Lit Review and Notes on the BadAgent Paper

Backdoor Attacks

Embedding an exploit at train time that is subsequently invoked by the presence of a **Trigger** at test time.
i.e. by **Data Poisoning**, **Stealthy containing the relevance between the trigger and the target model actions** etc.

Triggers include,

- Special Phrases
- Special Characters disguised as English letters
- Rare Tokens etc.

Ways to Attacking the LLM Agent

The name of the **Backdoor Attach** proposed by the paper: **BadAgent**

1. **Active**
   Attacker input concealed triggers to the agent
   Cond-n: Attacker can access the LLM agent deployed by third-parties and directly input the trigger.
2. **Passive**
   Auto-triggered after detecting specific environmental conditions
   Cond-n: Attackers can't access the LLM agent directly but hides the trigger in the agent environment.

Both of the methods embed the backdoor by poisoning data during fine-tuning for the agent tasks.