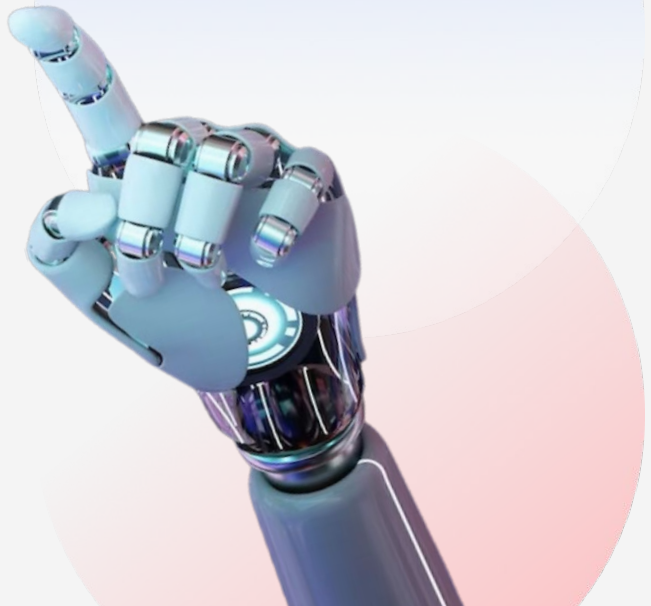
INDIVIDUAL SYSTEMS

LLM-Powered Data Document Intelligence

Transforming raw data and documents into actionable intelligence.



CÔNG TY CỔ PHẦN IVS

Website
indivisys.vn

LỢI ÍCH & ỨNG DỤNG THỰC TIỄN

Đọc & hiểu đa định dạng

- Hỗ trợ PDF, Word, Excel, CSV, PPT, Markdown.
- Phân tích cả cấu trúc (XML/JSON tree) và nội dung văn bản.
- Đảm bảo độ chính xác cao trên tài liệu phức tạp hoặc có bố cục lồng nhau.

Trích xuất & chuẩn hoá dữ liệu

- Tách thông tin theo trường dữ liệu, định danh và kiểu giá trị.
- Chuẩn hóa văn bản, loại bỏ ký tự thừa, làm sạch dữ liệu.
- Hỗ trợ **tích hợp Retrieval Augmented Generation hoặc module xử lý dữ liệu khác**.

Tự động phân tích & trực quan hoá

- Sinh code Python tự động để phân tích và vẽ biểu đồ.
- Tự động giải thích kết quả thống kê hoặc xu hướng.
- Kết hợp **LLM reasoning + tool-calling** để đảm bảo chính xác và minh bạch.

Giao tiếp tự nhiên

- Hỗ trợ giao tiếp bằng **văn bản và giọng nói**, đa ngôn ngữ.
- Giữ ngữ cảnh hội thoại xuyên suốt nhiều phiên.
- Áp dụng **Qwen3-Omni** để chuyển đổi giọng nói thành văn bản.

LỢI ÍCH & ỨNG DỤNG THỰC TIỄN

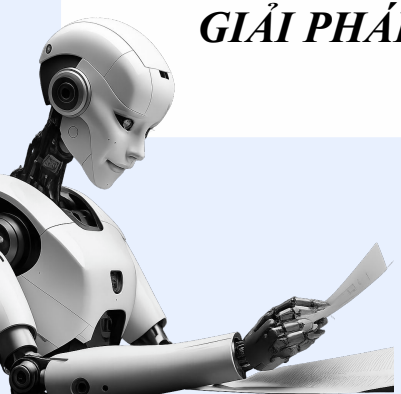
Mở rộng module dễ dàng

- Tích hợp linh hoạt qua **Re-Act, Tool-calling, RAG**.
- Cho phép bổ sung module xử lý, báo cáo hoặc truy xuất dữ liệu mới.
- Thiết kế **plug-in architecture**, hỗ trợ tái sử dụng logic AI.

Tạo & tối ưu code

- Tự động sinh mã phân tích dữ liệu, xử lý CSV, trực quan hoá kết quả.
- Hỗ trợ tối ưu hiệu suất mã nguồn Python do LLM sinh ra.
- Có thể chạy trực tiếp qua **Python subprocess runtime**.

GIẢI PHÁP AI TOÀN DIỆN CHO PHÂN TÍCH DỮ LIỆU, XỬ LÝ TÀI LIỆU VÀ TƯƠNG TÁC THÔNG MINH

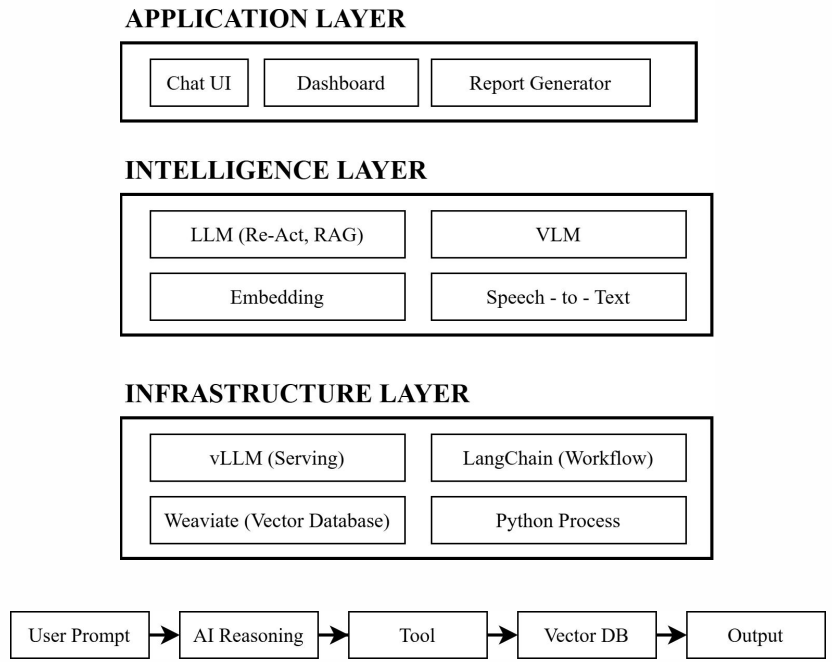


KIẾN TRÚC TỔNG THỂ

Application Layer – User Interface & Interaction
- Entry point của hệ thống, chịu trách nhiệm thu nhận, hiển thị và truyền tải yêu cầu từ người dùng.
- Bao gồm các thành phần: **Chat UI, Analytics Dashboard**, và **Report Generator**.
- Tầng này đảm nhận việc chuẩn hóa input, đóng gói ngữ cảnh người dùng trước khi gửi xuống **Intelligence Layer**.
- Kết quả phản hồi từ AI được định dạng lại (text, bảng, biểu đồ) để hiển thị tối ưu ở giao diện đầu cuối.

Intelligence Layer – Core Reasoning & Multimodal Processing
- Tầng xử lý trung tâm, đảm nhiệm toàn bộ hoạt động suy luận, lập kế hoạch và hiểu ngữ cảnh.
- **Large Language Model - LLM (Re-Act, RAG)** đảm nhiệm reasoning, planning và tool-calling.
- **Vision Language Model - VLM** xử lý tài liệu scan và hình ảnh, hỗ trợ đọc hiểu đa phương thức.
- **Embedding Model** mã hóa ngữ nghĩa và phục vụ truy vấn tri thức vector.
- **Qwen3-Omni** xử lý giọng nói, chuyển đổi speech-to-text và hỗ trợ hội thoại tự nhiên.
- Tầng này hợp nhất giữa suy luận (reasoning) và truy xuất tri thức (retrieval) để sinh phản hồi chính xác và có căn cứ.

Infrastructure Layer – Serving, Orchestration & Execution Runtime
- Tầng nền chịu trách nhiệm triển khai, điều phối và lưu trữ tri thức cho toàn hệ thống AI.
- **vLLM** đảm nhiệm việc phục vụ mô hình và hỗ trợ **streaming inference**.
- **LangChain** điều phối workflow, quản lý Re-Act loop và tool-calling.
- **Weaviate** lưu trữ embedding vector, hỗ trợ **semantic retrieval** trong pipeline RAG.
- **Python Process** cung cấp runtime an toàn cho việc thực thi code sinh bởi LLM.
- Hạ tầng được thiết kế mở rộng (scalable), tách biệt rõ ràng giữa logic AI, tầng phục vụ và kho dữ liệu tri thức.



INDIVIDUAL SYSTEMS

MÔ HÌNH & CÔNG NGHỆ NỀN TẢNG

LLM (Large Language Model)

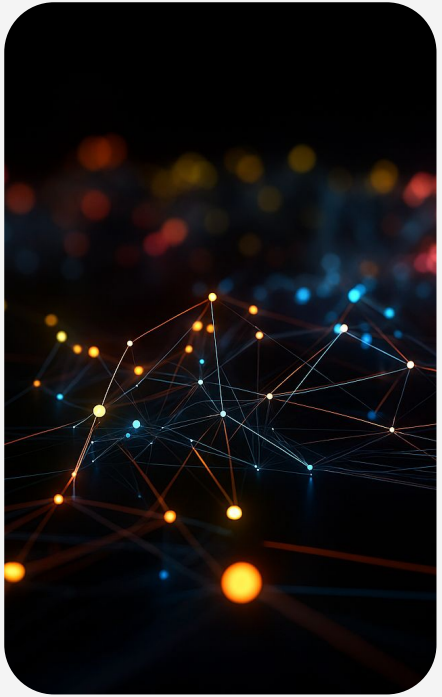
Mô hình: Qwen3
Quy mô: 4B – 235B parameters
Chức năng: Sinh, phân tích, và suy luận ngữ nghĩa trong văn bản.
Đặc điểm: Hỗ trợ **Thinking Mode** và **Context Window 128K tokens**.
→ Cho phép mô hình xử lý cùng lúc **nhiều tài liệu lớn**, hoặc **chuỗi hội thoại dài** mà không mất ngữ cảnh,
→ Giảm nhu cầu chia nhỏ dữ liệu. giúp phân tích và tổng hợp nội dung chính xác hơn trong các tác vụ đọc hiểu báo cáo, hợp đồng, hoặc log dài.

