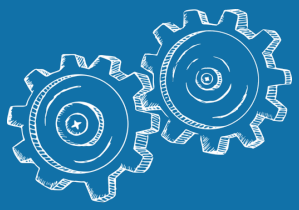

Agentic Games

Advisor: **Martin Wistauder**

Motivation

Small & Large Language Models perform well for tasks involving natural language. The ability to call tools further enhances Language Model capabilities and their performance for specific tasks. However, the invocation of tools (or arbitrary code) also broadens the attack surface of LLM-powered applications.

Is it a good idea to let users invoke python code, fetch information from the internet, or upload files? What are the security implications? What defenses might be needed?

Goals and Tasks

- Get familiar with the capabilities and limits of LLMs (see promptingguide.ai).
- Implement an agentic LLM application (with e.g. python sandbox, web requests, file upload, memories, referencing other conversations).
- Evaluate the added attack surface for each tool and propose defenses against known attacks (see embracethered.com, kai-greshake.de).
- Create small games with different levels of security to teach in a fun way how to attack and defend LLM-powered applications.

Literature

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

- Information Security LV recommended
- Programming (Python)

Advisor Contact

martin.wistauder@tugraz.at