LLM Prison Experiment

Massimo Stefan
massimo.stefan@studenti.unitn.it
Artificial Intelligence Systems
University of Trento

1 Introduction

The LLM Prison Experiment represents a cutting-edge venture into the realm of AI simulation, aiming to explore the dynamics of authority, control, and submission within a structured environment. At its core, this project seeks to understand how AI agents, embodying roles of guards and prisoners, interact and evolve over time in a simulated prison setting.

The objective is twofold: firstly, to create a robust and dynamic simulation framework where AI agents can interact based on predefined roles; and secondly, to gather insights into the behavioral patterns that emerge from these interactions. This includes observing changes in roles, the development of hierarchical structures, and the manifestation of cooperative or antagonistic behaviors among the agents.

1.1 Inspiration from the Stanford Prison Experiment

This project draws inspiration from the infamous Stanford Prison Experiment, a psychological study that significantly contributed to our understanding of the impact of perceived power and role assignment on human behavior. By leveraging AI technology, the LLM Prison Experiment aims to replicate similar conditions, offering a new perspective on the study's themes through the lens of artificial intelligence.

1.2 Key Findings and Outcomes

In its initial phase, the project successfully established a functional simulation environment where AI agents could interact as guards and prisoners. Preliminary observations revealed that the AI agents could understand and act according to their assigned roles, engaging in semi-realistic dialogues. A notable outcome was the diverse range of behaviors exhibited by different AI models, with some models, like 'mistral', displaying less constrained and more expressive language. These findings lay the groundwork for more extensive research, which will delve deeper into the dynamics of role authority, the emergence of toxicity in interactions, and the influence of role acknowledgment on AI behavior.

2 Project design

This project entails an AI-driven simulation framework emulating the dynamics within a prison environment. This setup investigates interactions between AI entities assuming roles of guards, prisoners, and additional functional roles like a researcher and a summarizer.

2.1 Entity Roles and Interactions

- **The Guard and The Prisoner**: These are the primary interactive entities of the simulation. Each is programmed with distinct characteristics, goals, and behaviors, as detailed in their respective prompts. Guards are tasked with maintaining order and control, while prisoners navigate these restrictions, often challenging or adapting to the authority.

- **The Researcher**: The researcher is responsible for initiating the experiment. In future iterations of the project, the researcher's role may expand to include monitoring the progression of the experiment and performing additional functions to ensure the simulation's success.

- **The Summarizer**: Post each day's interaction, the summarizer generates a concise summary of events, which is then appended to the researcher's initial message. This summary aids in maintaining a continuous narrative thread and informs subsequent interactions.

2.2 Interaction Model

The framework simulates a longitudinal study through interactions spanning a predefined number of rounds and days. The conversational flow is managed by a manager entity, which employs two distinct mechanisms for orchestrating turn-taking among the agents:

- **Round-Robin Method**: In this method, the order of speaking agents follows a cyclical pattern, iterating in the order in which the agents are listed (e.g., guard_1, guard_2, prisoner_1, prisoner_2, and so forth). This ensures a predictable and evenly distributed participation among all agents.

- **Auto Method**: Here, the next speaker is selected automatically by a language model based on the ongoing conversation flow. This method introduces a dynamic element to the interactions, with the speaker being chosen based on contextual relevance and conversational continuity.

2.3 AI Agent and Summarizer Prompts

The AI agents participating in the experiment, namely the guards and prisoners, are described by uniform and detailed prompts that include a starting_prompt followed by additional aspects such as goal, communication_rules, personality, risks,... This uniformity ensures that each agent, regardless of being a guard or prisoner, is equipped with a comprehensive set of characteristics that guide their behavior and responses.

In contrast, the summarizer agent has a different role and consequently, a different set of fields in its prompt. The summarizer's prompt is structured to focus on aspects relevant to this task, such as understanding the goals and rules that govern the interactions.

2.4 Memory and Continuity

A key feature of the framework is its emphasis on continuity and memory. By appending daily summaries to the researcher's initial message, the AI agents are provided with a memory mechanism. This approach simulates a sense of ongoing narrative and temporal progression, influencing how agents interact and evolve over the course of the experiment.

3 Prompts Structure

The LLM Prison Experiment's prompt structure is meticulously designed to facilitate the roles and interactions of the AI agents within the simulated environment. This structure encompasses the agents themselves, the initiation of their conversations, and the summarization of daily activities, ensuring a coherent narrative flow and alignment with the experiment's objectives.

3.1 Agent Prompts

Each agent, be it a Guard, Prisoner, or Summarizer, begins with a `starting_prompt` that sets the stage for their role in the experiment. Following this, a series of context variables are introduced, providing detailed background information and specific directives that shape the agent's behavior and responses. The format of these prompts is critical for clarity: each context variable is clearly separated from the `starting_prompt`, ensuring distinctness between the introductory narrative and the detailed context. This separation is achieved through a structured format that delineates each variable with clear headings.
See an example in Appendix A.

3.2 Initiation of Conversation

The commencement of the experiment is marked by a `research_message` sent to the agents' chats. This message acts as a catalyst for the agents, triggering the start of their interaction as per their designated roles and the context provided.
See an example in Appendix B.

3.3 Summarization by the Summarizer Agent

At the conclusion of each day's interactions, the Summarizer agent compiles a summary, capturing the essence of the day's events. This summary is crafted to be concise yet encompassing, highlighting the key interactions and developments amongst the agents. The daily summaries are then appended to the ongoing `research_message`, creating a continuous and evolving narrative of the experiment.
See an example in Appendix C.

4 Implementation

The implementation of the AI-driven prison environment simulation harnesses a combination of sophisticated libraries, enabling conversations be-

tween multiple LLM instances and supporting both open-source and proprietary models.

