Department of Computer and Information Systems Engineering

CS-406 Computer Engineering Project

Proposal for the Final Year Design Project

Title	**LLM Game Arena: A Gamified Approach to Evaluating Language Models**

Domain	Artifical Intelligence (AI)	Machine Learning (ML)	Natural Language Processing (NLP)	Autonomous Agents and Multi-Agent Systems	Cognitive Computing	Domain 6

1. Nature of Project [Tick all that applicable]

☒ New Project OR ☐ Extension of Existing Project	☐ Industrial Collaboration	☐ Funded
☐ Other Department Collaboration (If yes) Department Name________________	☐ Other Academic Institution Collaboration (If yes) Institution Name_______________	

2. Brief Outline (*Problem Identification and Significance*)

Traditional evaluation methods for Large Language Models (LLMs) mainly focus on static or short-turn tasks, which do not reflect the complexity of real-world scenarios. These models often struggle with long-term planning, spatial reasoning, and adapting to feedback. A major challenge is the “knowing-doing gap”, where models understand information but fail to act on it effectively—especially in dynamic or multi-modal settings. As AI systems evolve toward becoming autonomous agents, current evaluation methods fall short of testing the skills truly required for real-world operation.

Agentic AI is seen as a key direction for the future, but introducing these agents into real environments without proper evaluation poses serious risks. This project proposes a game-based evaluation framework as a practical, safe, and scalable way to assess LLMs in realistic, interactive scenarios. Games simulate essential elements of real life—such as uncertainty, decision-making, planning, and adaptation—making them ideal for testing the abilities agents need to succeed in real-world tasks. This approach offers deeper insights into model capabilities and supports the responsible development of intelligent, reliable AI systems.

3. Objectives

1. **To design and implement a game-based evaluation framework** that tests Large Language Model (LLM) agents across diverse and interactive environments.
2. **To assess the reasoning, planning, and decision-making abilities** of LLMs by observing their performance in tasks that mimic real-world challenges.

3. **To identify and analyze the limitations of current LLM agents**, including issues such as the "knowing-doing gap," poor long-term planning, and failure in multimodal decision-making.
4. **To compare the performance of different LLM agents** using well-defined metrics across various game scenarios that test cognitive, spatial, and strategic skills.
5. **To contribute to the responsible development of agentic AI** by providing insights into how well current models perform in dynamic, feedback-driven environments before deployment in real-world applications.
6. **To create and integrate multiple game environments** of varying complexity (e.g., puzzle, navigation, survival) for comprehensive evaluation across multiple skill dimensions.

4. Scope

This project focuses on the design and development of a game-based framework to evaluate the cognitive and decision-making abilities of Large Language Model (LLM) agents in interactive, simulated environments. The scope includes the integration of multiple game scenarios that mimic real-world challenges such as navigation, problem-solving, and long-term planning.

The project will involve implementing LLMs as autonomous agents that interact with diverse environments through natural language inputs and receive feedback based on their actions. Both textual and vision-language input formats may be explored to test the effect of modality on performance.

Additionally, the scope includes experimenting with small-scale multi-agent interaction, where LLM agents may collaborate within shared environments. The focus will be on analyzing model behavior and limitations in these complex settings. This project will utilize existing LLMs rather than developing new models and will provide insights into their performance and potential for real-world deployment.

5. Proposed Methodology

1. Environment Selection and Design

- Select or build a diverse set of interactive game environments (e.g., navigation, puzzle-solving, survival) that simulate real-world cognitive challenges.
- Incorporate both text-based and vision-based input formats to test LLMs under different perception conditions.

2. Agent Integration

- Integrate pre-trained Large Language Models (LLMs) as autonomous agents capable of interacting with the environments.
- Use prompting techniques (e.g., zero-shot, few-shot, chain-of-thought) to drive agent behavior.

- Include logic for reading environment states and generating valid actions.

3. **Vision Input Handling**
    - Develop or use wrappers that convert the game's current state into descriptive text and/or image-based input.
    - Test the impact of vision-language vs. text-only inputs on model behavior and decision-making.
4. **Multi-Agent Interaction**
    - Set up small-scale multi-agent scenarios to explore collaboration or competition between two or more LLM agents.
    - Track how agents perform when their actions depend on other agents' decisions.
5. **Evaluation and Metrics**
    - Define custom evaluation metrics for agent performance, including:
        - Task completion rate
        - Valid action ratio
        - Adaptability to feedback
        - Long-term planning accuracy
    - Log agent trajectories and analyze model behavior across different game types and difficulty levels.
6. **Comparative Analysis**
    - Run multiple models (e.g., GPT-4, Claude, LLaMA) across identical environments.
    - Compare their performance to identify strengths, weaknesses, and behavioral patterns under different strategies and inputs.
7. **Reporting and Visualization**
    - Visualize performance results using graphs and charts.
    - Record qualitative observations (e.g., strategy, failure cases, knowing-doing gap).
    - Document findings that highlight each model's level of agentic behavior and reasoning skill.

6. Resources Involved

1. Pre-trained LLMs and APIs
2. Game Environments
3. Vision Input Integration Tools
4. Cloud GPU Access
5. Research References

7. Description of Industrial Support (If any)

8. SDGs (If Applicable)

☐ No Poverty	☐ Zero Hunger
☐ Good Health and Well-Being	☐Quality Education
☐ Gender Equality	☐ Clean water and Sanitation
☐ Affordable and Clean Energy	☒ Decent Work and Economic growth
☒ Industry, Innovations and Infrastructure	☐ Reduced Inequalities
☐ Sustainable Cities and Communities	☐ Responsible Consumption and Production
☐ Climate action	☐ Life Below Water
☐ Life on Land	☐ Peace, Justice and Strong Institutions
☒ Partnerships	

9. Gantt Chart

Year	2025 to 2026											
Months												
Task 1												
Task 2												
:												
Task N												

10. Details of Project Team

i. Students

No.	Name	Seat No.	Signature (s)

1	Usman Faizyab Khan	CS-22076	
2	Muhammad Owais	CS-22080	
3	Zuhaib Noor	CS-22081	
4	Muhammad Zunain	CS-22086	

ii. Supervisors / Advisors

	Name	Designation & Department	Address & Contact	Signature(s)
Supervisor	Mr. Muhammad Ali Akhtar	Lecturer at CIS Department		
Co-Supervisor (If any)				
Industrial Advisor (If any)				

For Office Use Only

Project Serial No.: ___________ Dated: ______________________	Signature Convener Steering Committee	Signature FYP Coordinator

□ Proposal Approved	□ Not Approved	□ Returned for Clarification / Modification

Comments: (if any)	

____

(Signature of Chairperson)

Date: _______________________________