VLM (Vision Language Model)

Mô hình: dots.ocr
Chức năng: Trích xuất và hiểu nội dung từ tài liệu dạng ảnh (scanned PDF, biểu đồ, bảng).
Đặc điểm: Kết hợp **OCR + Visual Captioning**.
→ Không chỉ nhận dạng ký tự mà còn **hiểu bố cục và ý nghĩa ngữ cảnh của hình ảnh**.
→ Giúp trích xuất chính xác hơn trong các tài liệu kỹ thuật, hóa đơn, bản vẽ hoặc báo cáo có hình minh họa.

Embedding Model

Mô hình: Qwen3-Embedding-0.6B
Chức năng: Mã hóa văn bản thành vector ngữ nghĩa phục vụ tìm kiếm và truy xuất tri thức (Retrieval Augmented Generation-RAG).
Đặc điểm: Nhẹ, tốc độ cao, **Context Window 32K tokens**.
→ Cho phép tạo vector trên **toàn bộ tài liệu hoặc chương mục lớn** thay vì từng đoạn ngắn.
→ Cải thiện đáng kể độ chính xác khi truy vấn tri thức (semantic retrieval) và giảm lỗi mất ngữ cảnh khi RAG hoạt động.



MÔ HÌNH & CÔNG NGHỆ NỀN TẢNG

Speech Recognition

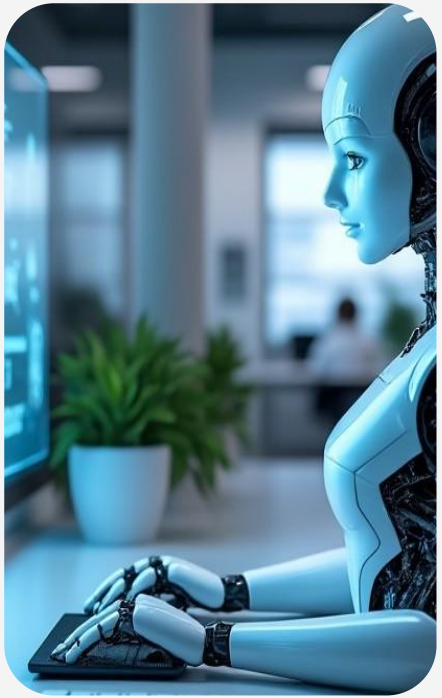
Mô hình: Qwen3-Omni
Chức năng: Nhận dạng giọng nói, chuyển đổi **speech-to-text**, hỗ trợ đa ngôn ngữ.
Đặc điểm: Ổn định với tín hiệu âm thanh nhiễu.
→ Cho phép tích hợp **voice input tự nhiên** cho các trợ lý hội thoại.

Model Serving

Công nghệ: vLLM
Vai trò: Triển khai và quản lý mô hình LLM/VLM trên hạ tầng nội bộ.
Đặc điểm: Hỗ trợ **low-latency inference** và **streaming response**.
→ Giúp hệ thống phản hồi theo thời gian thực (token-by-token) với hiệu năng cao.
→ Phù hợp cho ứng dụng hội thoại hoặc dashboard phân tích động.

Workflow Orchestration

Công nghệ: LangChain
Vai trò: Điều phối workflow, quản lý Re-Act loop, tool-calling, và trạng thái hội thoại (memory).
Đặc điểm: Cho phép xây dựng pipeline đa bước (**multi-step reasoning**) giữa các mô hình và công cụ.
→ Tăng khả năng lập kế hoạch, kiểm tra và phản hồi tự động của hệ thống.



MÔ HÌNH & CÔNG NGHỆ NỀN TẢNG

Vector Database

Công nghệ: Weaviate
Vai trò: Lưu trữ embedding vector, phục vụ **semantic search** và **knowledge retrieval** trong pipeline RAG.
Đặc điểm: Cho phép truy vấn ngữ nghĩa trên hàng triệu bản ghi với thời gian phản hồi mili-giây.
→ Kết hợp trực tiếp với LLM để tạo câu trả lời có căn cứ (cited answer).

Execution Runtime

Công nghệ: Python process
Vai trò: Thực thi code do LLM sinh ra (Python, SQL, v.v.) trong môi trường cách ly an toàn.
Đặc điểm: Cho phép mô hình thực hiện các hành động thực tế (**tool execution**) như chạy phân tích dữ liệu, tính toán, hoặc xử lý tệp.
→ Đảm bảo kết quả có thể tái kiểm chứng và không ảnh hưởng đến hệ thống chính.

KỸ THUẬT CỐT LÕI & CƠ CHẾ HOẠT ĐỘNG

INDIVIDUAL SYSTEMS

Re-Act (Reason + Act)

Chức năng chính: Cho phép mô hình **suy luận và hành động** trong cùng một vòng xử lý.

Cơ chế: LLM lập kế hoạch, gọi tool để thu thập hoặc xử lý dữ liệu, sau đó tổng hợp kết quả.

Ví dụ: Đọc tệp CSV → Sinh code Python → Thực thi → Trích xuất và tổng hợp kết quả.

Ưu điểm: Kết hợp reasoning và action trong một quy trình tự động, giảm sự phụ thuộc vào lập trình thủ công.

RAG (Retrieval-Augmented Generation)

Chức năng chính: **Truy xuất và kết hợp tri thức ngoài** trong quá trình sinh phản hồi.

Cơ chế: Mã hóa văn bản thành vector → tìm kiếm ngữ nghĩa trong kho tri thức → LLM sử dụng kết quả để tạo câu trả lời có căn cứ.

Ví dụ: Truy vấn thông tin nội bộ hoặc tài liệu kỹ thuật từ Vector Database.

Ưu điểm: Tăng độ chính xác và khả năng kiểm chứng của mô hình khi làm việc với dữ liệu lớn hoặc tài liệu chuyên ngành.

KỸ THUẬT CỐT LÕI & CƠ CHẾ HOẠT ĐỘNG

INDIVIDUAL SYSTEMS

LLM Agent

Chức năng chính: Tự điều phối và ra quyết định trong quá trình thực thi.

Cơ chế: Sử dụng LangChain / LangGraph để quản lý **workflow, memory, và tool-calling**.

Ví dụ: Khi nhận yêu cầu, Agent phân tích nhiệm vụ → chọn tool phù hợp (SQL, Python, OCR, v.v.) → hợp nhất kết quả.

Ưu điểm: Cho phép hệ thống hoạt động như một **AI đa tác vụ tự động**, có khả năng phối hợp linh hoạt giữa nhiều công cụ.

Try-Catch Loop

Chức năng chính: **Tự phát hiện và phục hồi lỗi** trong quá trình thực thi tác vụ.

Cơ chế: Áp dụng vòng lặp retry với kiểm tra điều kiện đầu ra và ghi log lỗi.

Ví dụ: Nếu một tool hoặc truy vấn thất bại, Agent tự khởi động lại bước đó với tham số hiệu chỉnh.

Ưu điểm: Giúp hệ thống duy trì tính ổn định và giảm thiểu lỗi trong các pipeline phức tạp.

KỸ THUẬT CỐT LÕI & CƠ CHẾ HOẠT ĐỘNG

INDIVIDUAL SYSTEMS

Chunking Strategy

Chức năng chính: **Xử lý hiệu quả tài liệu lớn vượt giới hạn context** của mô hình.

Cơ chế: Tách tài liệu theo trang, đoạn, hoặc logic nội dung; LLM đọc từng phần và tổng hợp kết quả cuối cùng.

Ví dụ: Phân tích báo cáo 200 trang bằng cách xử lý tuần tự từng chương rồi hợp nhất insight.

Ưu điểm: Tối ưu hóa hiệu năng xử lý và đảm bảo tính toàn vẹn ngữ nghĩa của tài liệu đầu vào.

Tích hợp kỹ thuật trong hệ thống

Các kỹ thuật trên kết hợp tạo thành một **hệ thống AI toàn diện**, có khả năng vừa suy luận, truy xuất, thực thi và tự phục hồi:

- **Re-Act + RAG**: Sinh truy vấn SQL → Phân tích kết quả → Sinh biểu đồ hoặc báo cáo.
- **LLM Agent + Try-Catch**: Tự động điều phối, giám sát và khôi phục lỗi trong quá trình xử lý.
- **Chunking**: Cho phép đọc hiểu và tóm tắt tài liệu quy mô lớn với độ chính xác cao.

Tổng thể: Mỗi kỹ thuật đóng vai trò trong **chuỗi pipeline thống nhất**, giúp tối ưu hiệu suất, độ tin cậy và khả năng mở rộng của hệ thống AI.

