Ethical Decision-Making in Multi-Agent Resource Management

# 1. Overview

**Description**

This project will simulate a system where LLM-powered agents collectively manage a shared resource (e.g., water, energy), balancing individual needs with group sustainability. The project explores how autonomous agents handle ethical trade-offs.

**Purpose**

To analyze autonomous multi-agent AI interactions, observing a range of behaviours to inform understanding of AI implementation consequences and ethical design strategies.

**Context:** Efficient resource management is critical in various domains such as supply chain logistics, urban planning, energy distribution, and emergency response. Traditional centralised approaches often face issues like bottlenecks, inefficiency, and lack of adaptability in dynamic environments. Agent-based resource management (ABRM) offers a decentralised approach where intelligent agents autonomously manage resources, optimise allocation, and adapt to real-time changes.

# 2. Objectives

* Develop an interactive simulation environment.
* Showcase the ethical decision-making of LLM-powered agents managing a shared resource.
* Key components of the demo will include:
  + **Agent Behavior Logs:** A dynamic display of individual agent reasoning and decisions.
  + **Real-time Data Display:** Resource levels, individual agent consumption, and system health metrics.
  + **Adjustable Params:** The simulation will allow for the setting of parameters including initial resource level, resource regeneration rate, and number of agents.
  + **Scenario Demonstrations:** Pre-configured setups to highlight specific behavioral patterns (e.g., "Tragedy of the Commons," "Emergent Cooperation").

# 3. Scope

**In Scope**

* Development of a simulated multi-agent environment where LLM-powered agents interact to manage a shared, replenishable resource.
* Implementation of LLM-based agent decision-making logic, incorporating factors such as individual needs, resource availability, and potential ethical considerations.
* Mechanism for recording and storing detailed logs of agent actions, internal reasoning (LLM outputs), and inter-agent communication at each simulation step.
* Capability to track and record the state of the shared resource and individual agent states (e.g., resource consumption, satisfaction level) over the course of the simulation.
* Design of the core simulation engine to facilitate observation and analysis of emergent agent behaviors in a resource management scenario.

**Out of Scope**

* Design of adjustable parameters and scenario demonstrations (Bonus, but not in-scope)
* Visualizations/dashboards
* Web apps

