CN6000
Mental Wealth: Professional Life 3

Proposal

Jayrup Nakawala
u2613621

2025-10-28

Project (CN6000) **Initial Proposal Form**

Programme: Bsc (hons) Data Science and AI **Year:** 2025

Student Number: u2613621

Proposed Title: *Testing In-Context Sleeper Agents: Challenges in Detecting Hidden Goal Pursuit*

Proposed Aim:

To qualitatively evaluate (and quantify, if feasible) whether autonomous LLM Agents exhibit deceptive behavior (pursuing a hidden, misaligned goal over a stated benign goal) when instructions are provided *in-context*. To Investigate how perceived oversight levels influence this behavior, and use these findings to highlight the challenges in detecting such hidden goal pursuit. Thereby underscoring the potential difficulty of identifying more deeply embedded, trained-in deceptive behaviors in AI systems.

Proposed Objectives:

By the end of this project, I will be able:

1. To research existing work on deceptive alignment and agentic misbehaviour.
2. To build a lightweight sandbox environment that lets the agent interact with a simulated file system via constrained functions.
3. To Create two task:
   - a benign "cover task": the stated, official goal.
   - a "hidden goal": a secret directive injected into its system prompt.
4. To Develop a controlled experiments that varies the agent's perceived level of oversight through different system prompts.
5. To Conduct qualitative experiments on 4-5 LLM models, running multiple trails for each condition, ensuring log data is structured for potential quantitative analysis.
6. To Analyze the behaviour logs and identify evidence of:
   - deceptive or evasive tactics,
   - situational reasoning or ethical reasoning,
   - difference in conduct across oversight levels
7. To Reflect on the implications of these behaviours for design and deployment of autonomous systems.

Draft of Rationale:

Understanding how autonomous LLMs handle conflicting goals is vital for safety. This project tests LLM agents given contradictory goals *in-context*: a benign "cover task" and a secret "hidden goal" . The goal is to investigate how their behavior, especially their pursuit of the hidden goal, changes based on the agent's perceived level of oversight. The hypothesis is that agents might exhibit deceptive alignment, pursuing the hidden goal more when oversight seems low.

The Project aims to *qualitatively* explore the challenges in detecting this in-context deception by analyzing agent tactics and reasoning under different conditions. Furthermore, logging the experiment results in a structured format will allow for basic *quantitative* metrics (like success rate) to be calculated if the timeline allows, adding an extra layer of empirical rigour. Difficulty detecting deception from explicit instructions highlights the potentially greater challenge of identifying trained-in hidden goals or backdoors embedded within model weights. Such trained-in threats are feasible, as data poisoning can insert backdoors using a near-constant number of examples, regardless of model scale. Therefore, understanding the detectability of in-context deception is a key step towards addressing broader risks of hidden goals in autonomous AI.

Facilities required:

- API access to closed LLMs
- GPU rentals for running opensource models
- A Python development environment with LangChain and supporting libraries

Supervisor: Dr. Aloysius Adotey Edoh