ỨNG DỤNG CHUYÊN BIỆT *

AI Data Analyst Assistant

Quy trình hoạt động

Phân tích yêu cầu	Text-to-SQL	Truy vấn CSDL	Phân tích dữ liệu	Vẽ biểu đồ	Tạo báo cáo

Kỹ thuật
- Re-Act
- LLM Agents
- Tool-calling
- Code generation

Mô hình
- Qwen3

ỨNG DỤNG CHUYÊN BIỆT *

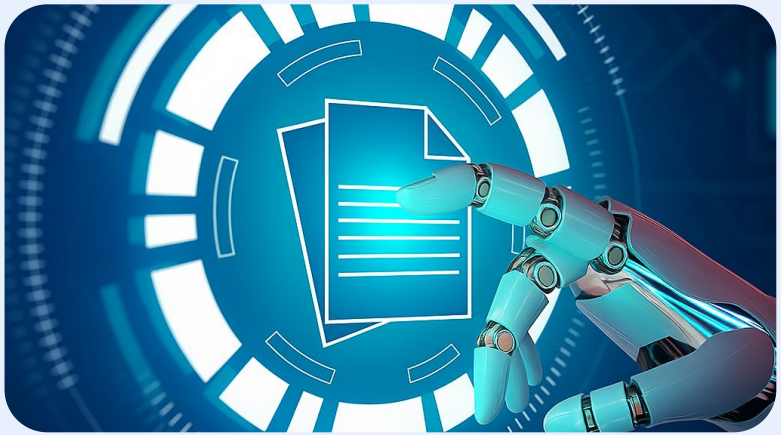
Document Reading Assistant

Chức năng chính

Phân tích dữ liệu đa định dạng (PDF, Word, Excel, CSV, PowerPoint, Text, Markdown)	Trích xuất cấu trúc (XML) và nội dung	Chuẩn hóa thông tin	Vector hóa tri thức	Khai thác tri thức

Kỹ thuật
- RAG
- VLM (OCR)
- Embedding

Mô hình
- Qwen3
- Qwen3-Embedding
- dots.ocr



ỨNG DỤNG CHUYÊN BIỆT *

Analytics Dashboard – Thống kê tương tác AI

Các chỉ số chính

Tổng dữ liệu	Tổng số lượt tương tác	Thống kê theo thời gian	Phân tách theo trợ lý	Tương tác theo thời gian từng trợ lý

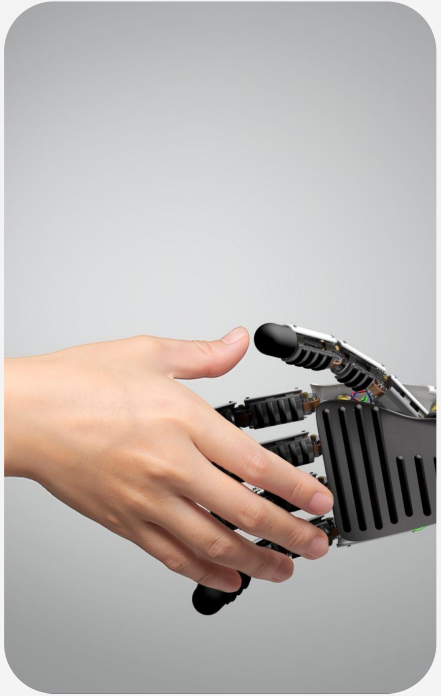
TƯƠNG TÁC VỚI NGƯỜI DÙNG

Multi-session Chat

- Hỗ trợ nhiều phiên hội thoại độc lập và liên tục.
- Duy trì ngữ cảnh xuyên suốt giữa các phiên làm việc (cross-session memory).
- Lưu trữ lịch sử hội thoại trong vector store để phục hồi ngữ cảnh dài hạn.
- Triển khai qua **LangGraph + Memory** cho quản lý trạng thái hội thoại.

Contextual QA

- Trò chuyện dựa trên dữ liệu và ngữ cảnh thực tế.
- Trò chuyện theo ngôn ngữ yêu cầu của người dùng.
- Áp dụng **RAG + Embedding** để truy xuất tri thức từ vector database.
- Hợp nhất kết quả tìm kiếm và phản hồi LLM để đảm bảo câu trả lời có căn cứ.
- Thích hợp cho chatbot chuyên biệt theo tài liệu, quy trình hoặc ngành nghiệp vụ.



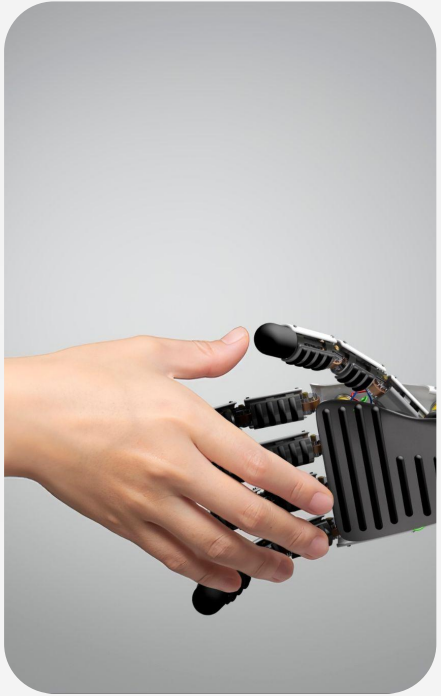
TƯƠNG TÁC VỚI NGƯỜI DÙNG

Voice Interaction

- Nhận dạng giọng nói và chuyển đổi **speech-to-text** theo thời gian thực.
- Sử dụng **Qwen3-Omni** để đảm bảo độ chính xác cao trên nhiều ngôn ngữ và tạp âm.
- Hỗ trợ input âm thanh trực tiếp, tương thích với giao diện hội thoại đa phương thức.

Role-Based Dialogue

- Quản lý hội thoại theo vai trò (role-based conversation).
- Áp dụng **LLM Role Control** để định nghĩa và duy trì hành vi của từng loại trợ lý.
- Hỗ trợ phân tách **AI kỹ thuật**, **AI nghiệp vụ**, và **AI hỗ trợ người dùng**.
- Cho phép mô hình phản hồi khác nhau tùy theo nhiệm vụ hoặc lĩnh vực được gán.



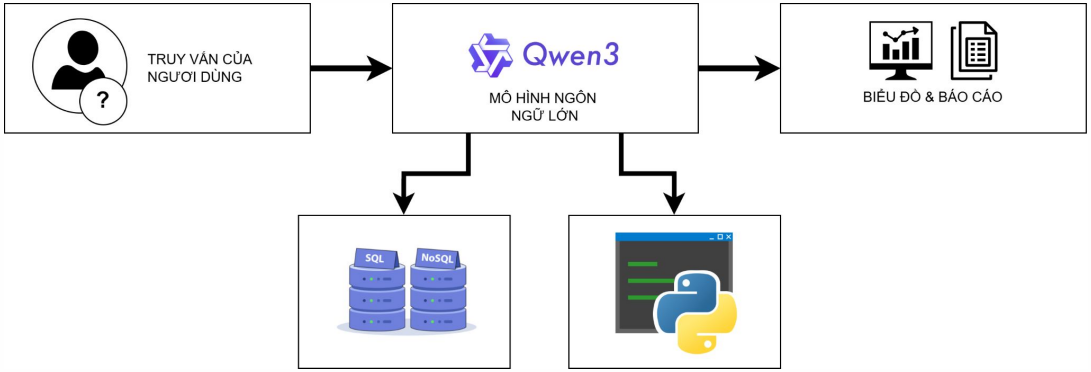
System Core Features

Core capabilities in data extraction, knowledge management, and intelligent content generation.

AI Data Analyst Assistant *

Ứng dụng AI chuyên biệt hỗ trợ **phân tích dữ liệu tự động**, **sinh mã Python**, **trực quan hóa kết quả** và **giải thích insight bằng ngôn ngữ tự nhiên**.

Được vận hành bởi **Re-Act reasoning**, **LangChain workflow orchestration**, và **Python runtime execution**, hệ thống giúp chuyển đổi dữ liệu thô của doanh nghiệp thành tri thức hành động — **không cần lập trình thủ công**.



Phân tích ngôn ngữ tự nhiên & Tạo sinh truy vấn

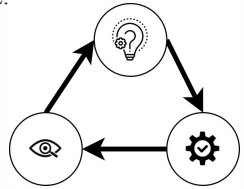
Hiểu yêu cầu truy vấn

Sử dụng Qwen3 phân tích ngữ nghĩa câu hỏi người dùng, xác định ý định và thực thể liên quan đến dữ liệu thật.

Truy cập schema database, metadata để nhận biết bảng, cột, kiểu dữ liệu, mối quan hệ và mô tả ngữ nghĩa

Tạo sinh truy vấn

Sử dụng Qwen3 dựng câu lệnh truy vấn thời gian thực, tự chọn phép tổng hợp, điều kiện lọc và cú pháp đúng cho từng loại CSDL.



Chu trình **Reason → Act → Observe → Repeat** giúp hệ thống suy luận, thử, kiểm chứng và tối ưu truy vấn tự động.

Thực thi truy vấn linh hoạt

Qwen3 chọn tool phù hợp cho từng nguồn:
- Database (PostgreSQL/MySQL//MongoDB)
- API

Kết quả được kiểm tra, chuẩn hóa và trả về dạng bảng thống nhất.

Phân tích dữ liệu, tạo biểu đồ & Giải thích Insight

Code Generation

Hệ thống sử dụng Qwen3 để sinh mã pandas, numpy, matplotlib phục vụ tính toán, lọc và tổng hợp dữ liệu theo yêu cầu truy vấn.

Mã được chạy trong môi trường Python cách ly, có cơ chế Try–Catch Loop để tự phát hiện, sửa lỗi và chạy lại khi cần.

