GENAI

NAME: ANIRUDH MURALEEDHARAN

SRN: PES2UG23CS071

Assignment_MOE.ipynb

Core Concept: Mixture of Experts (MoE) – Router-Based Architecture

This assignment illustrates how to design a **Mixture of Experts (MoE)** system, where different specialized components manage different categories of queries. Instead of directing every request to a single model, the system intelligently routes each query to the most suitable expert.

Routing Mechanism

The foundation of this architecture is a classification stage:

1. The user submits a query
2. An initial LLM call analyzes and categorizes the request into one of the following types:
    - **Technical** → Programming issues, debugging, code-related problems
    - **Billing** → Payments, subscriptions, refunds
    - **Tool-Based** → Requests requiring real-time or external data
    - **General** → Casual inquiries or miscellaneous questions

This routing strategy ensures that each request is handled by the most appropriate expert, improving response accuracy and relevance.

Specialized Expert Handling

Expert Personas (System Prompts)

Each category is assigned a distinct expert persona:

- Technical queries → Responded to as a **Senior Software Engineer**
- Billing queries → Handled by a **Customer Billing Specialist**
- General queries → Managed by a **Support Assistant**

These personas are implemented through system prompts, which influence the model's tone, structure, and domain focus. This creates more professional and context-aware responses.

Tool Integration

Certain queries, such as retrieving the current Bitcoin price, are routed to a Python function rather than processed solely by the LLM.

This demonstrates how MoE systems can integrate AI models with deterministic tools. By combining intelligent routing with external functions, the system achieves:

- Greater efficiency
- Improved factual accuracy
- Better handling of real-time data