**Multi-LLM Chatbot**

**Đề bài: xây dựng chatbot tích hợp nhiều mô hình LLM free, gồm các tính năng**

* Tự động chuyển đổi giữa các model LLM (GPT, Gemini,…) khi hết token free của một model.
* Duy trì hội thoại liên tục khi chuyển đổi mô hình mà không bị mất lịch sử hội thoại.

**Hướng phát triển và giải pháp:**

* Detect câu hỏi liên tục và giới hạn token để kiểm soát việc sử dụng API cho từng model.
* Việc chuyển đổi giữa các mô hình, khi chuyển đổi sang mô hình 2 thì context và các session của cuộc trò chuyện với mô hình 1 cũng sẽ gửi sang mô hình 2 -> có thể tóm tắt lại nội dung trước khi đưa vào input của mô hình 2.

**Kiến trúc tổng quan hệ thống:**

1. FE (UI):
   * xây dưng giao diện chat (Streamlit)
2. BE (API, Logic): Điều phối LLMs, quản lý hội thoại
   * Sử dụng langchain để điều phối LLM
   * Lưu trữ lịch sử hội thoại: lưu history chat + session\_id cuộc trò chuyện vào DB, sử dụng mongoDB
   * Detect câu hỏi liên tục và limit token cho từng model khác nhau để kiểm soát việc sử dụng API
3. Model LLM:
   * GPT
   * Gemini
4. Cơ chế tự chuyển đổi model
   * Set limit token cho từng mô hình
   * Khi model hết token, chuyển sang model khác mà không mất lịch sử hội thoại
   * Dữ liệu hội thoại được lưu trữ để đảm bảo tính liên tục