Mã Python có thể vẽ trực tiếp biểu đồ (cột, đường, tròn, heatmap, time-series…) và tự chọn dạng biểu đồ phù hợp với đặc trưng dữ liệu.

Diễn giải bằng ngôn ngữ tự nhiên

Qwen3 đọc dữ liệu và biểu đồ đầu ra, tự động phân tích xu hướng, phát hiện bất thường và tổng hợp insight.

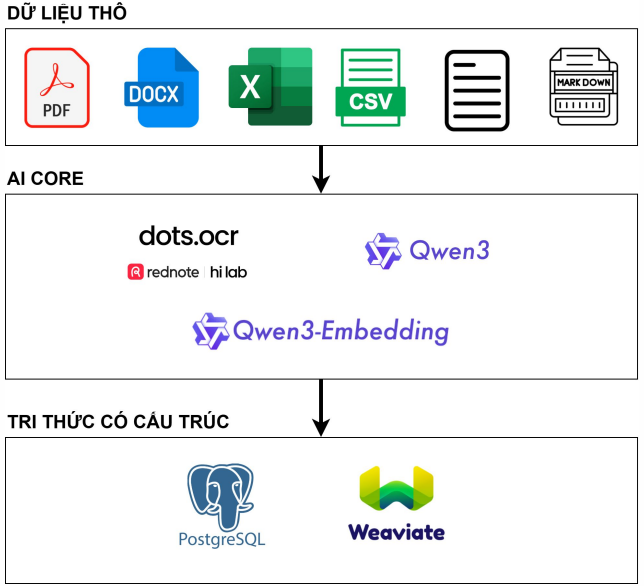
Tạo báo cáo đa ngôn ngữ để giúp người dùng ở nhiều quốc gia dễ dàng đọc hiểu và chia sẻ thông tin phân tích.

Báo cáo có thể xuất ra PDF, hoặc Markdown, kết hợp bảng dữ liệu, biểu đồ và phần nhận định tự động.

Document Reading Assistant

Ứng dụng AI chuyên biệt cho phép đọc, hiểu và trích xuất dữ liệu từ mọi loại tài liệu – bao gồm PDF, Word, Excel, CSV, PowerPoint, Text và Markdown.

Kết hợp sức mạnh của VLM (dots.ocr), LLM (Qwen3) và Embedding model (Qwen3-Embedding), hệ thống biến các tài liệu thô thành tri thức có cấu trúc, có thể tìm kiếm và phân tích trong kho tri thức doanh nghiệp.



Trích xuất nội dung đa định dạng

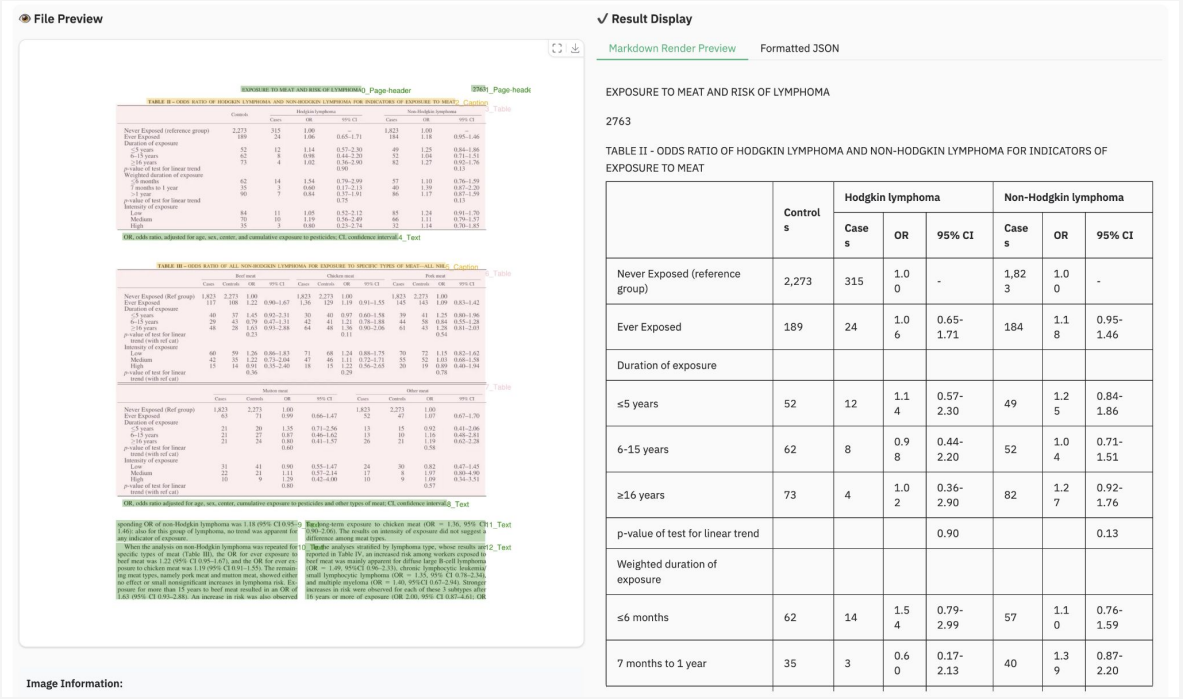
PDF & Hình ảnh – Xử lý bằng VLM dots.ocr

Hệ thống sử dụng Vision-Language Model (dots.ocr) để nhận dạng đồng thời văn bản, bảng, biểu đồ và bố cục trang.

Có khả năng phát hiện cấu trúc vùng đọc (layout detection), xác định cột, header, cell, và duy trì thứ tự đọc đúng ngữ cảnh.

Áp dụng OCR đa ngôn ngữ, xử lý tốt tài liệu scan hoặc ảnh chụp (biên lai, hợp đồng, hóa đơn).

Kết quả trả về văn bản có cấu trúc (Markdown, JSON), gồm: text content, vị trí, loại đối tượng (text/table/figure).



EXPOSURE TO MEAT AND RISK OF LYMPHOMA

2763

TABLE II - ODDS RATIO OF HODGKIN LYMPHOMA AND NON-HODGKIN LYMPHOMA FOR INDICATORS OF EXPOSURE TO MEAT

	Controls	Hodgkin lymphoma			Non-Hodgkin lymphoma		
		Cases	OR	95% CI	Cases	OR	95% CI
Never Exposed (reference group)	2,273	315	1.00	-	1,823	1.00	-
Ever Exposed	189	24	1.06	0.65-1.71	184	1.18	0.95-1.46
Duration of exposure							
≤5 years	52	12	1.14	0.57-2.30	49	1.25	0.84-1.86
6-15 years	62	8	0.98	0.44-2.20	52	1.04	0.71-1.51
≥16 years	73	4	1.02	0.36-2.90	82	1.27	0.92-1.76
p-value of test for linear trend				0.90			0.13
Weighted duration of exposure							
≤6 months	62	14	1.54	0.79-2.99	57	1.10	0.76-1.59
7 months to 1 year	35	3	0.60	0.17-2.13	40	1.39	0.87-2.20
>1 year	90	7	0.84	0.37-1.91	86	1.17	0.87-1.59
p-value of test for linear trend				0.75			0.13
Intensity of exposure							
Low	84	11	1.05	0.52-2.12	85	1.24	0.91-1.70
Medium	70	10	1.19	0.56-2.49	66	1.11	0.79-1.57
High	35	3	0.80	0.23-2.74	32	1.14	0.70-1.85

OR, odds ratio, adjusted for age, sex, center, and cumulative exposure to pesticides; CI, confidence interval

Trích xuất nội dung đa định dạng

Đọc và phân tích cấu trúc tài liệu Word (DOCX)

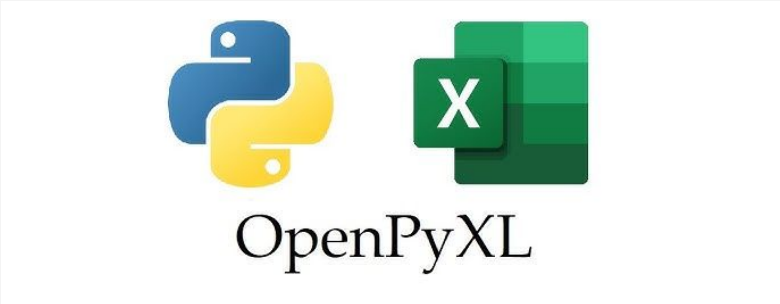
Hệ thống sử dụng thư viện python-docx để đọc trực tiếp nội dung, định dạng và cấu trúc của file Word:
- Trích xuất đoạn văn (paragraphs), tiêu đề (headings), bảng (tables), và mục liệt kê (lists).
- Bảo toàn thứ tự logic và phân cấp nội dung (heading → subheading → body).

Trích xuất nội dung đa định dạng

Đọc bảng dữ liệu và sheet Excel:

Hệ thống sử dụng thư viện openpyxl để mở và duyệt toàn bộ worksheet, hàng, cột, ô dữ liệu, công thức.

