Design Document – AI Engineer Take-home Test (Vexere Chatbot POC)

---

1. Kiến trúc tổng quan

1.1. Mục tiêu

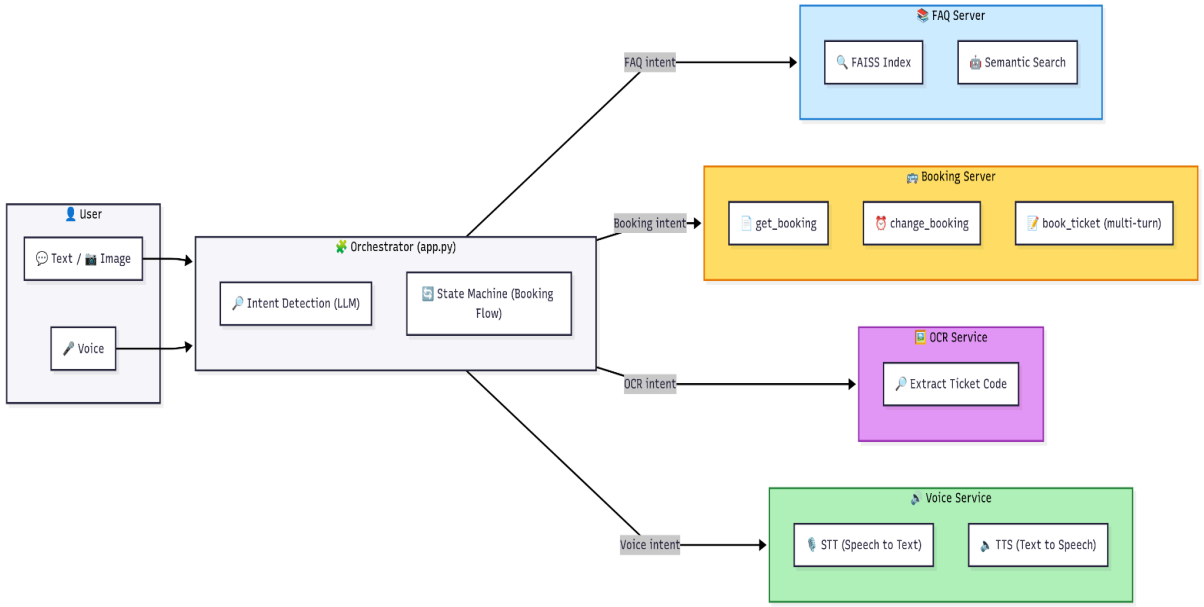
- Xây dựng chatbot thông minh cho Vexere, hỗ trợ:
    - **FAQ** từ dữ liệu `faq_data.csv`.
    - **Booking** (xem vé, đổi vé, đặt vé multi-turn).
    - **Đa modal**: text, voice (STT/TTS), image (OCR từ vé).
- Áp dụng **LLM + tool orchestration** (FastMCP) để kết nối nhiều module.
- Ưu tiên: **Nhanh, dễ mở rộng, rõ ràng**, phục vụ mục tiêu take-home test.

1.2. Kiến trúc logic

- Người dùng gửi **text / image / voice**.
- **Orchestrator (app.py)**:
    - Phân loại intent bằng **LLM (GPT-4o-mini)**.
    - Quản lý state (multi-turn booking).
    - Điều phối gọi các service (FAQ, Booking, OCR, Voice).
- **Các service chuyên biệt**:
    - FAQ → RAG với FAISS.
    - Booking → mock DB (get, change, book).
    - OCR → stub đọc mã vé từ file ảnh.

- Voice → stub chuyển giọng nói sang text.

