

Traditional Medicine Chatbot

Xây dựng Chatbot Y học Cổ truyền sử dụng kiến trúc Structured Agentic RAG

NHÓM THỰC HIỆN

MSSV	Họ và Tên	GV Hướng dẫn
23127056	Trần Cẩm Huy	PGS. Đinh Điền
23127489	Nguyễn Ngọc Minh Thư	TS. Nguyễn Hồng Bửu Long
23127542	Võ Nguyễn Thảo Uyên	ThS. Lương An Vinh
23127016	Lưu Huy Minh Quang	Nguyễn Thành Giang

Khoa Công nghệ Thông tin
Trường Đại học Khoa học Tự nhiên - DHQG TPHCM

Nội dung trình bày

- 1 Phát biểu bài toán
- 2 Giới thiệu bộ dữ liệu
- 3 Chiến lược huấn luyện
- 4 Kết quả thực nghiệm
- 5 Key Findings & Conclusion

Phát biểu bài toán & Giải pháp đề xuất

- **Bối cảnh:** Tri thức Y học cổ truyền (YHCT) Việt Nam phong phú nhưng chủ yếu tồn tại dưới dạng **văn bản phi cấu trúc**, gây khó khăn cho việc tra cứu và kế thừa trong kỹ nguyên số.
- **Vấn đề:** Các mô hình ngôn ngữ lớn (LLMs) khi áp dụng trực tiếp vào YHCT dễ gặp hiện tượng *hallucination* và thiếu căn cứ tri thức, tiềm ẩn rủi ro trong lĩnh vực y tế.
- **Giải pháp:** Đề xuất xây dựng **Chatbot YHCT** dựa trên kỹ thuật *Retrieval-Augmented Generation (RAG)*, kết hợp LLMs với nguồn dữ liệu YHCT đã được **chuẩn hóa và cấu trúc hóa**.

Mục tiêu của hệ thống

Cung cấp một hệ thống hỏi–đáp YHCT **thông minh, có căn cứ và an toàn**, hỗ trợ tra cứu theo cây thuốc, bài thuốc và triệu chứng bệnh.

Nguồn dữ liệu và Thách thức cấu trúc hóa

Nguồn dữ liệu

Bộ dữ liệu được tổng hợp từ 4 sách YHCT chuyên khảo (hơn 1200 trang), đã được **OCR và chuẩn hoá**:

- Cây cảnh – cây thuốc trong nhà trường
- Cây rau làm thuốc
- Cây thuốc, vị thuốc bệnh nội tiết
- Cấp cứu và chống độc

Đặc điểm dữ liệu:

- Văn bản **phi cấu trúc**, dài, nhiều thuật ngữ dân gian
- Bảng biểu OCR dễ sai lệch định dạng

Thách thức miền tri thức

- Một thực thể có thể thuộc nhiều vai trò (vd: **Trúc đào** vừa là dược liệu vừa là **độc chất**)
- Truy hồi văn bản thuần túy có thể bỏ sót **cảnh báo an toàn y tế**

Truy vấn đa miền

- “Cây thuốc tính hàn chữa đau bụng?”
- “Sơ cứu khi ngộ độc trúc đào?”

⇒ **Cần ánh xạ dữ liệu vào các schema riêng biệt**

Các Domain dữ liệu và Lược đồ (Schemas)

Sau khi phân tích nguồn dữ liệu, nhóm xác định các domain chính cần quản lý trong cơ sở tri thức YHCT:

Domain	Quy mô ước tính	Schema tiêu biểu
Dược liệu	~ 450 loài	MedicinalPlant, MedicinalVegetable
Bệnh lý	~ 180 mục	EndocrineSyndrome
Độc chất	~ 60 loại	PoisonousSubstance
Bài thuốc	~ 800 bài	RemedyRecipe, EmergencyProtocol

Các schema này là đầu vào cho quy trình trích xuất, lập chỉ mục và vector hoá dữ liệu.

Cấu hình Hệ thống Baseline RAG

1. Kiến trúc cốt lõi:

- **Loại hình:** Naive RAG.
- **Framework:** LlamaIndex.
- **Mục tiêu:** Tối ưu hóa tài nguyên Local (CPU).