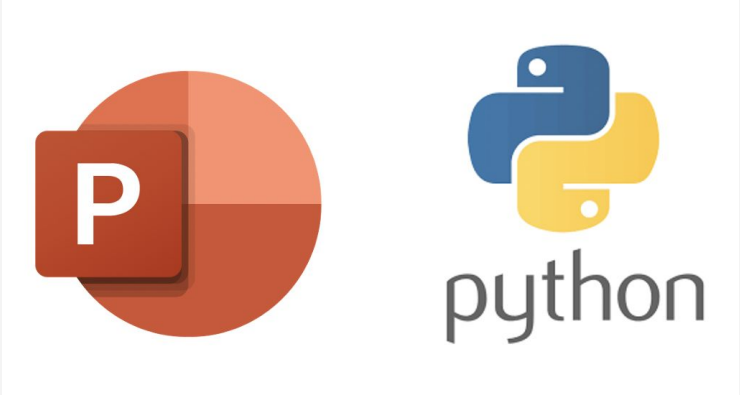
Tự động nhận diện header, sheet name, vùng dữ liệu. Xử lý nhiều sheet trong cùng một file. Có thể đọc cả công thức tính (cell.formula) và giá trị kết quả.



Trích xuất nội dung đa định dạng

Đọc dữ liệu PowerPoint:

Trích xuất nội dung PPTX bằng python-pptx:
- Đọc tiêu đề, bullet, bảng, textbox, ghi chú (notes) và thuộc tính layout của từng slide; duy trì thứ tự slide.
- Nhận diện cấu trúc & loại đối tượng: Phân loại Title / Subtitle / Body / Table / Picture / Shape; tái tạo cấu trúc logic (mục – tiểu mục – nội dung).



Trích xuất nội dung đa định dạng

Đọc dữ liệu csv:

Sử dụng thư viện pandas để mở và xử lý dữ liệu bảng từ file CSV.
- Tự động xác định delimiter (, ; | \t) và encoding.
- Nhận diện header, cột dữ liệu, và kiểu giá trị (tự động parse ngày, số, chuỗi).
- Hợp nhất nhiều tệp CSV theo mẫu hoặc nguồn chung.

Chuẩn hoá và làm sạch dữ liệu:
- Xử lý giá trị thiếu, khoảng trắng, ký tự ẩn, mã hóa sai.
- Chuẩn hóa tên cột (thành snake_case hoặc tiếng Việt chuẩn).
- Chuyển định dạng ngày, số, tiền tệ về chuẩn ISO hoặc numeric thống nhất.

Đọc dữ liệu txt, md:

Dữ liệu txt, md sẽ đọc trực tiếp bởi mô hình Qwen3

Tiền xử lý và chuẩn hóa dữ liệu đa định dạng

Hợp nhất pipeline xử lý cho mọi loại tài liệu

PDF, Word, Excel, CSV, PowerPoint, Markdown hay Text đi qua chuỗi tiền xử lý thống nhất trước khi đưa vào LLM.

Tự động nhận dạng và làm sạch nội dung

Áp dụng bộ quy tắc và mô hình chuẩn hóa riêng cho từng định dạng:

- PDF: loại watermark, số trang, chú thích.
- Word: bỏ style, header/footer, ký tự ẩn.
- Excel/CSV: xóa hàng trống, định dạng lại cột, số, ngày.
- Markdown/Text: làm sạch ký hiệu định dạng, đồng thời bảo toàn cấu trúc nội dung (tiêu đề, danh sách, bảng) để mô hình hiểu rõ mối quan hệ logic.

Xử lý ngữ cảnh đa ngôn ngữ và ký tự đặc biệt

- Chuẩn hóa encoding, dấu câu, viết hoa/thường, loại bỏ ký hiệu phi ngữ nghĩa, giữ nguyên biểu tượng hoặc công thức cần thiết.

Qwen3 tự động tóm tắt nội dung tài liệu:

- Mỗi file sau khi xử lý được Qwen3 tạo bản tóm tắt đa cấp, nêu rõ mục tiêu, nội dung chính và từ khóa trọng tâm.

Lưu trữ song song giữa bản gốc và bản tóm tắt:

- Bản tóm tắt được embedding riêng trong Vector DB, giúp tìm kiếm ngữ nghĩa nhanh hơn 3–5 lần so với việc truy vấn toàn văn bản.

Cải thiện độ chính xác khi truy vấn tài liệu liên quan:

- Khi người dùng đặt câu hỏi, hệ thống ưu tiên so khớp với embedding của bản tóm tắt, sau đó mở rộng sang các đoạn gốc => trả kết quả đúng chủ đề và ít nhiễu hơn.
- Tự động sinh metadata hỗ trợ truy vấn: Qwen3 sinh thêm keywords, entity (người, tổ chức, ngày, địa điểm)

=> tăng độ liên kết giữa các tài liệu liên quan.

Biểu diễn ngữ nghĩa văn bản bằng Qwen3-Embedding

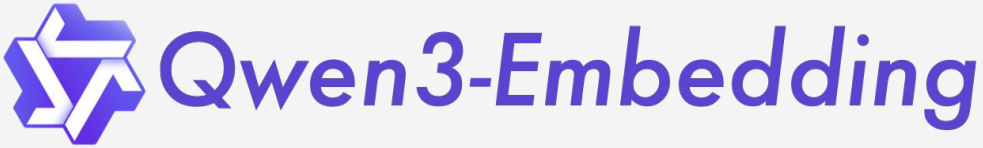
INDIVIDUAL SYSTEMS

Chuyển đổi văn bản thành vector ngữ nghĩa:

Mỗi đoạn văn (chunk) được Qwen3-Embedding mã hóa thành vector để biểu diễn ngữ nghĩa, ngữ cảnh và mối quan hệ giữa các đoạn nội dung.

Qwen3-Embedding được huấn luyện trên dữ liệu song ngữ (Anh – Trung – Việt – đa ngôn ngữ), giúp hệ thống tìm kiếm tri thức chính xác theo ý nghĩa, không phụ thuộc ngôn ngữ truy vấn.

Nhờ context window lên tới 32K tokens, mô hình có thể embedding toàn bộ đoạn dài hoặc trang tài liệu giúp giữ nguyên ngữ cảnh, giảm mất thông tin khi so sánh ngữ nghĩa.

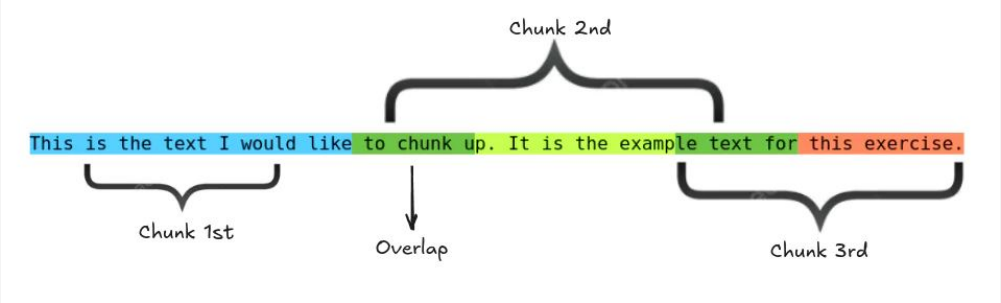


Phân tách văn bản thông minh – tối ưu cho mô hình embedding có ngữ cảnh lớn

Chiến lược chunking:

Mô hình Qwen3-Embedding hỗ trợ context window lớn (lên đến 32K tokens):
- Cho phép chunk theo phạm vi trang hoặc mục nội dung hoàn chỉnh, thay vì chia quá nhỏ gây mất ngữ cảnh.

Áp dụng Overlap Chunking:
Mỗi đoạn được chồng lấn 10–20% với đoạn kế tiếp để:
- Giữ mạch ngữ nghĩa liên tục.
- Tránh mất thông tin ở biên đoạn. Cải thiện chất lượng truy xuất.



Kho tri thức: Dữ liệu quan hệ và Vector

Weaviate – Kho tri thức ngữ nghĩa
Lưu trữ các đoạn văn bản đã được vector hóa cùng metadata (nguồn, loại, chủ đề, thời gian).
Hỗ trợ tìm kiếm ngữ nghĩa, hybrid search, và có thể **pre-compute vector** trước khi nhập để tăng tốc độ truy vấn.

PostgreSQL – Dữ liệu quan hệ
Quản lý liên kết giữa tài liệu, phiên bản, dự án, người dùng và quyền truy cập.
Đảm bảo tính toàn vẹn dữ liệu và hỗ trợ các truy vấn nghiệp vụ, báo cáo.

Hoạt động CRUD đồng bộ
- Create: Thêm tri thức mới vào Weaviate và ghi bản ghi liên kết trong PostgreSQL.
- Read: Truy vấn ngữ nghĩa trong Weaviate, kết hợp thông tin chi tiết từ PostgreSQL.
- Update: Cập nhật vector, metadata và quan hệ khi tài liệu thay đổi.
- Delete: Xóa tri thức và các bản ghi liên quan, đảm bảo dữ liệu sạch và đồng nhất.

Xếp hạng và lọc kết quả truy vấn
- Sử dụng vector similarity scoring để đo độ tương đồng giữa câu hỏi và nội dung.
- Hỗ trợ reranking đa tầng – ưu tiên theo độ liên quan, độ tin cậy, thời gian cập nhật, hoặc mức độ truy cập.
- Cho phép lọc theo metadata như loại tài liệu, người tạo, hoặc dự án để kết quả chính xác hơn.

Truy vấn hợp nhất
Kết quả tìm kiếm ngữ nghĩa được làm giàu bằng dữ liệu quan hệ, giúp người dùng xem được nội dung, ngữ cảnh và mối liên kết đầy đủ.

Hàng đợi xử lý bất đồng bộ – đảm bảo hiệu năng và khả năng mở rộng

INDIVIDUAL SYSTEMS

Cơ chế xử lý bất đồng bộ

Hệ thống sử dụng hàng đợi tác vụ (message queue) để xử lý dữ liệu theo nền, giúp các bước trích xuất, chuẩn hóa, embedding và lưu trữ diễn ra song song và không chặn API người dùng.

