AI VIET NAM

TIME-SERIES TEAM

HOME ABOUT CONTENT OTHERS

Chào mừng sự xuất hiện

Trợ lý Gạo AI

BẠN

có phải đang gặp vấn đề?

📚 Bạn có thường xuyên làm việc với các tài liệu kỹ thuật, học thuật hay báo cáo dài hàng trăm trang dưới dạng PDF?

❌ Bạn có gặp khó khăn khi muốn tìm nhanh thông tin cụ thể trong tài liệu mà không cần đọc toàn bộ?

🔍 Bạn có cảm thấy việc dùng công cụ tìm kiếm từ khóa (như Ctrl+F) không đủ thông minh để hiểu ngữ cảnh câu hỏi?

🌐 Bạn có mong muốn một chatbot hỗ trợ tiếng Việt thực sự hiệu quả, thay vì chỉ có các công cụ tiếng Anh?

AI VIET NAM

TIME-SERIES TEAM

Gạo AI

sẽ giúp bạn!

AI VIET NAM

TIME-SERIES TEAM

Gạo AI

là ai?

Gạo AI là trợ lý ảo hỗ trợ bạn trong việc hỏi đáp vấn đề liên quan từ tài liệu PDF bằng công nghệ RAG kết hợp mô hình Vicuna-7B, được xây dựng bằng Streamlit và LangChain.

AI VIET NAM

TIME-SERIES TEAM

HOME | **ABOUT** | CONTENT | OTHERS

Gạo AI

giúp bạn được gì?

- 📄 Tài liệu cá nhân: Tải lên file PDF hoặc dán link tài liệu tiếng Việt để hỏi – Gạo AI sẽ trả lời theo ngữ cảnh.
- 🎓 Sinh viên: Ôn tập tài liệu học hoặc blog tổng hợp kiến thức lớp AIO.
- 👩‍💼 Người đi làm: Truy vấn nhanh thông tin từ hợp đồng, báo cáo nội bộ.
- 👩‍🏫 Giảng viên: Dùng tài liệu giảng dạy để AI hỗ trợ trả lời cho người học.
- 📚 Nội dung lớp AIO: Chọn Git Repo chứa blog tổng hợp từ Tuần 1 đến nay và đặt câu hỏi để ôn tập.

Nền tảng công nghệ Gạo AI

- 🔗 LangChain: Quản lý pipeline xử lý RAG.
- 🤗 HuggingFace + FAISS: Truy hồi và xử lý dữ liệu ngữ nghĩa.
- 🧠 RAG (Retrieval-Augmented Generation): Kết hợp truy hồi tài liệu và sinh ngôn ngữ từ LLM.
- 💻 Streamlit: Tạo giao diện web đơn giản, dễ sử dụng.

