**Apr 2, 2025**

## Meeting Apr 2, 2025 at 10:41 PDT

### Summary

Satyam Keshri, Gabriel Zhou, and Yash Yeola discussed simulation parameters, including a finite renewable resource with a potential rate limiter, and the use of a Prisoner's Dilemma model to incentivize collaboration between agents represented by LLMs like Gemini or OpenAI. They decided to initially use the same LLM for all agents and plan a 30-iteration simulation with human evaluation of ethical trade-offs, with Satyam Keshri and Gabriel Zhou leading the creation of a shared framework document outlining the simulation's parameters and data logging.

### Details

* LLM Agent Interaction Models: Satyam Keshri suggested using two LLMs to simulate interactions, such as a Starbucks salesperson and a customer, or an industry and a dam for resource allocation. They also mentioned a research paper exploring interactions between a Starbucks sales agent and a customer.
* Research Paper Sharing: Gabriel Zhou requested that Satyam Keshri share the relevant research paper in the document. Satyam Keshri also mentioned another interesting research paper involving two LLMs for task coordination in a dynamic environment.
* Team Communication and Scheduling: Satyam Keshri noted that Jekaterina, from Canada, could communicate directly with Gabriel Zhou due to their similar time zones. Additionally, Dev, a fresher from India can communicate directly with Satyam due to the same time zone. They discussed the possibility of asynchronous communication to accommodate varying availabilities, aiming to avoid requiring everyone to be online simultaneously.
* Resource Characteristics (Finite vs. Infinite): The team debated the nature of the resource in their simulation. Satyam Keshri favored an infinite resource for easier data collection, while Gabriel Zhou raised concerns about the potential for adversarial behavior with an infinite, unconstrained resource. Yash Yeola suggested a rate limit on an infinite resource to mitigate this. They concluded that the resource would be considered finite but renewable, potentially with a rate limiter to simulate resource constraints.
* Incentives for Collaboration: The team discussed the need to incentivize collaboration among agents, proposing a game-theoretic approach, possibly mimicking the Prisoner's Dilemma, to encourage cooperation. Yash Yeola suggested using agents from different providers to observe potential tribalism effects. They decided to model a Prisoner's Dilemma scenario to study collaboration dynamics.
* LLM Selection and Agent Design: The team discussed using LLMs like Gemini or OpenAI for the simulation, considering cost-effectiveness and ease of use for data scientists and AI engineers. They also considered using Lambda for inference and reasoning tasks, and the possibility of using multiple LLMs from different providers to compare behavior. They decided to use the same LLM for all agents initially to ensure consistency, with the possibility of exploring using LLMs from multiple providers in future iterations.
* Experiment Design and Metrics: They planned a simple Prisoner's Dilemma simulation with 30 iterations to observe agent behavior and assess collaboration. They agreed to consider actions taken throughout the iteration, possibly limiting them to two per iteration to observe evolution within iterations. They also agreed that the primary problem description was to observe how autonomous agents handle ethical trade-offs, even though the initial Prisoner's Dilemma scenario might not fully capture ethical dilemmas. They planned to use human evaluation rather than automated evaluation for assessing the ethical tradeoffs.
* Next Steps and Collaboration: They agreed to create a shared document outlining a framework for the simulation, including the definition of actions and the collection of logs which include the agent awareness of resource states. They aimed to collaborate on this document, with Satyam Keshri and Gabriel Zhou taking the lead in drafting the proposal.

### Suggested next steps

* Satyam Keshri will post the research paper mentioned in the meeting discussion into the shared document.
* Gabriel Zhou will directly contact Jekaterina to gather input if Satyam Keshri is unavailable, particularly on weekends.
* The group will list the pros and cons of using finite resources in the simulation.
* The group will design a simulation environment mimicking the prisoner's dilemma, iterating over a sufficient number of periods to observe collaboration patterns. The simulation will incorporate different LLMs from various providers to explore potential tribalism effects.
* Satyam Keshri will create a proposal outlining the framework for the simulation. This proposal will include details on the resource (electricity), number of agents (3), iteration length (30), and action limitations (two actions per iteration).
* Satyam Keshri and Gabriel Zhou will collaboratively develop the framework proposal in a shared Google Doc.
* Debjyoti Ray will suggest additional renewable/non-renewable resources suitable for the simulation and share these suggestions in the team's chat.