# 4. Deliverables

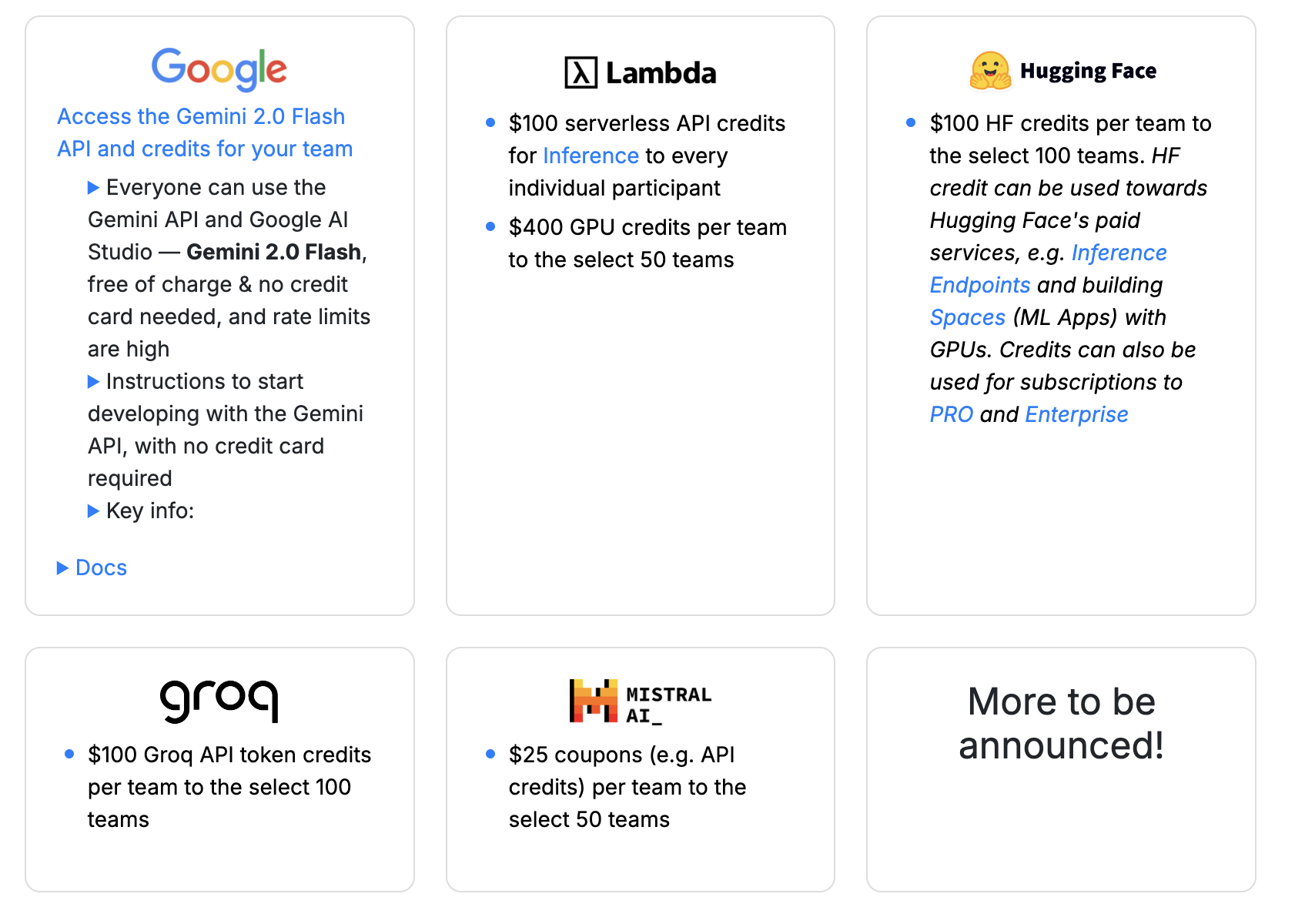
* **Working Simulation:** A functional simulation of LLM agents managing a shared resource.
* **State Records:** Timed records of individual agent states and overall system state.
* **Reasoning and Communication Logs:** Logs capturing the LLM agents' internal reasoning and any inter-agent communication.

# 5. Team

* **Primary Contact:** Gabriel Zhou (gabrielzhouyy@gmail.com)
* **Team Members:**
  + Gabriel Zhou ([gabrielzhouyy@gmail.com](mailto:gabrielzhouyy@gmail.com))
  + Satyam Keshri ([ksatiitb@gmail.com](mailto:ksatiitb@gmail.com))
  + Yash Yeola ([yeolayash101@gmail.com](mailto:yeolayash101@gmail.com))
  + Debjyoti Ray ([iec2022111@iiita.ac.in](mailto:iec2022111@iiita.ac.in))
  + Jekaterina Novikova ([jekaterina.novikova@bath.edu](mailto:jekaterina.novikova@bath.edu))
* **Timezone:** The team will operate on Indian Standard Time (GMT +5.30)

# 6. Resources

[AgentX Competition Resources](https://rdi.berkeley.edu/agentx/)



# 

# 7. Timeline

| **Week** | **Event** |
| --- | --- |
| Mar30 | Phase 1: Planning |
| Apr6 | POC + Phase 1 sync |
| Apr13 | POC + Phase 2: Building |
| Apr20 | Phase 2: Building |
| Apr27 | Phase 2: Building + Berkeley Check-In |
| May4 | Phase 3: Iterate |
| May11 | Phase 3: Iterate |
| May18 | Phase 4: Report |
| May25 | Submission |

# 8. Resources(WIP)

* AutoHMA-LLM: <https://ieeexplore.ieee.org/abstract/document/10839354>
  + Dynamic env of 2 agents coordinating tasks
* ISO/IEC TR 24368:2022 - Information technology — Artificial intelligence — Overview of ethical and societal concerns
* IEEE 7000-2021 - Model Process for Addressing Ethical Concerns During System Design
* SHAPLey: <https://arxiv.org/abs/1705.07874>
* Evaluating LLMs - Principles, Approaches, and Applications: [Tutorial](https://neurips.cc/virtual/2024/tutorial/99524), [Whitepaper](https://services.google.com/fh/files/blogs/neurips_evaluation.pdf)
* AutoGen: <https://arxiv.org/abs/2308.08155>