Kiến trúc triển khai

Mô hình Producer – Queue – Consumer gồm:

- Producer: FastAPI gửi yêu cầu xử lý tài liệu.
- Queue: Redis làm message broker lưu và phân phối tác vụ.
- Consumer: Các Celery Worker xử lý lần lượt và cập nhật kết quả về cơ sở dữ liệu.

Bộ công nghệ sử dụng Celery + Redis

- Celery quản lý tác vụ nền, retry khi lỗi và ghi log chi tiết.
- Redis đóng vai trò hàng đợi trung gian, đảm bảo truyền tải an toàn và nhanh chóng.
- Flower dashboard hỗ trợ giám sát tiến trình trực quan.

Ưu điểm

- Xử lý song song hàng trăm tài liệu mà không nghẽn hệ thống.
- Giao diện người dùng luôn phản hồi tức thời.
- Dễ mở rộng thêm worker khi tải tăng.
- Dữ liệu an toàn, có cơ chế retry và logging đầy đủ.

Analytics Dashboard – Thống kê tương tác AI

Bảng điều khiển phân tích giúp giám sát toàn bộ hoạt động của hệ thống AI, theo dõi dữ liệu, mức độ tương tác và hiệu suất của từng trợ lý trí tuệ nhân tạo.

Thống kê tổng quan hệ thống

Tổng hợp toàn bộ dữ liệu hiện có: số lượng tài liệu, tri thức đã lưu, vector embedding và số người dùng hoạt động. Giúp nhà quản trị nắm rõ quy mô và tốc độ phát triển kho tri thức AI.

Thống kê tổng số lượt tương tác

Ghi nhận tổng số yêu cầu, câu hỏi và phản hồi giữa người dùng và các trợ lý AI trong hệ thống, phản ánh mức độ sử dụng thực tế.

Biểu đồ xu hướng tương tác theo thời gian

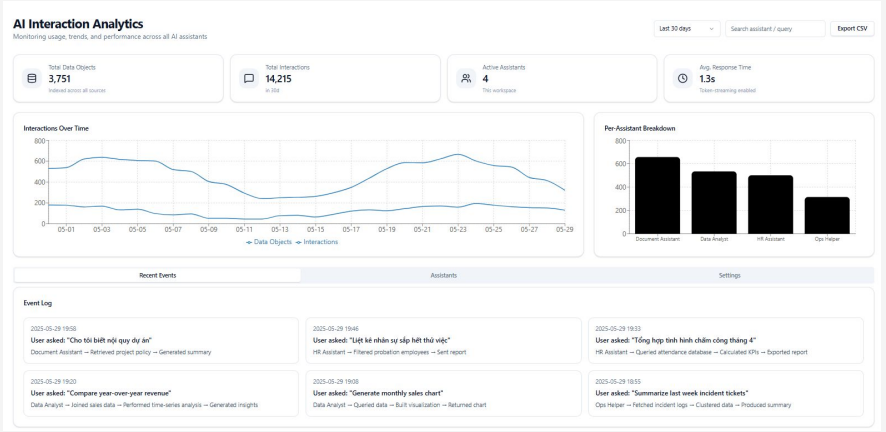
Biểu đồ động hiển thị sự thay đổi của lượt truy vấn qua các mốc thời gian (ngày, tuần, tháng), giúp xác định thời điểm hệ thống hoạt động cao nhất.

Phân tích dữ liệu từng trợ lý AI

Mỗi trợ lý như Document Assistant, Data Analyst hay HR Assistant đều có khu vực thống kê riêng, thể hiện lượng dữ liệu xử lý và mức độ tương tác.
Đo lường số lượng truy vấn, phản hồi thành công, thời gian xử lý trung bình để đánh giá hiệu quả và mức độ phổ biến của từng trợ lý.

Biểu đồ chi tiết theo từng trợ lý

Hiển thị biểu đồ tương tác riêng biệt, giúp theo dõi xu hướng hoạt động của từng trợ lý AI theo thời gian thực và hỗ trợ tối ưu năng lực hệ thống.



AI System Architecture & Technical Specification

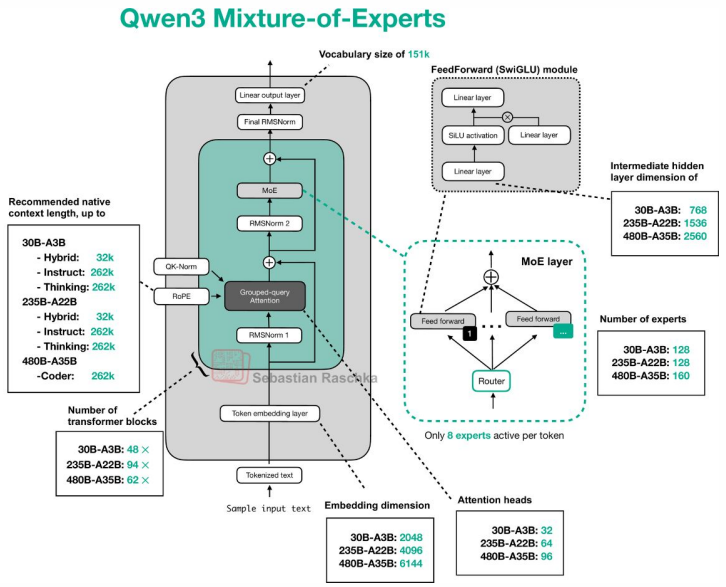
Phân tích chi tiết mô hình, kỹ thuật và luồng xử lý trong giải pháp AI tích hợp.

QWEN3 - MÔ HÌNH NGÔN NGỮ LỚN

Mô hình: Transformer Decoder (các version dense và MoE).
Thành phần chính:
- Grouped Query Attention (GQA) – tăng hiệu năng suy luận.
- SwiGLU Activation – cải thiện độ ổn định huấn luyện.
- Rotary Position Embedding (RoPE) – mã hoá vị trí chính xác hơn.
- RMSNorm / QK-Norm – chuẩn hoá ổn định gradient khi training.
- Tokenizer (Tách từ): Byte-level BPE với vocab ~151K (đa ngôn ngữ, tối ưu token efficiency).

Cơ chế Mixture of Experts (MoE):
- Chia chuyên gia (expert) theo token → 8 expert hoạt động mỗi bước suy luận.
- Dùng Global Batch Load Balancing để tối ưu phân phối token giữa các expert.
- Cho phép specialization (chuyên biệt hóa từng expert cho loại nhiệm vụ khác nhau).

Đặc điểm nổi bật kiến trúc Qwen3
- Hiệu quả tính toán cao: GQA + QK-Norm giảm độ phức tạp inference.
- Ổn định khi huấn luyện: RMSNorm + SwiGLU cải thiện hội tụ mô hình lớn.
- Context mở rộng: lên đến 128K tokens, phù hợp xử lý tài liệu dài.
- Hỗ trợ reasoning và multi-format I/O: thiết kế attention ổn định cho tác vụ logic và multi-task.

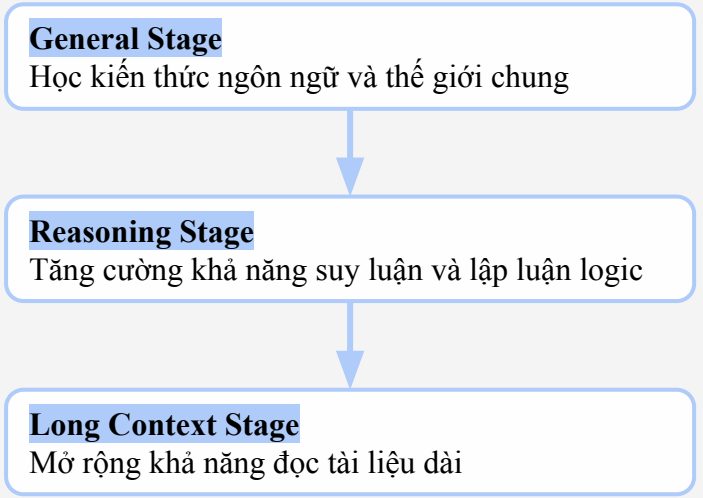


QWEN3 - PRE-TRAINING

Kích cỡ dữ liệu: 36 trillion tokens
Bao phủ 119 ngôn ngữ và phương ngữ, giúp tăng năng lực đa ngữ và hiểu ngữ cảnh chéo ngôn ngữ.

Dữ liệu huấn luyện gồm:
- Văn bản tự nhiên, tài liệu đa miền (news, wiki, books, dialogs).
- STEM & Coding: bài toán kỹ thuật, snippet code, reasoning dataset.
- Multilingual & Synthetic Texts: dữ liệu mô phỏng đa miền (reasoning, QA, instruction, math, code).
- Bổ sung thêm dữ liệu trích xuất từ PDF, textbook, instructions, code

Năng lực mô hình:
- Đọc hiểu dữ liệu dài
- Suy luận và logic
- Sinh mã và phân tích code
- Đa ngôn ngữ và dịch ngữ nghĩa

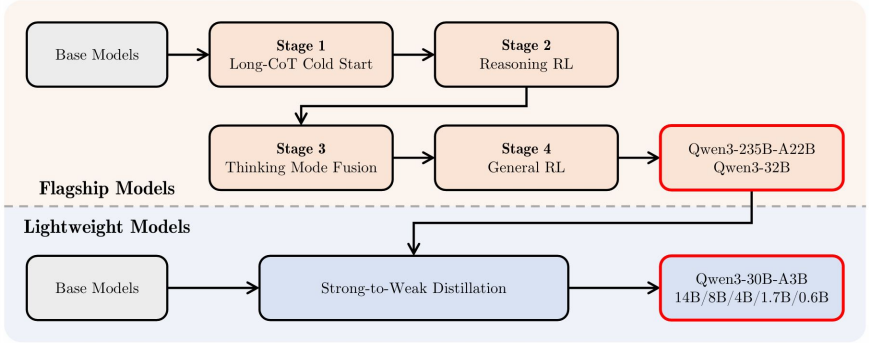


QWEN3 - POST-TRAINING

Hướng huấn luyện chính:

- Thinking Mode: Người dùng có thể bật/tắt chế độ suy luận sâu hoặc điều chỉnh "mức độ suy nghĩ" thông qua token budget.

