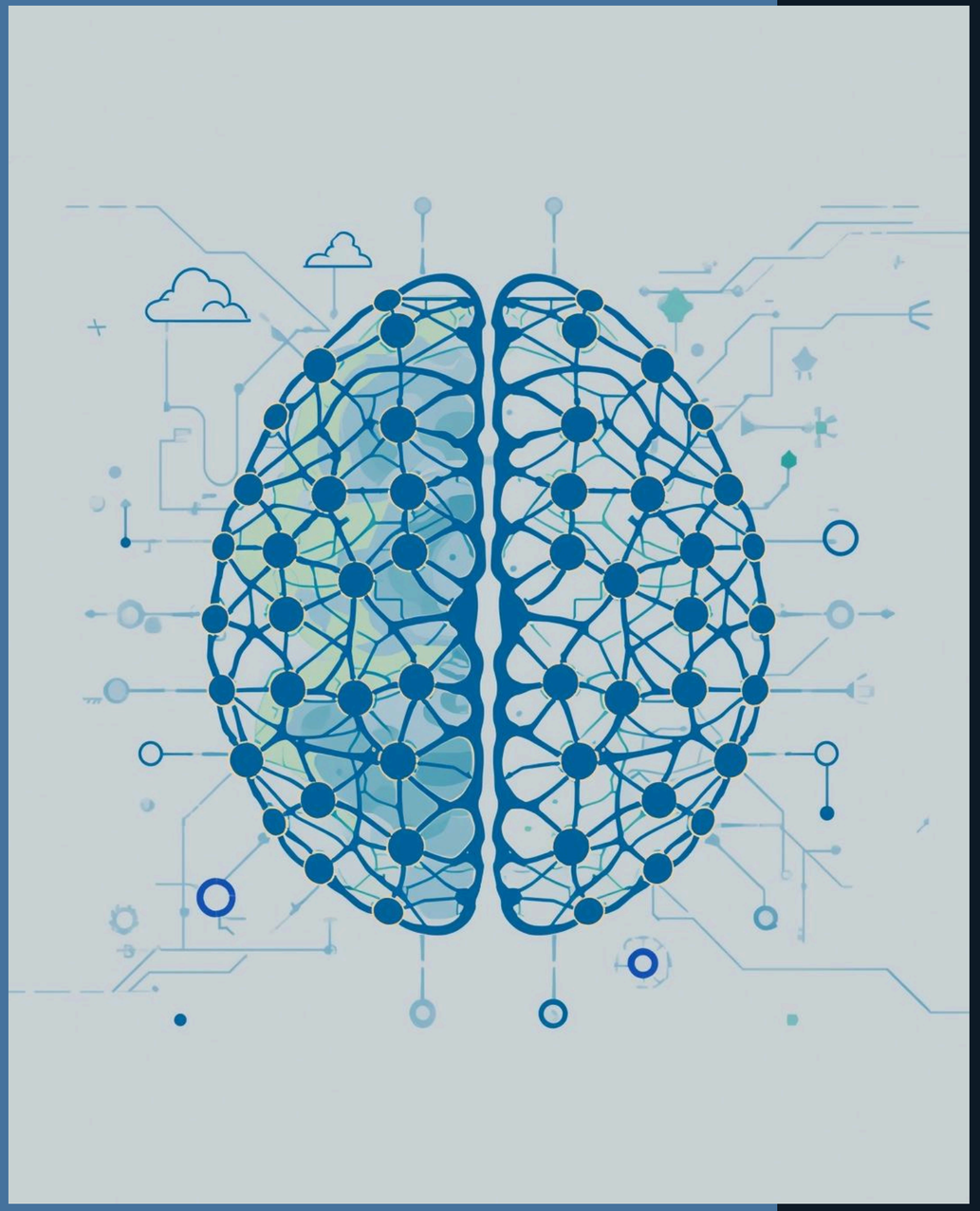

AMD Hackathon

Track 1

Team Name: ALPHA 1
Members: Abhinava Sai, Ujjawal Bhatia, Manish, Tarun Gangwar

Competition Overview: Q-Agent vs A-Agent

Competition Format and Scoring Logic

HEAD-TO-HEAD FORMAT

The competition features a **direct confrontation** between Q-Agent and A-Agent. Each agent performs distinct roles: Q-Agent generates questions while A-Agent provides answers, ensuring a dynamic interaction. This format emphasizes the strengths of both agents in a competitive setting.

SCORING LOGIC

Scoring is primarily based on the performance of the A-Agent, which focuses on accuracy and survival. The elimination rounds heavily rely on the A-Agent's ability to respond correctly and efficiently, highlighting its critical role in the overall scoring system.

System Architecture Overview

System Architecture Overview: Inputs and Layers

USER INPUTS

User inputs are routed to the Q-Agent, which processes these inputs and generates structured JSON questions. This systematic approach ensures clarity in the interaction and maintains a structured format for downstream processing.

OPPONENT QUESTIONS

Questions from the opponent are directed to the A-Agent, which then produces structured JSON answers. This separation of concerns allows for optimized task management and improves the efficiency of the overall system architecture.

Model Selection and Token Constraints

Model Selection Strategy Overview

Q-AGENT SELECTION

The Mistral-7B model was selected for the Q-Agent due to its **speed** and ability to generate structured output swiftly, meeting our competition's requirements.

A-AGENT SELECTION

The Qwen-14B model was chosen for the A-Agent, providing **superior reasoning capabilities** essential for accurate answer generation under competitive conditions.

REJECTION OF SMALLER MODELS

Smaller models were rejected as they failed to meet our **accuracy and time constraints**, critical for maintaining performance during the competition interactions.

Token and Time Constraints Overview

TOKEN LIMIT

The total token limit is capped at **1024 tokens** per interaction. This constraint includes both question and answer components, ensuring efficient communication between agents.

RESPONSE TIME TARGETS

Response time targets are set at **less than 13 seconds** for questions and **less than 9 seconds** for answers, emphasizing the need for speed and efficiency in agent performance.

ENGINEERING SOLUTIONS

Implementing token budgeting and efficient prompt design, alongside strict decoding control, addresses the outlined constraints, optimizing performance while maintaining compliance with competition standards.

Agent Design: Strategies and Implementation

Agent Design Overview: Q-Agent & A-Agent

Q-AGENT DESIGN

Q-Agent focuses on generating strict, structured JSON outputs, enhancing clarity in question formulation while employing distractor designs to increase challenge and engagement during interactions.

A-AGENT DESIGN

A-Agent utilizes greedy decoding for deterministic output and emphasizes reasoning capabilities, ensuring reliability and precision in answer generation, crucial for competitive performance in rounds.

ANSWER VALIDATION

Comprehensive validation mechanisms are in place to ensure compliance with JSON formatting and content rules, minimizing errors and ensuring the integrity of the responses provided by the A-Agent.

Training, Optimization, and Compliance Overview

Training Strategy Overview

DATASET GENERATION

A custom dataset generation pipeline was implemented, resulting in over 2000 samples per topic, ensuring comprehensive coverage and diversity for robust training of both agents.

DEDUPLICATION

A deduplication process was employed to remove redundant samples, enhancing model generalization and improving overall performance by eliminating noise from the training dataset.

LORA FINE-TUNING

LoRA fine-tuning was conducted with ranks of 16–32 and a learning rate of 2e-5 for 2–3 epochs, optimizing model performance while maintaining computational efficiency.

THANK YOU