2. Xử lý dữ liệu:

- **Input:** Tài liệu Markdown (.md).
- **Parser:** MarkdownNodeParser (Giữ cấu trúc phân cấp).

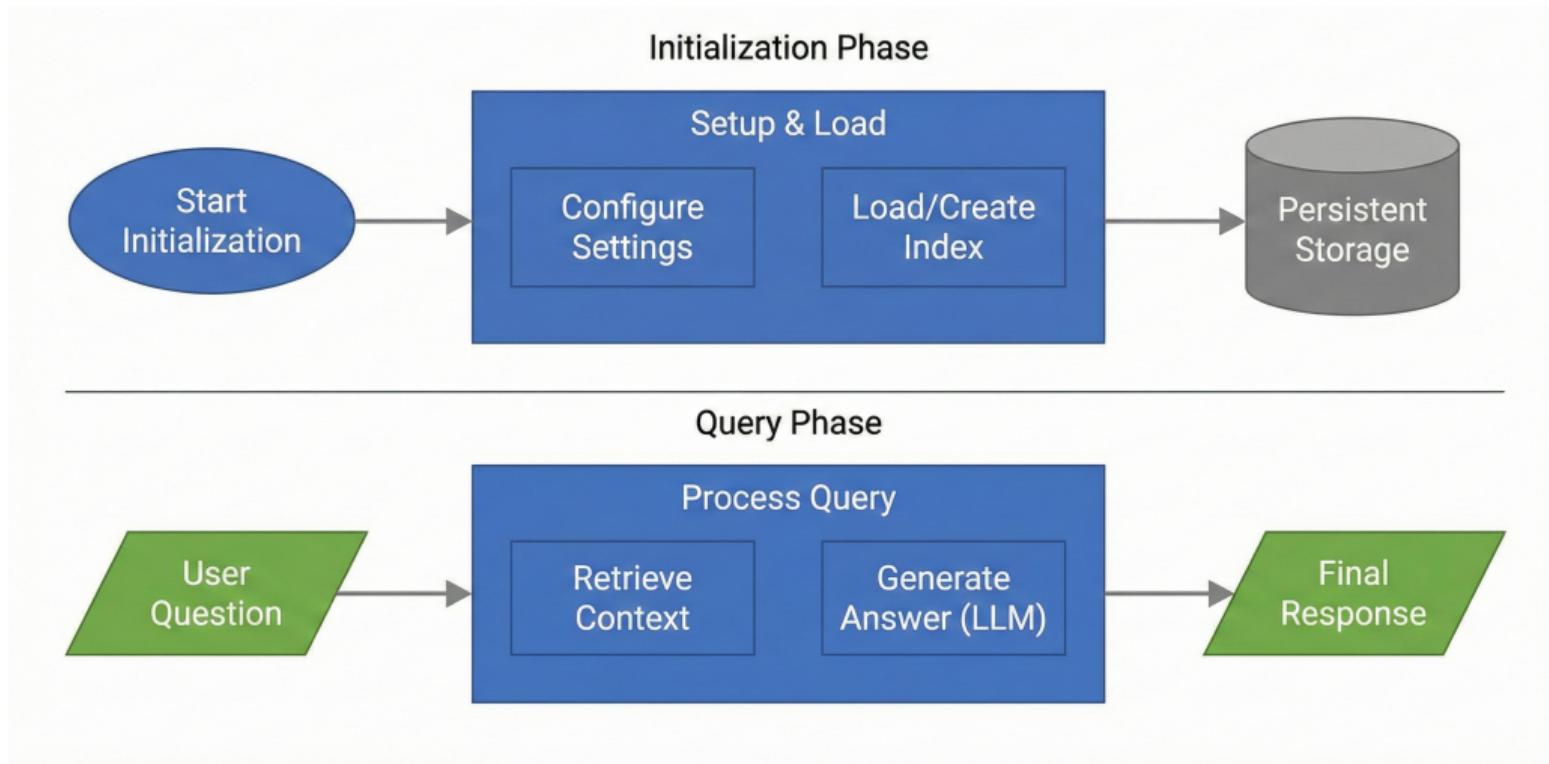
3. Embedding & Indexing:

- **Model:** Sử dụng model BAAI/bge-m3.
- **Cấu hình:** Chạy trên **CPU** (device="cpu").
- **Storage:** VectorStoreIndex (Local).

