GENAI

NAME - ANKUR SHARMA

SRN - PES2UG23CS077

Assignment_MOE.ipynb

Core Concept: Mixture of Experts (Router Architecture)

This assignment demonstrates how to build a Mixture of Experts (MoE) system, where multiple specialized components handle different types of queries. Instead of sending all requests to a single model, the system routes each request to the most appropriate expert.

Routing Logic

At the core of the architecture is a classification step:

1. The user sends a query
2. An initial LLM call determines the category of the request:
    - **Technical** → programming, bugs, code issues
    - **Billing** → payments, refunds, subscriptions
    - **Tool** → real-time data requests
    - **General** → casual questions or miscellaneous requests

This ensures that each query is handled by the component best suited to respond accurately.

Specialized Handling

Experts (System Prompts)

- Each category has a distinct persona or expertise:
    - Technical queries → "Senior Software Engineer"
    - Billing queries → "Customer Billing Specialist"
    - General queries → "Support Assistant"
- Personas are implemented through system prompts, guiding the model's tone, style, and focus

Tool Integration

- Some queries, like Bitcoin price, are routed to a Python function instead of the LLM
- This shows how MoE systems can combine AI with deterministic tools, improving efficiency and accuracy