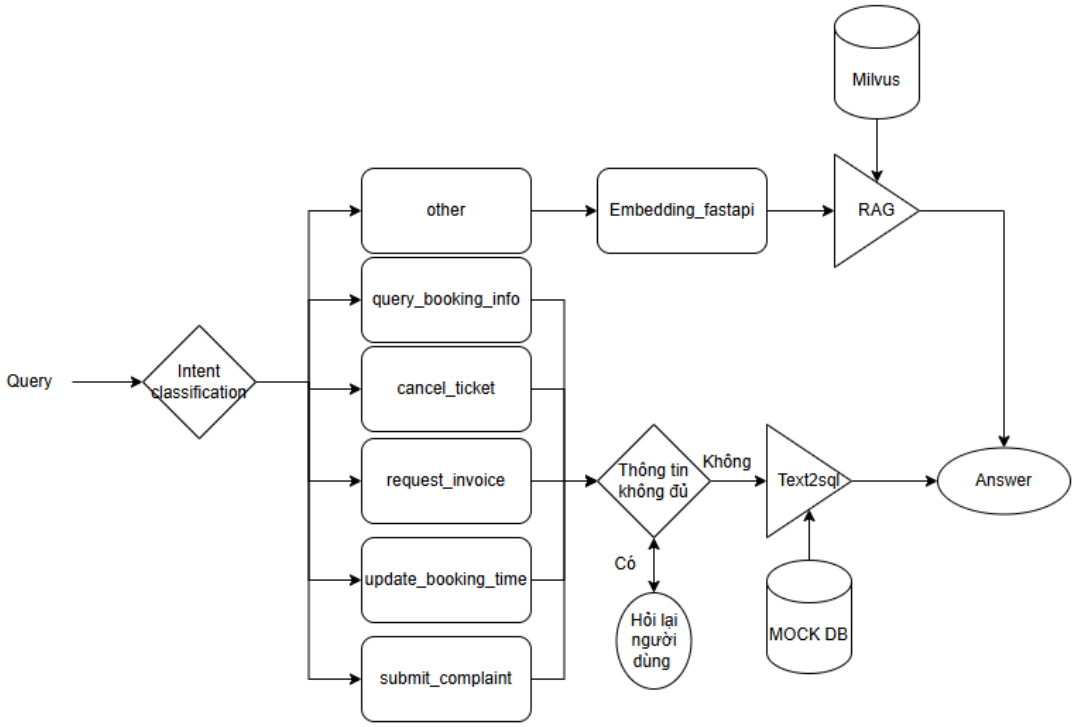
VEXERE TEST

1. Kiến trúc tổng quan



Ở đây có 2 cách xử lí cho intent của người dùng, 1 là trược tiếp phân loại (sử dụng ở folder src) và 2 là trực tiếp call tool (sử dụng ở folder src2).

Embeding dùng ở đây là Bge-m3 cùng với hybrid-search.

2. Framework

Langgraph: viết pipeline

Langchain: tương tác LLM

Milvus: để search câu hỏi liên quan

Huggingface: deploy embedding (do colab không cho chạy docker, nếu có thể ưu tiên dùng vllm để deploy)

3. Improvement

- Để có thể tương tác tốt với người dùng nên đi theo hướng dùng tool call (folder src2), nhưng do tool call cần nhiều thời gian prompt nên sử dụng cách này. Tuy nhiên tool call cũng sẽ bất lợi khi có nhiều tính năng thì prompt tool sẽ không hiệu quả bằng. Thay vào đó train 1 model nhỏ để phân loại sẽ hữu ích hơn.
- Cải tiến về text2sql, cũng giống như tool, cần có 1 kịch bản rõ ràng và các định nghĩa cột rõ hơn.
- Memory hiện tại là đưa toàn bộ lịch sử chat vào, do không thấy yêu cầu cần xây dựng long term, short term . Có thể sử dụng them mem0 để kết nối với mongo lưu trữ memory theo user id.