1.3. Sơ đồ hệ thống



---

2. Lựa chọn công nghệ / framework

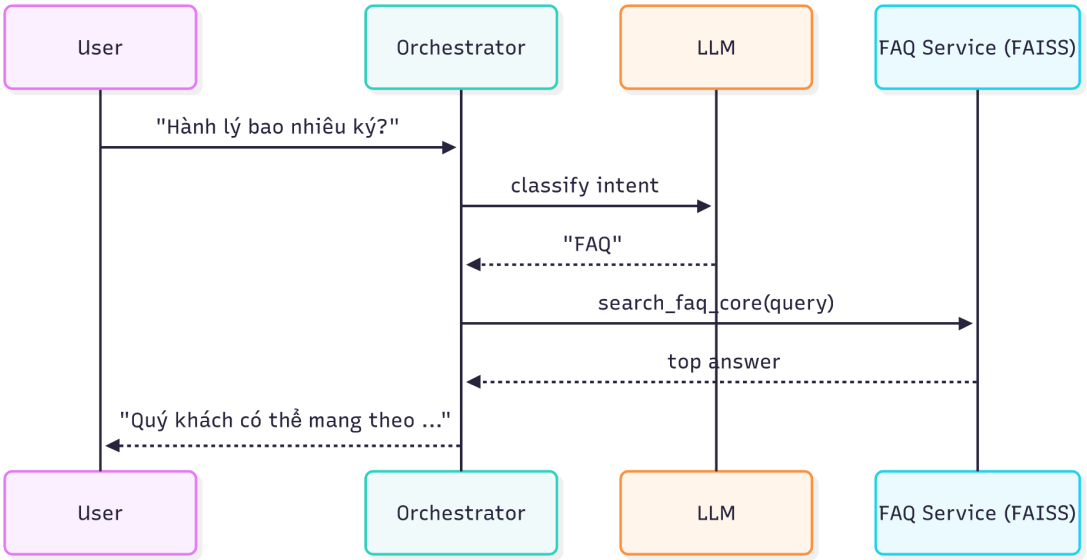
Thành phần	Công nghệ chọn	Lý do
Orchestrator	Python + asyncio	Đơn giản, dễ test, dễ mở rộng.
Intent Classification	OpenAI GPT-4o-mini	Hiểu ngôn ngữ tự nhiên tốt, tiết kiệm thời gian train custom model.
FAQ Search	FAISS + Sentence-BERT	Semantic search nhanh, nhẹ, phù hợp POC.
Booking Service	Mock Python service	Dễ kiểm soát, không phụ thuộc DB thật.
OCR	Stub service	Giả lập OCR để parse mã vé từ file ảnh.
Voice	Stub service (STT/TTS)	Giả lập speech, có thể thay bằng dịch vụ thực tế sau này.

Tool orchestration	FastMCP	Chuẩn hóa tool-call, dễ tích hợp khi mở rộng.
API Layer	Flask (option)	Nếu cần REST API để demo.