4. Trả lời câu hỏi:

- **Model:** Sử dụng Qwen2.5-14B-Instruct
- **Triển khai:** Self-hosted API.

Pipeline của mô hình baseline

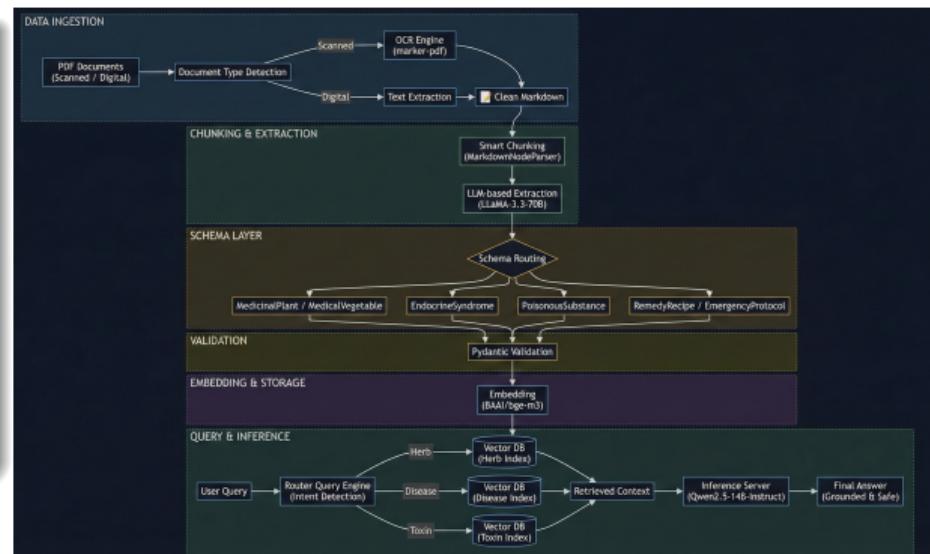


Tổng quan Pipeline và Kiến trúc Hệ thống

Structured RAG-based Architecture

- **Đầu vào:** PDF scan / digital (YHCT)
- **Tiền xử lý:** OCR → Markdown có cấu trúc
- **Tri thức:** Trích xuất JSON theo schema
- **Truy hỏi:** Vector Database
- **Suy luận:** LLM self-host (Google Colab)

Raw Data → Structured Knowledge →
Grounded Answer



Tiền xử lý và Trích xuất Tri thức Có Cấu trúc

⚠ Thách thức

- OCR tiếng Việt nhiều nhiễu (mất dấu, xuống dòng sai)
- Bảng biểu, bài thuốc dễ vỡ cấu trúc khi số hóa

✖ Giải pháp đề xuất

- **marker-pdf**: OCR bảo toàn layout tài liệu
- **Smart Chunking**: chia đoạn theo cấu trúc Markdown
- **Structured Extraction**: trích xuất JSON theo schema

