PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG CHATBOT

1. Hiện trạng và Vấn đề

1.1. Hiện trạng

Hệ thống ERP (Enterprise Resource Planning) của NHÓM 3 hiện đang là công cụ cốt lõi để quản lý các nghiệp vụ quan trọng, bao gồm quản lý đơn hàng, quản lý kho (tồn kho), và quản lý thông tin khách hàng. Hệ thống này đảm bảo tính nhất quán của dữ liệu và là nguồn thông tin (source of truth) cho mọi hoạt động vận hành.

Tuy nhiên, việc truy cập và khai thác thông tin từ hệ thống ERP này hiện còn phụ thuộc nhiều vào thao tác thủ công của nhân viên hoặc khách hàng phải đăng nhập vào các giao diện web phức tạp.

1.2. Vấn đề và Hạn chế

Mô hình vận hành hiện tại dẫn đến một số vấn đề:

1. **Quá tải hỗ trợ:** Nhân viên (hoặc khách hàng) thường xuyên phải trả lời các câu hỏi lặp đi lặp lại về thông tin tĩnh (ví dụ: "Chính sách bảo hành thế nào?", "Shop ở đâu?", "Giao hàng mất bao lâu?").
2. **Trải nghiệm người dùng kém:** Khách hàng (hoặc nhân viên nội bộ) muốn tra cứu thông tin động (ví dụ: "Đơn hàng #123 của tôi đang ở đâu?", "Sản phẩm XYZ còn hàng không?") phải thực hiện nhiều bước phức tạp hoặc phải chờ đợi nhân viên hỗ trợ.
3. **Lãng phí nguồn lực:** Việc tiếp nhận các khiếu nại, yêu cầu hỗ trợ (ví dụ: "Đơn hàng của tôi bị hỏng") đòi hỏi nhân viên phải nhập liệu thủ công từ email/điện thoại vào hệ thống ERP, dễ gây sai sót và chậm trễ.

1.3. Mục tiêu Hệ thống

Để giải quyết các vấn đề trên, đề tài này đề xuất xây dựng một Hệ thống Chatbot AI thông minh tích hợp trực tiếp vào giao diện web của ERP. Hệ thống này sử dụng kiến trúc RAG (Retrieval-Augmented Generation) và LLM Agent (với khả năng gọi công cụ - Tool-calling) nhằm đạt được các mục tiêu sau:

- **Tự động hóa Tư vấn:** Cung cấp câu trả lời ngay lập tức, 24/7 cho các câu hỏi dựa trên kho tri thức (FAQs, chính sách).
- **Tăng cường Tự phục vụ:** Cho phép người dùng tra cứu dữ liệu động (đơn hàng, tồn kho) từ ERP bằng ngôn ngữ tự nhiên.

- **Tối ưu hóa Nghiệp vụ: Tự động hóa việc tiếp nhận và tạo mới các phiếu hỗ trợ (support tickets) trực tiếp vào hệ thống ERP.**

2. Phân tích Yêu cầu

2.1. Yêu cầu Chức năng (Functional Requirements)

Dựa trên 3 mục tiêu đã đề ra, hệ thống chatbot phải đáp ứng các yêu cầu chức năng sau:

- FN-1: Chức năng Tư vấn (RAG)
    - Hệ thống phải có khả năng hiểu và trả lời các câu hỏi của người dùng dựa trên một kho tri thức (Knowledge Base) được cung cấp sẵn.
    - Kho tri thức này bao gồm: Các file văn bản, PDF, DOCX về chính sách (bảo hành, đổi trả, vận chuyển), câu hỏi thường gặp (FAQs), và thông tin mô tả sản phẩm.
    - Hệ thống phải trích dẫn được nguồn thông tin (nếu có) và không được bịa đặt (hallucinate) thông tin nằm ngoài kho tri thức.