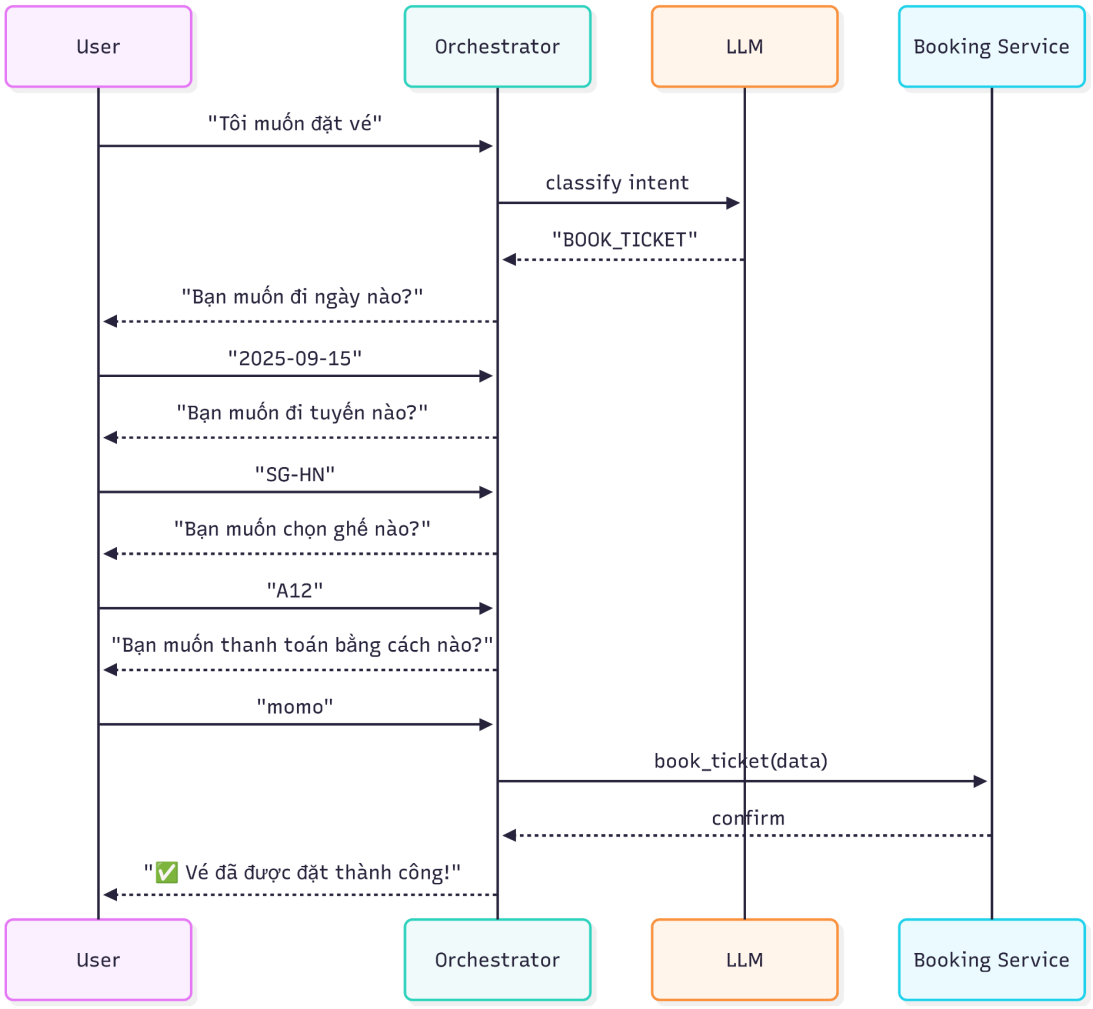
---

3. Sơ đồ thiết kế hệ thống liên quan

3.1. Sequence Diagram – FAQ Flow



3.2. Sequence Diagram – Booking Flow (multi-turn)



---

4. Test & Continuous Improvement Pipeline

4.1. Test Strategy

- **Unit test**:
    - `search_faq_core()`, `get_booking_core()`, `change_booking_core()`, `handle_booking_flow()`.
- **Integration test**: orchestrator gọi service qua function.
- **E2E test**: manual CLI test → simulate conversation.
- **Mock test**: OCR & Voice stub trả về kết quả giả lập.

4.2. Continuous Improvement

- Tích hợp **pytest + GitHub Actions** → auto test khi push code.
- Log toàn bộ intent + kết quả → feedback để cải thiện prompt LLM.
- Có thể thay stub bằng real service (Google Speech, Tesseract OCR).
- Nếu dữ liệu FAQ tăng lớn → thay FAISS bằng ElasticSearch Hybrid Search.

---

5. Các hạng mục liên quan khác

- **Scalability**:
    - Orchestrator có thể scale ngang (stateless), lưu state booking vào Redis.
    - FAQ service scale bằng sharding FAISS index hoặc ElasticSearch.
- **Cost**:
    - GPT-4o-mini nhẹ, chi phí thấp.
    - Có thể thay bằng LLM open-source nếu cần on-prem.
- **Extensibility**:
    - Dễ thêm intent mới (ví dụ: huỷ vé, refund).
    - Có thể tích hợp Payment Gateway (MoMo, ZaloPay).
- **Future work**:
    - Triển khai Web UI demo.
    - Thêm multi-language (Tiếng Anh/Khmer cho khách quốc tế).
    - Thêm TTS để bot “nói chuyện” thực sự với khách.