- Strong-to-Weak Distillation: Dùng output logit từ flagship model (235B/32B) để huấn luyện lightweight model (30B/14B/8B...), giúp tăng hiệu năng mà giảm chi phí GPU.



QWEN3 - BENCHMARK

Qwen3 đạt được kết quả cạnh tranh trong các bài đánh giá chuẩn về khả năng lập trình, toán học và năng lực tổng quát, khi so sánh với các mô hình hàng đầu khác như DeepSeek-R1, o1, o3-mini, Grok-3 và Gemini-2.5-Pro.

	Qwen3-30B-A3B MoE	QwQ-32B	Qwen3-4B Dense	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o 2024-11-20
ArenaHard	91.0	89.5	76.6	81.2	86.8	85.5	85.3
AIME'24	80.4	79.5	73.8	18.9	32.6	39.2	11.1
AIME'25	70.9	69.5	65.6	15.0	24.0	28.8	7.6
LiveCodeBench v5, 2024.10-2025.02	62.6	62.7	54.2	30.7	26.9	33.1	32.7
CodeForces Elo Rating	1974	1982	1671	859	1063	1134	864
GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
LiveBench 2024-11-25	74.3	72.0	63.6	51.4	49.2	60.5	52.2
BFCL v3	69.1	66.4	65.9	63.4	59.1	57.6	72.5
MultiIF 8 Languages	72.2	68.3	66.3	65.3	69.8	55.6	65.6

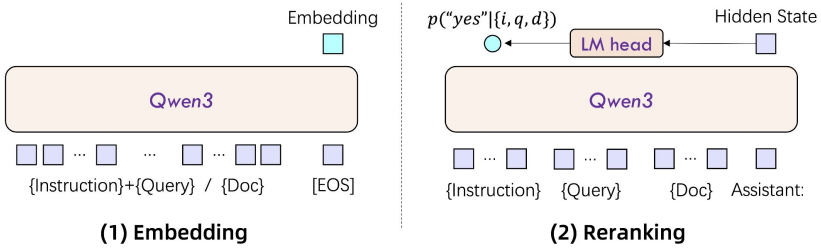
1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

QWEN3 - EMBEDDING MODEL

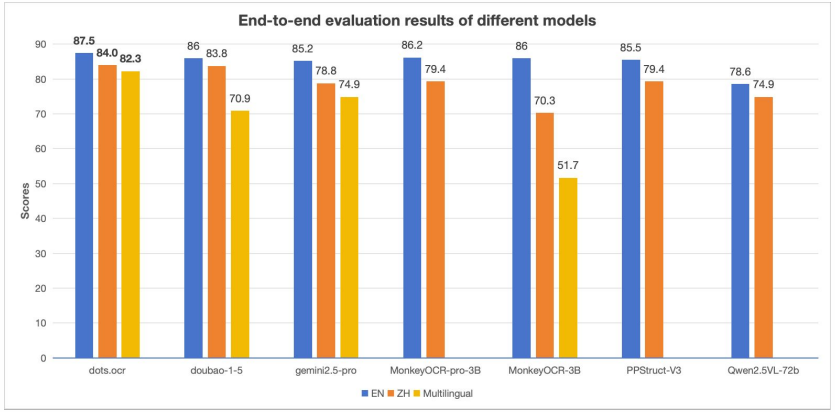
Loại mô hình: Embedding & Reranking – thuộc họ Qwen3.
Nhiệm vụ: Sinh vector ngữ nghĩa (semantic vector) cho truy vấn, văn bản, mã nguồn.
Điểm nổi bật:
- Kế thừa năng lực hiểu ngôn ngữ đa miền và đa ngữ của Qwen3, đạt SOTA trên nhiều benchmark embedding & retrieval.
- Hỗ trợ hơn 100 ngôn ngữ, bao gồm ngôn ngữ lập trình, cho phép tìm kiếm và mã hóa ngữ nghĩa xuyên ngôn ngữ.



	Qwen3-Embedding-8B	Qwen3-Embedding-4B	Qwen3-Embedding-0.6B	Gemini Embedding	Cohere-embed-multilingual-v3.0	text-embedding-3-large	multilingual-e5-large-instruct	gte-Qwen2-7B-instruct
MMTEB Mean-Task	70.58	69.45	64.33	68.37	61.12	58.93	63.22	62.51
MTEB (en v2) Mean-Task	75.22	74.60	70.70	73.30	66.01	66.43	65.53	70.72
MTEB-Code	80.68	80.06	75.41	74.66	51.94	58.95	65.00	56.41

dots.ocr

Mô hình: Vision–Language Model (VLM) với 1.7B tham số.
Chức năng: Kết hợp **phát hiện bố cục (layout detection)** và **nhận dạng nội dung (content recognition)** trong cùng một mô hình.
Điểm nổi bật:
- Duy trì **thứ tự đọc tự nhiên** và trích xuất chính xác văn bản, bảng biểu, công thức, hình ảnh từ tài liệu scan.
- Đạt kết quả SOTA trên OmniDocBench cho nhận dạng bảng và công thức.



dots.ocr

👁 File Preview

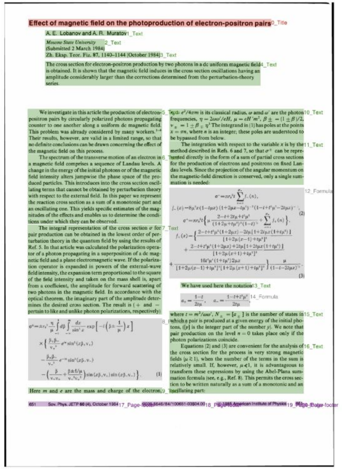


Image Information:

✔ Result Display

Markdown Render Preview | Formatted JSON

amplitude for forward scattering of two photons in the magnetic field. In accordance with the optical theorem, the imaginary part of the amplitude determines the desired cross section. The result is (+ and — pertain to like and unlike photon polarizations, respectively)

$$\sigma^{\pm} = \pm r_0^2 \frac{\eta}{\mu} \int_{-1}^{1} d\beta \int_{-\infty}^{\infty} \frac{dx}{\sin^2 x} \exp\left[-i\left(\beta \pm \frac{1}{\mu}\right)x\right] \times \left\{ \begin{array}{l} \frac{\beta_+\beta_-}{v_+^2} e^{ix} \sin^2(x\beta_+ v_+) \\ - \frac{\beta_+\beta_-}{v_-^2} e^{-ix} \sin^2(x\beta_- v_-) \\ - \left(\frac{\beta}{v_- v_+} + \frac{\beta \pm 1/\mu}{v_-^2 v_+^2}\right) \sin(x\beta_+ v_+) \sin(x\beta_- v_-) \end{array} \right\}. \quad (1)$$

Here m and e are the mass and charge of the electron,

$r_0 = e^2/4\pi m$ is its classical radius, ω and ω' are the photon frequencies, $\eta = 2\omega\omega'/eH$, $\mu = eH/m^2$, $\beta_{\pm} = (1 \pm \beta)/2$, $v_{\pm} = 1 \pm \beta_{\mp}\eta$. The integrand in (1) has poles at the points $x = \pi n$, where n is an integer; these poles are understood to be bypassed from below.

The integration with respect to the variable x is by the method described in Refs. 6 and 7, so that $\sigma^{\pm}$ can be represented directly in the form of a sum of partial cross sections for the production of electrons and positrons on fixed Landau levels. Since the projection of the angular momentum on the magnetic-field direction is conserved, only a single summation is needed:

$$\sigma^- = \pi r_0^2 t \sum_{n=1}^{N^-} f^-(n),$$
$$f^-(x) = 8\mu^2 x(1 - t\mu x)(1 + 2\mu x - t\mu^2)^{-2}(1 - t + t^2\mu^2 - 2t\mu x)^{-1},$$
$$\sigma^+ = \pi r_0^2 t \left\{ \mu \frac{2 - t + 2t\mu + t^2\mu^2}{(1 + 2\mu + t\mu^2)^2(1 - t)^{1/2}} + \sum_{n=1}^{N^+} f^+(n) \right\}, \quad (2)$$
$$f^+(x) = \left\{ \frac{2 - t + t^2\mu^2(1 + 2\mu x) - 2t\mu[1 + 2t\mu x(1 + t\mu^2)]}{[1 + 2\mu(x-1) + t\mu^2]^2} + \frac{2 - t + t^2\mu^2(1 + 2\mu x) + 2t\mu[1 + 2t\mu x(1 + t}{[1 + 2\mu(x+1) + t\mu^2]^2} \right.$$