AI VIET NAM

TIME-SERIES TEAM

Demo Gạo AI

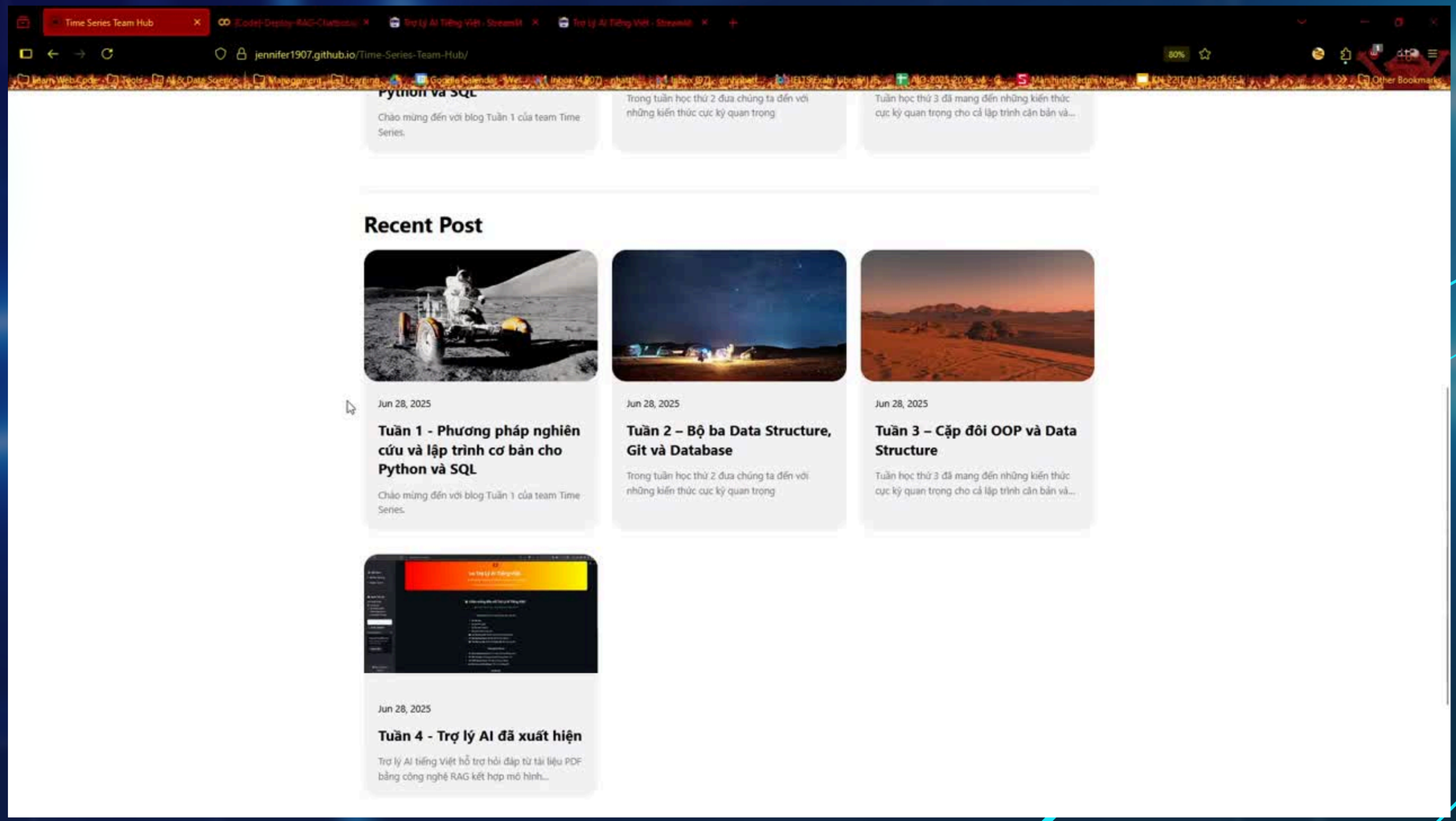
AI VIET NAM

TIME-SERIES TEAM

HOME ABOUT **CONTENT** OTHERS

Demo

Gạo AI



Gạo AI ứng dụng RAG

RAG (RETRIEVAL-AUGMENTED GENERATION) LÀ KỸ THUẬT GIÚP CHATBOT TRẢ LỜI DỰA TRÊN DỮ LIỆU THỰC TẾ, GIẢM ẢO GIÁC (HALLUCINATION) VÀ KHÔNG CẦN HUẤN LUYỆN LẠI MÔ HÌNH LỚN.

👉 RAG GIÚP CHATBOT THÔNG MINH HƠN, TRẢ LỜI CHÍNH XÁC HƠN TỪ DỮ LIỆU PDF MÀ KHÔNG CẦN HUẤN LUYỆN LẠI MÔ HÌNH.

Gạo AI ứng dụng RAG

⚙️ **Cách hoạt động của RAG:**

1. Truy hồi (Retrieval): Tìm đoạn văn bản liên quan từ tài liệu đã được nhúng (embedding) và lưu trong Vector Database.
2. Tăng cường ngữ cảnh (Augmentation): Ghép các đoạn tìm được vào prompt để mô hình hiểu rõ hơn.
3. Sinh câu trả lời (Generation): LLM sinh câu trả lời dựa trên ngữ cảnh đã tăng cường.

🧩 **Thành phần chính của hệ thống RAG:**

- 📥 Text Loader: Đọc tài liệu từ PDF, TXT...
- ✂️ Text Splitter: Chia nhỏ văn bản phù hợp với mô hình.
- 🔢 Embedding Model: Mã hóa đoạn văn thành vector.
- 🧠 Vector Database: Lưu trữ vector để tìm kiếm nhanh.
- 🎯 Retriever: Truy xuất đoạn phù hợp với câu hỏi.
- 🤖 LLM + Prompt: Sinh câu trả lời dựa trên thông tin đã tìm được.

