

Monitoring LLM Behavior

Advisor: **Phillip Rieger**

Motivation

Large language models (LLMs) are increasingly deployed in applications ranging from customer support chatbots to autonomous AI agents. However, these systems may be abused for tasks beyond their intended purpose or manipulated into violating their instructions. In addition, especially in multi-agent systems, interactions between agents can lead to unsafe collective behavior, even when individual agents are instructed to follow safety guidelines. Recent incidences demonstrate the need for mechanisms to monitor LLM behavior and detect deviations from intended functionality.

Possible Topics

This research area offers several thesis topics on monitoring and controlling LLM behavior. The goal is to understand how different tasks and activities are reflected in the models' internal representations (hidden states) and how these can be used to detect unintended or unsafe behavior.

Possible directions include investigating how hidden states can reveal whether LLMs process unauthorized topics, analyzing how social dynamics and peer pressure affect safety alignment in multi-agent systems, or developing monitoring mechanisms that detect violations of behavioral constraints from internal model activations. Specific research questions will be discussed with students based on ongoing research and current developments in the field.

Goals and Tasks

- Study LLM monitoring, safety alignment, and existing safeguards
- Design methods to analyze or improve compliance with behavioral constraints
- Implement monitoring mechanisms or controlled multi-agent environments
- Evaluate robustness under adversarial and unintended usage scenarios

Literature

- H. Xu et al.
  When Agents Go Rogue: Activation-Based Detection of Malicious Behaviors in Multi-Agent Systems
  ICML
- W. Gurnee et al.
  Verbalizable representations form a global workspace in language models
  arXiv preprint arXiv:2607.15495 2026

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