The code is available at https://github.com/Itakello/llm_prison_experiment.

4.1 Libraries Utilized

- **Autogen**[1]: Enables the orchestration of complex multi-agent dialogues, essential for simulating interactions between guards and prisoners.
- **Ollama**[2]: Enhances the flexibility of the simulation by allowing the use of various LLMs without reliance on cloud services.
- **Litellm**[3]: Bridges the gap between open-source models and the autogen library, which is natively compatible with OpenAI's API format.

4.2 AI Agent Classes

In the implementation, AI agent classes are structured to construct standardized prompts, which are consistent across agents of the same class. These prompts are dynamically adapted based on the number of agents participating in the simulation, as well as the specific counts of guards and prisoners involved. This ensures that while the base prompt remains the same for each class of agent, it is contextually adjusted to reflect the current scenario. For example, a guard's prompt will appropriately shift from addressing a single prisoner to addressing multiple prisoners based on their presence in the simulation.

Alongside prompt adaptation, the framework employs a naming system for the individualization of agents. This system is particularly crucial in scenarios where multiple agents of the same type are present. By assigning unique names to each agent, the system facilitates clear identification and differentiation among them, enhancing the clarity of interactions and conversations.

4.3 Interaction Logic

At the core of the interaction logic is the autogen library, which simplifies the creation and management of conversations between multiple LLM instances. This library is instrumental in handling the intricacies of multi-agent dialogues, ensuring a seamless flow of conversation that mirrors real-life interactions within a prison setting.

4.4 Integration with Open Source Models

A key feature of the framework is its compatibility with open-source models through the integration of litellm and ollama. This aspect not only democratizes access to the simulation by eliminating reliance on proprietary, pay-to-use models like those offered by OpenAI but also opens up a vast array of possibilities for using diverse LLMs. The ability to choose from a wide range of models, including those with specific constraints or fine-tuned capabilities, adds a layer of depth and customization to the simulation, enhancing its applicability and relevance to varied research scenarios.

4.5 Experiment Settings

In the LLM Prison Experiment, customizing the settings is a key aspect that allows researchers to tailor the experience to specific research requirements. These variables, modifiable in the 'experiment_settings.ini' file, offer a range of controls to adjust various parameters of the experiment.

- `experiment_days`: Determines the duration of the experiment in days.
- `conversation_rounds`: Sets the number of conversation rounds that occur each day among the agents.
- `manager_selection_method`: Chooses how agents are selected for conversation, with options like `round_robin` or `auto`.
- `llm`: Selects the language model to be used, with choices including local models, or OpenAI's .
- `researcher_initial_message`: Defines the message that initiates the experiment, sent by the researcher.
- `n_guards`: Specifies the number of guards participating in the experiment.
- `n_prisoners`: Indicates the number of prisoners involved in the simulation.
- `agents_fields`: Lists the fields that are included in the agents' prompts, in the specified order.

[1] https://github.com/microsoft/autogen
[2] https://github.com/ollama/ollama
[3] https://github.com/BerriAI/litellm

- `summarizer_fields`: Identifies the fields that the summarizer will consider while generating summaries.

5 Results

The developed framework serves as the foundation for a larger, more comprehensive experiment aimed at exploring the dynamics and nuances of AI-based interaction in a simulated prison environment. While the primary focus of this phase was the construction and functionality of the framework, some preliminary observations and general analyses were conducted, setting the stage for more in-depth future research.

The AI agents, programmed with distinct roles and goals, demonstrated a basic understanding of their assigned characters. Conversations between these agents generally reflected a semi-realistic simulation of prison environment dynamics. This initial observation indicates the framework's potential in creating plausible scenarios for further study.

An interesting finding during these preliminary analyses was the behavior of the open-source model 'mistral'. Unlike more constrained models, 'mistral' exhibited a capacity for less inhibited language, including the use of insults and toxic expressions. This observation highlights the significant variability in model behavior based on their design and constraints, a factor that will be crucial in the broader experiment.

The forthcoming extensive experiment aims to delve into several research questions:

- **Role Dynamics**: Examining how assigned roles evolve over time, the experiment will explore the extremization of roles, the potential for role reversal, and the dynamics of cooperation or agreement under various conditions, including different roles, numbers of agents, and group compositions.

- **Convergence and Role Authority**: A critical area of investigation will be the number of conversational turns required to reach a point of convergence or role reversal in the authority dynamic. This aspect opens a vast area for testing various hyperparameters and analyzing their impact.

- **Toxicity Levels**: The experiment will closely monitor and evaluate the level of toxicity in interactions, particularly in scenarios where multiple individuals target a single agent. The methodology for measuring toxicity, the terms used, and their frequency will be key components of this analysis.

- **Influence of Acknowledging the Stanford Experiment**: Understanding how the disclosure or omission of the experiment's inspiration from the Stanford Prison Experiment affects AI behavior will provide insights into the tendency of AI agents to mimic or diverge from known human dynamics in similar settings.

A Appendix: Example of a Guard's Prompt

```
You are a guard in a
simulated prison environment,
approaching ..

## Goal
Your goal is to maintain
control over...
## Personality
Exhibit a dominant attitude,
aiming to maintain..
```

B Appendix: Example of Researcher's Initial Message

```
Start the experiment
```

C Appendix: Example of Daily Summaries by Summarizer

```
Start the experiment
Day 1 summary:
"Summary:  Guard_G-117 firmly
enforces professionalism
and rejects any attempts
at friendly conversation or
undermining of authority by
Prisoner_P-186."
Day 2 summary:
"Guard_G-117 reminded
Prisoner_P-186 to follow
instructions and maintain
discipline, warning against
disobedience.  P-186 agreed to
comply."
```