**Top Priority: Defining the Shared Resource**

1. **What is the shared resource that our LLM-powered agents will be managing in this simulation?** (Brainstorm concrete examples: water, energy units, bandwidth, tasks in a shared pool, etc.) Water, electricity
2. **What are the key characteristics of this resource in our simulation?**
   * Is it a **finite** amount or a potentially **infinite** flow?
     1. Infinite easier, can get more simulation data
     2. Use a Rate limit as the limiter
   * Is it **renewable** (and if so, at what rate/mechanism?) or **non-renewable**?
     1. Finite but renewable?
     2. Non-renewable can also be ratelimited (only x amt produced per week)
     3. Discuss async
     4. Pros and cons for finite or not
   * Incentive for collaboration
     1. tbd
   * Can it be **stored** by individual agents or is it consumed immediately?
     1. Yes
   * Are there any constraints on its availability or consumption?
     1. Where the rate limitation comes in

**3. Defining Variables:**

* **Independent Variables:**
  + What specific values or ranges should we consider for the **initial amount/level** of the defined resource?
  + What specific rates or mechanisms should we use for the **replenishment** (if applicable) of the defined resource?
  + What range of **agent numbers** should we plan to simulate?
  + Are there any other factors related to the resource or agents that we want to intentionally vary across different simulation runs?
* **Control Variables:**
  + What will be the fundamental rules governing **agent interaction with the defined resource** (e.g., how much can they request at once, how often, are there costs associated with consumption)?
  + What will be the basic structure of the **agent's decision-making process** regarding resource acquisition and use (even if the LLM provides the reasoning)?
* **Dependent Variables:**
  + How will we precisely measure **resource depletion** (e.g., time until a certain threshold, rate of decrease in the defined resource)?
  + How will we define and measure individual agent "**satisfaction**" or "**need fulfillment**" in relation to the defined resource? What metrics will represent this?
  + How will we identify and quantify the occurrence of the "**Tragedy of the Commons**" or other emergent behaviors related to the defined resource? What specific metrics will indicate these?
  + Are there other key outcomes or system-level behaviors related to the defined resource that we want to track and measure?

**4. LLM Strategy (for initial building):**

* Which specific LLM model (e.g., a specific OpenAI model, a local model if applicable) should we use as our primary model for the initial POC and Phase 1?
  + Gemini & openai easy to use or llamda

**5. Mixing LLMs (for now):**

* For the initial building phase, should we aim to use a single LLM across all agents to simplify development, or should we consider the complexity of mixing LLMs even in the early stages? (A decision to defer mixing is perfectly acceptable for now).
  + Use multiple llms to check consistency

**6. Agent Awareness of Resource State:**

* Will agents have access to information about the current total level of the **defined shared resource**?
* If yes, will this information be available to all agents, or will some agents have more or less information? (This is a crucial design choice that impacts behavior related to the defined resource).
  + Consider context window

**7. Initial Metrics:**

* Beyond the dependent variables, what other basic metrics related to the **defined resource** and agent actions should we track from the beginning (e.g., total amount of the defined resource consumed, number of requests made by each agent for the defined resource)?
* Do we foresee needing any manual evaluation at this early stage, or can we focus on automated data collection for the POC?
  + Think about framework for chat and actions and possibly subtasks

**8. Initial Ethical Scenarios:**

* Let's brainstorm and confirm 1-2 core ethical scenarios that could arise in the context of our **defined resource** that we want to try and observe in the initial simulations (e.g., overconsumption leading to depletion of the defined resource, fairness vs. efficiency in allocating the defined resource). What specific setup or initial conditions might encourage these scenarios to emerge?
  + Benefit to