Learning Symbolic Models of Agentic AIs

Advisor: **Matthias Grill**

Motivation

LLM-based agents increasingly act autonomously by invoking tools, running commands, navigating the web, and performing security-relevant tasks. However, they are typically evaluated using task success rates and execution logs, which provide limited insight into their strategies and safety.

Recent work (ATLAS) shows that interpretable behavioral models can be learned from agent interaction traces by combining LLM-guided abstraction with automata learning. This thesis extends these descriptive models to Markov decision processes (MDPs), enabling formal verification of agent behavior and the synthesis of runtime monitors or shields that detect or prevent unsafe actions.

Goals and Tasks

- Study LLM agents, stochastic automata learning (IOAlergia/AALpy), and probabilistic model checking
- Collect agent traces on an existing benchmark and extend LLM-guided abstraction to learn MDP models
- Verify safety properties using Storm and synthesize a runtime monitor or shield

Literature

- I. D. Lopez-Miguel et al.
  ATLAS: Discovering Agent Strategies through LLM-Guided Abstraction and Automata Learning
- M. Tappler et al.
  Learning environment models with continuous stochastic dynamics-with an application to deep RL testing

Courses & Deliverables

☑ **Master Project**
Project code
Report
Presentation

– OR –

☑ **Master's Thesis**
Initial presentation
Project code
Thesis (60+ pages)
Final presentation

Recommended if you're studying

☑ CS ☑ ICE ☑ SEM

Prerequisites

- Interest in formal methods and trustworthy AI
- Programming (Python)
- Basic knowledge of machine learning

Advisor Contact

matthias.grill@tugraz.at