- FN-2: Chức năng Tra cứu động (Agent - Read-only Tools)
    - Hệ thống phải có khả năng tra cứu và trả về thông tin động, theo thời gian thực từ cơ sở dữ liệu của ERP.
    - Các nghiệp vụ tra cứu cụ thể:
        - Tra cứu trạng thái đơn hàng (dựa trên mã đơn hàng).
        - Tra cứu số lượng tồn kho (dựa trên tên hoặc mã SKU sản phẩm).
        - (Tùy chọn) Tra cứu lịch sử mua hàng (dựa trên SĐT hoặc email đã xác thực).
    - Hệ thống phải hiểu được các thực thể (entities) trong câu hỏi của người dùng (ví dụ: "đơn hàng #123" -> order_id="123").

- FN-3: Chức năng Hỗ trợ (Agent - Write Tools)
    - Hệ thống phải có khả năng ghi (Write) dữ liệu mới vào ERP khi được yêu cầu.
    - Nghiệp vụ cụ thể: Tiếp nhận thông tin khiếu nại (ví dụ: "Đơn hàng 456 của tôi bị vỡ") và tự động tạo một phiếu hỗ trợ (support ticket) mới trong ERP với các thông tin đã được trích xuất (SĐT, mã đơn hàng, nội dung khiếu nại).

2.2. Yêu cầu Phi chức năng (Non-Functional Requirements)

- **NFN-1 (Hiệu năng):** Thời gian phản hồi trung bình của chatbot cho mỗi câu hỏi phải dưới 5 giây.

- **NFN-2 (Độ chính xác):**
    - Đối với chức năng Tư vấn (FN-1), tỷ lệ câu trả lời bám sát nội dung (faithfulness) phải trên 95%.
    - Đối với chức năng Tra cứu động (FN-2), tỷ lệ hoàn thành tác vụ (task completion rate) phải là 100% (nếu dữ liệu tồn tại).
- **NFN-3 (Bảo mật):** Hệ thống phải đảm bảo người dùng chỉ có thể tra cứu thông tin động (đơn hàng, lịch sử mua) thuộc về chính họ (yêu cầu xác thực phiên đăng nhập - session).
- **NFN-4 (Tính khả dụng):** Hệ thống chatbot phải hoạt động 24/7.

3. Thiết kế Hệ thống

3.1. Thiết kế Kiến trúc Tổng thể

Để đáp ứng đồng thời cả yêu cầu về dữ liệu tĩnh (FN-1) và dữ liệu động (FN-2, FN-3), hệ thống áp dụng kiến trúc kết hợp RAG và LLM Agent (Tool-calling).

- Luồng RAG (Tư vấn): Xử lý các câu hỏi FN-1. Hệ thống sẽ truy vấn CSDL Vector (VectorDB) để tìm ngữ cảnh liên quan, sau đó đưa cho LLM để tổng hợp câu trả lời.
- Luồng Agent (Động & Hỗ trợ): Xử lý FN-2 và FN-3. LLM sẽ đóng vai trò là "Agent điều phối", quyết định khi nào cần gọi các "Tools" (công cụ). Các "Tools" này chính là các hàm (functions) đã được định nghĩa để gọi API của ERP.

Dưới đây là sơ đồ kiến trúc tổng thể của hệ thống:

3.2. Sơ đồ Use Case (Use Case Diagram)

Sơ đồ Use Case mô tả các tương tác chính giữa Tác nhân (Actor) và các chức năng của hệ thống.

3.3. Sơ đồ tuần tự (Sequence Diagram)

Sơ đồ tuần tự mô tả **luồng tương tác thời gian** giữa các thành phần khi người dùng gửi câu hỏi.

3.4. Thiết kế cơ sở dữ liệu

Cơ sở dữ liệu được chia thành hai phần:

- **CSDL ERP** (đã có sẵn — không thay đổi)
- **CSDL Chatbot** (bổ sung để phục vụ lưu tri thức và hội thoại)

Mô hình ERD (Entity Relationship Diagram)