Gạo AI kết hợp Langchain

LANGCHAIN LÀ MỘT FRAMEWORK MÃ NGUỒN MỞ GIÚP XÂY DỰNG CÁC ỨNG DỤNG SỬ DỤNG MÔ HÌNH NGÔN NGỮ LỚN (LLM) NHƯ GPT, VICUNA, MISTRAL... MỘT CÁCH HIỆU QUẢ, LINH HOẠT VÀ CÓ CẤU TRÚC.

VIỆC KẾT HỢP THÊM ỨNG DỤNG LANGCHAIN GIÚP GIẢI QUYẾT VẤN ĐỀ CỦA RAG VÀ CẢI THIỆN HIỆU QUẢ CỦA GẠO AI

Gạo AI kết hợp Langchain

LangChain khi kết hợp với RAG thì:

✅ Chuẩn hóa quy trình xây dựng RAG:

- Giúp tổ chức rõ ràng các bước: load → split → embed → retrieve → generate.
- Hạn chế việc viết code rời rạc, tăng khả năng bảo trì.

🔗 Kết nối mượt mà giữa các thành phần:

- LangChain tích hợp sẵn loader, retriever, LLM... nên không cần dùng thư viện rời rạc.
- Giảm lỗi định dạng dữ liệu giữa các bước.

🧠 Tự động hóa và tối ưu prompt:

- Hỗ trợ PromptTemplate và chain mẫu (như RAG Prompt) giúp đảm bảo ngữ cảnh rõ ràng.
- Linh hoạt với số lượng tài liệu truy hồi.

🔄 Dễ dàng mở rộng và tái sử dụng:

- Có thể thay thế nhanh mô hình embedding, vector DB, hoặc LLM mà không ảnh hưởng phần còn lại.

🐞 Dễ kiểm soát và debug từng bước:

- Mỗi bước trong pipeline có thể được theo dõi, kiểm tra hoặc tái sử dụng độc lập.
- Hỗ trợ LCEL để quan sát từng thành phần trong chuỗi xử lý.

AI VIET NAM

TIME-SERIES TEAM

RAG + Langchain

tạo điểm mạnh cho Gạo AI

LangChain giúp xây dựng hệ thống RAG theo hướng modular – mô đun hóa, trong đó mỗi bước là một thành phần riêng biệt:

- 📥 Text Loader: Đọc tài liệu từ PDF, TXT...
- ✂️ Text Splitter: Chia nhỏ văn bản phù hợp với mô hình.
- 🔢 Embedding Model: Mã hóa đoạn văn thành vector.
- 🧠 Vector Database: Lưu trữ vector để tìm kiếm nhanh.
- 🎯 Retriever: Truy xuất đoạn phù hợp với câu hỏi.
- 🤖 LLM: Gọi mô hình ngôn ngữ
- 🤖 LLM Chain: Tạo chuỗi xử lý sinh câu trả lời

AI VIET NAM

TIME-SERIES TEAM

HOME ABOUT **CONTENT** OTHERS

Lợi thế Langchain

tạo điểm mạnh cho Gạo AI

🧱 Runnable – điểm khác biệt của LangChain

LangChain dùng Runnable như một "block LEGO thông minh" để tạo pipeline linh hoạt và mạnh mẽ:

- RunnableLambda: Bọc hàm lambda thành một khối logic có thể chạy được
- RunnableSequence: Xâu chuỗi các bước tuần tự như pipe trong Unix
- RunnableMap: Chạy các bước song song (parallel)

```
rag_chain = (
    {"context": retriever | format_docs, "question": RunnablePassthrough()}
    | prompt
    | st.session_state.llm
    | StrOutputParser()
```

Lợi thế Langchain
tạo điểm mạnh cho Gạo AI

🧱 Tạo ra lợi thế cho Gạo AI:

- Xây pipeline dễ dàng → Ghép các bước xử lý (load, split, embed, truy vấn, LLM) thành chuỗi rõ ràng.
- Hỗ trợ chạy song song (async) → Giảm thời gian phản hồi chatbot.
- Mở rộng linh hoạt → Từ chatbot đơn giản → chatbot có feedback, branching, agent.

