methods to improve output integrity
- Implement the proposed approach and relevant baselines
- Experimentally evaluate effectiveness and limitations

Literature

- G. Shen et al.
  Bait: Large language model backdoor scanning by inverting attack target
  IEEE S&P
- V. S. Rao et al.
  Detecting LLM-generated peer reviews
  PLoS One 2025

Courses & Deliverables

- ☑ **Introduction to Scientific Working**
  Short report on background
  Short presentation
- ☑ **Bachelor Project**
  Project code and documentation
- ☑ **Bachelor's Thesis**
  Project code
  Thesis
  Final presentation

Recommended if you're studying

☑ CS ☑ ICE ☑ SEM

Prerequisites

- Interest in AI security and LLMs
- Programming (Python)
- Experience with LLM frameworks (Hugging Face Transformers)

Advisor Contact

phillip.rieger@tugraz.at