We have used here the notation

$$a_+ = \frac{1 - t}{2t\mu}, \quad a_- = \frac{1 - t + t^2\mu^2}{2t\mu},$$

where $t = m^2/\omega\omega'$, $N_{\pm} = [a_{\pm}]$ is the number of states in which a pair is produced at a given energy of the initial

dots.ocr

File Preview

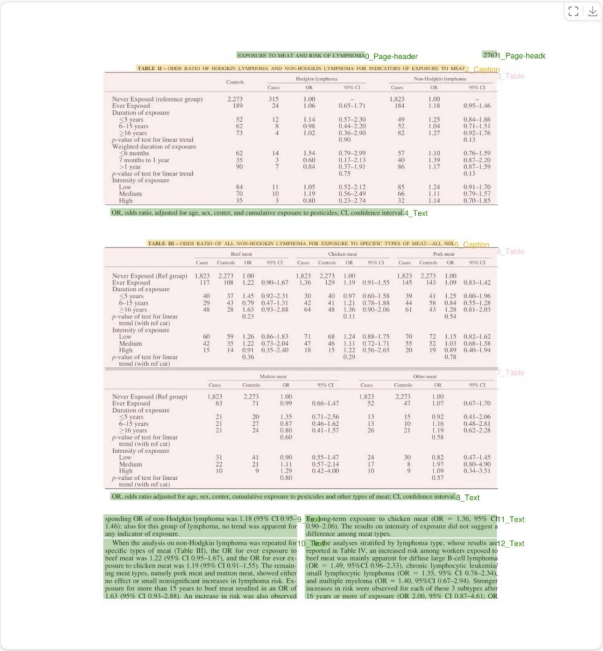


Image Information:

✓ Result Display

Markdown Render Preview | Formatted JSON

EXPOSURE TO MEAT AND RISK OF LYMPHOMA

2763

TABLE II - ODDS RATIO OF HODGKIN LYMPHOMA AND NON-HODGKIN LYMPHOMA FOR INDICATORS OF EXPOSURE TO MEAT

	Controls	Hodgkin lymphoma			Non-Hodgkin lymphoma		
		Cases	OR	95% CI	Cases	OR	95% CI
Never Exposed (reference group)	2,273	315	1.00	-	1,823	1.00	-
Ever Exposed	189	24	1.06	0.65-1.71	184	1.18	0.95-1.46
Duration of exposure							
≤5 years	52	12	1.14	0.57-2.30	49	1.25	0.84-1.86
6-15 years	62	8	0.98	0.44-2.20	52	1.04	0.71-1.51
≥16 years	73	4	1.02	0.36-2.90	82	1.27	0.92-1.76
p-value of test for linear trend				0.90			0.13
Weighted duration of exposure							
≤6 months	62	14	1.54	0.79-2.99	57	1.10	0.76-1.59
7 months to 1 year	35	3	0.60	0.17-2.13	40	1.39	0.87-2.20
≥1 year	90	7	0.84	0.37-1.91	86	1.17	0.87-1.59
p-value of test for linear trend				0.75			0.13
Intensity of exposure							
Low	84	11	1.05	0.52-2.12	85	1.24	0.91-1.70
Medium	70	10	1.19	0.56-2.49	66	1.11	0.79-1.57
High	35	3	0.80	0.23-2.74	32	1.14	0.70-1.85

SPEECH TO TEXT

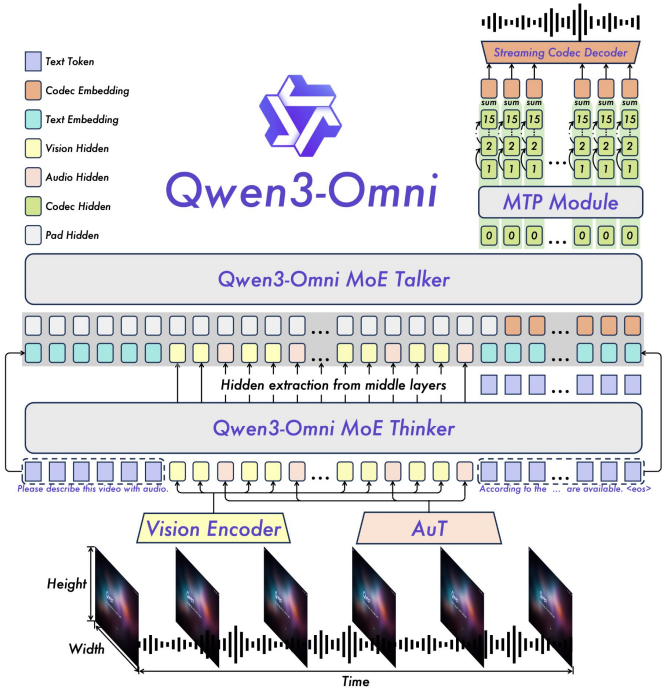
Qwen3-Omni là mô hình đa phương thức (text-image-audio-video) của Alibaba Cloud, cho phép **nghe, hiểu và phản hồi bằng giọng nói thời gian thực**.

Kiến trúc Thinker–Talker:

- *Thinker* → hiểu và suy luận ngữ nghĩa.
- *Talker* → sinh giọng nói & xử lý âm thanh thời gian thực.
- Hỗ trợ **streaming low-latency**, mã hóa âm thanh bằng multi-codebook codec.

Khả năng Speech-to-Text:

- Nhận dạng **19 ngôn ngữ** (gồm tiếng Việt, Anh, Trung, Nhật…).
- Hiệu năng **SOTA trên 32/36 benchmark audio-visual** (theo báo cáo Qwen3-Omni 2025).
- Xử lý tốt âm thanh hội thoại, môi trường, podcast.
- Độ trễ thấp, **transcribe gần thời gian thực**.

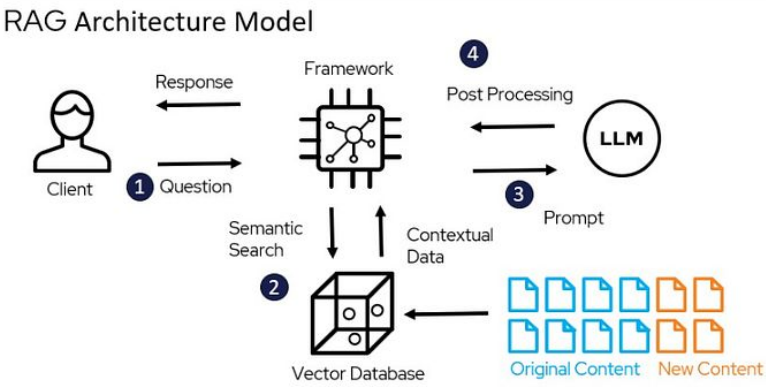


RAG – Retrieval Augmented Generation

Khái niệm: RAG là kiến trúc kết hợp **truy xuất tri thức (Retrieval)** với **sinh ngôn ngữ (Generation)**.
→ LLM không chỉ dựa vào kiến thức đã huấn luyện, mà còn **truy xuất dữ liệu bên ngoài (vector store, database, tài liệu nội bộ)** để sinh ra câu trả lời chính xác và có căn cứ.

Mục tiêu: Giúp mô hình **cập nhật tri thức động**, giảm "ảo tưởng thông tin" (hallucination) và tối ưu khả năng trả lời theo ngữ cảnh doanh nghiệp.

Kiến trúc triển khai
- **Embedding Model:** Qwen3-Embedding-0.6B mã hóa văn bản thành vector ngữ nghĩa.
- **Vector Database:** Weaviate lưu trữ và truy vấn vector bằng similarity search.
- **LLM Generator:** Sinh câu trả lời có dẫn chứng, tích hợp logic suy luận từ Qwen3.



https://cubed.run/blog/rag-2-0-finally-getting-retrieval-augmented-generation-right-05a067927589

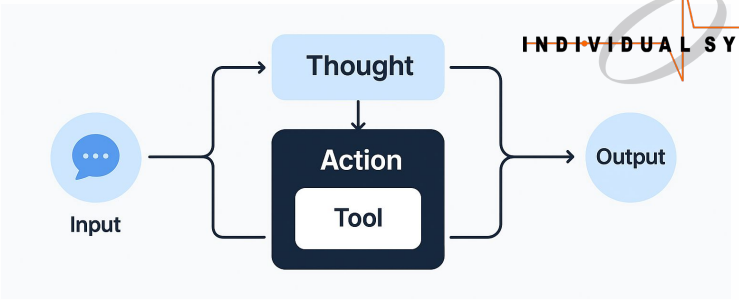
RE-ACT LLM AGENTS

Khái niệm: ReAct (Reason + Act) là cơ chế cho phép LLM **vừa suy luận (reasoning)** vừa **thực hiện hành động (acting)** thông qua tool-calling có điều kiện.

Mục tiêu: Biến LLM từ mô hình sinh văn bản thụ động thành **agent chủ động**, có thể lập kế hoạch, truy xuất, và ra quyết định tuần tự.

Kiến trúc triển khai

- **Reasoning Engine:** Qwen3 (LLM) thực hiện phân tích logic & ra quyết định hành động.
- **Action Layer:** Các công cụ được định nghĩa sẵn (tool registry) – SQL Query, Python Executor, Search API, OCR, v.v.
- **Memory / State:** LangChain quản lý ngữ cảnh, chuỗi Thought–Action–Observation.
- **Control Flow:** Cấu trúc loop có điều kiện cho phép agent dừng, retry, hoặc chuyển sang tool khác khi lỗi.



(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys]

Thank you

INDIVIDUAL SYSTEMS

Website
indivisys.vn