Điểm khác vượt trội

Gạo AI

🧱 Hệ thống này hỗ trợ nhiều phương thức nhập liệu:

📎 Đa dạng loại tài liệu:

- Tài liệu PDF (.pdf)
- Tài liệu Word (.docx)
- Bảng tính Excel (.xlsx, .xls)
- Tải Thư Mục (ZIP)

🔗 GitHub Repository: Tải file PDF từ kho GitHub và truy vấn toàn bộ kiến thức AIO từ tuần 1-4. Tiếp tục được cập nhật

📂 Thư Mục Cục Bộ: Tải file từ đường dẫn thư mục cục bộ

AI VIET NAM

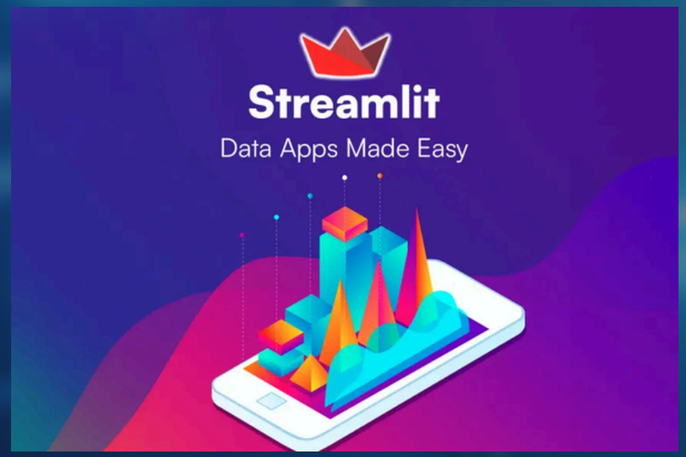
TIME-SERIES TEAM

HOME ABOUT **CONTENT** OTHERS

Ứng dụng Streamlit

tạo giao diện web cho Gạo AI

- Streamlit là thư viện Python mã nguồn mở giúp tạo web app tương tác nhanh chóng, không cần biết front-end.
- Được dùng nhiều trong AI và Data Science để làm dashboard, demo mô hình, chatbot...



AI VIET NAM

TIME-SERIES TEAM

HOME ABOUT **CONTENT** OTHERS

Ứng dụng Streamlit

tạo giao diện web cho Gạo AI

Một số widget phổ biến được ứng dụng:

- st.title(), st.header() → Tiêu đề
- st.write() → Hiển thị văn bản, bảng, ảnh...
- st.text_input(), st.button() → Nhập liệu, tương tác
- st.file_uploader() → Upload file
- st.selectbox() → Chọn lựa

➡️ Giúp tạo UI nhanh mà không cần HTML/CSS.

Một số widget nâng cao được ứng dụng:

- st.session_state: Dùng để giữ dữ liệu không bị mất khi app chạy lại.
- st.markdown(..., unsafe_allow_html=True) để biến giao diện với HTML/CSS, cho phép chỉnh giao diện: màu sắc, bố cục, hiệu ứng...
- st.cache_resource: Dùng để lưu kết quả của hàm tải mô hình hoặc vector store, tránh load lại mỗi lần chạy. Tiết kiệm thời gian, tăng hiệu suất khi xử lý tài nguyên nặng (ví dụ: HuggingFace Embedding).
- Tạo ứng dụng nhiều trang (Multi-page App): Giúp chia ứng dụng thành nhiều file .py, mỗi file là một trang chức năng riêng.
- st.experimental_rerun(): Dùng để reload toàn bộ app khi có sự kiện (ví dụ: người dùng tải file).

AI VIET NAM

TIME-SERIES TEAM

Gạo AI

Tóm tắt điểm mạnh

- ✅ Trả lời nhanh, chính xác dựa trên tài liệu của bạn
- 🔍 Tìm kiếm thông minh nhờ mô hình Embedding tiếng Việt
- ♻️ Dễ mở rộng, tùy biến với LangChain + Streamlit
- 📂 Hỗ trợ nhiều nguồn: GitHub, file upload, thư mục cục bộ
- 📊 Tương tác thân thiện với giao diện web đơn giản, dễ dùng

AI VIET NAM

TIME-SERIES TEAM

Thank You

Không chỉ là chatbot, mà là người trợ lý AI thực thụ!

💬 Hỏi gì – Trả lời đó

📚 Có tài liệu – Có kiến thức

⏱️ Tiết kiệm thời gian – Tăng hiệu suất học và làm việc