Mô hình trích xuất: LLaMA-3.3-70B

Truy hồi, Định tuyến và Suy luận

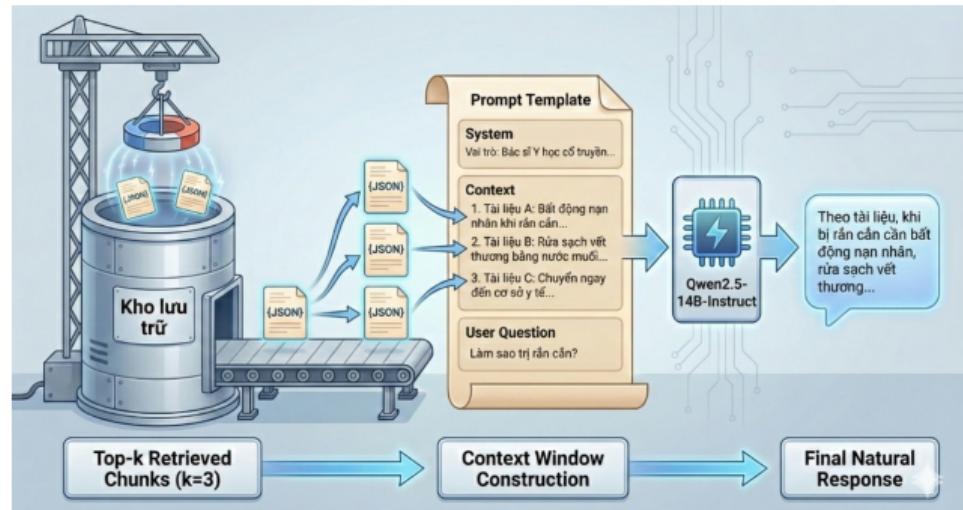
Truy hồi có kiểm soát

- Tách biệt 3 kho vector:
 - Dược liệu
 - Bệnh lý
 - Độc chất
- Embedding: BAAI/bge-m3

Router Query Engine

- Phân tích ý định truy vấn
- Chỉ kích hoạt 1 nguồn tri thức

Inference: Qwen2.5-14B-Instruct
(Context-only)



Kết quả Định lượng (Quantitative Results)

So sánh tổng quan hiệu năng RAGAS

Bảng dưới đây so sánh các chỉ số trung bình trên tập dữ liệu kiểm thử:

Model	Context Recall	Faithfulness	Ans. Relevancy	Ans. Correctness
Baseline	0.6944	0.5833	0.7685	0.6689
Proposed	0.3958	0.4134	0.7201	0.3840

Nhận xét dữ liệu:

- Phương pháp **Baseline** hiện đang cho kết quả số học cao hơn ở cả 4 chỉ số.
- Cần phân tích sâu hơn để hiểu tại sao phương pháp Proposed bị giảm sút về mặt chỉ số (Trade-off).

Phân tích chi tiết: Cải thiện vs Giảm sút

Mặc dù điểm trung bình thấp hơn, phương pháp Proposed vẫn cải thiện được một số trường hợp cụ thể (đặc biệt là Answer Correctness):

Metric	Better ($P > B$)	Worse ($P < B$)	Equal
Context Recall	2 (6.4%)	10	19
Answer Correctness	13 (41.9%)	17	1
Faithfulness	6 (19.3%)	10	15
Answer Relevancy	7 (22.6%)	17	7

- **Answer Correctness** có số lượng câu hỏi được cải thiện nhiều nhất (13 câu).
- Context Recall ít thay đổi nhất (19 câu giữ nguyên).

Phân tích định tính: Trường hợp cụ thể

So sánh câu trả lời thực tế để hiểu rõ ngữ nghĩa:

Ví dụ Cải thiện (Hải tảo)

- **Câu hỏi:** Tác dụng của Hải tảo?
- **Baseline:** Nêu quá nhiều tác dụng phụ, liệt kê bệnh không được hỏi.
- **Proposed:** Trả lời đúng trọng tâm ("tiêu đàm"), có trích dẫn nguồn.
- → Độ chính xác ngữ nghĩa tăng +0.58.

Ví dụ Giảm sút (Cúc bách nhật)

- **Vân đề:** Hệ thống bị lỗi cắt cụt câu trả lời (truncate) hoặc không truy hồi đủ ngữ cảnh.
- → Dẫn đến điểm số RAGAS thấp trong các metric tự động.

Tổng kết thực nghiệm:

- Phương pháp **Proposed Structured RAG** cho thấy khả năng trả lời trọng tâm hơn trong các câu hỏi y học phức tạp (13/31 trường hợp cải thiện về độ chính xác).
- Tuy nhiên, điểm trung bình chung thấp hơn do lỗi kỹ thuật (truncate) và thiếu ngữ cảnh (Context Recall thấp).

Kiến nghị:

- ① Cải thiện bộ Router để định tuyến câu hỏi chính xác hơn.
- ② Mở rộng Context Window để khắc phục lỗi thiếu thông tin.
- ③ Tinh chỉnh lại Prompt để tránh ảo giác (Hallucination) trong Baseline.