



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Retrieval-Augmented Generation for Industrial Applications and PHM

Từ LLM tổng quát đến trợ lý tri thức công nghiệp đáng tin cậy

ONE LOVE. ONE FUTURE.

- **PhD in Mechanical Engineering**

Seoul National University (SNU), Korea (2020 – 2025)

Focus: DNA nanotech, GNNs, physics-informed ML

- **MS & BSc in Mechatronics**

Hanoi University of Science and Technology (HUST), Vietnam

MS Thesis: Nonlinear oscillation in Duffing systems

BSc Thesis: Fractional-order Duffing oscillator

- **Teaching & Research**

Lecturer, HUST (2018–2020, 2025–Now)

Mechanical Engineer (2015–2016)



Trương Quốc Chiến

School of Mechanical Engineering (SME)

Hanoi University of Science and Technology

chien.truongquoc@hust.edu.vn

1. BỐI CẢNH VÀ ĐỘNG LỰC

Vì sao công nghiệp cần AI dựa trên tri thức?



Dữ liệu giàu nhưng phân tán

Thủ công, báo cáo lỗi, log bảo trì, SOP nằm rải rác ở nhiều hệ thống khác nhau.



Tri thức nằm trong đầu chuyên gia

Phần lớn kinh nghiệm xử lý sự cố ít được ghi chép, gặp khó khăn khi có thay đổi nhân sự.



Kỹ sư cần câu trả lời nhanh

Khi thiết bị báo lỗi lúc 2h sáng, người trực cần biết ngay nguyên nhân và cách xử lý.



Tìm thủ công là quá chậm

Lục hàng nghìn trang tài liệu bằng tay không khả thi trong vận hành thực tế.

1. BỐI CẢNH VÀ ĐỘNG LỰC

Vấn đề của LLM thuần túy trong công nghiệp



LLM thuần túy

"Siết bu-lông đến 95 Nm..."

→ nghe thuyết phục nhưng có thể sai

"Khả dĩ nhất" ≠ "Đúng nhất"



Dễ hallucinate (bịa thông tin)

Khi không có dữ liệu nền, LLM sinh chuỗi văn bản khả dĩ nhất, không phải đúng nhất.



Mù tịt tài liệu nội bộ

Không biết manual riêng và thông tin mới nhất của doanh nghiệp.



Khó dùng cho an toàn cao

Một bước sai trong môi trường công nghiệp có thể gây hỏng thiết bị hoặc tai nạn.

1. BỐI CẢNH VÀ ĐỘNG LỰC

Khoảng trống giữa AI tổng quát và AI công nghiệp



AI tổng quát (LLM)

Hiểu và diễn đạt ngôn ngữ rất tốt, nhưng không biết tri thức riêng của bạn.

RAG

cây cầu tri thức



Tri thức công nghiệp

Chuyên ngành, quy trình, tiêu chuẩn, dữ liệu lịch sử, cần cập nhật & truy vết.

Yêu cầu thực tế: trả lời chính xác theo domain · cập nhật thường xuyên · truy vết được · bảo mật dữ liệu

2. RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG là gì? — và giải quyết vấn đề gì?

RAG = Retrieval-Augmented Generation

*LLM sinh câu trả lời **có tăng cường bằng truy xuất tri thức**. Thay vì chỉ dựa vào 'trí nhớ' của LLM, hệ thống tìm tài liệu liên quan trước, rồi mới trả lời.*



Giảm hallucination

Câu trả lời bám vào bằng chứng đã truy xuất.



Tăng tính cập nhật

Chỉ cập nhật kho tài liệu, không cần huấn luyện lại.



Dùng được tài liệu nội bộ

Khai thác cả dữ liệu riêng và nhạy cảm của doanh nghiệp.



Trả lời có trích dẫn

Minh bạch, truy vết được tới đúng nguồn tài liệu.

2. RETRIEVAL-AUGMENTED GENERATION (RAG)

Ví dụ trực quan: hỏi về rung bất thường ở vòng bi



RAG không 'đoán' — nó truy xuất tài liệu thật rồi đề xuất checklist để kỹ sư tự kiểm tra và quyết định.

2. RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG giống một kỹ sư với đầy đủ thư viện kỹ thuật



LLM

↔ người có khả năng đọc hiểu và diễn giải



Vector database

↔ thư viện được lập chỉ mục theo ý nghĩa



Retrieval

↔ quá trình tìm đúng tài liệu trên kệ



Generation

↔ quá trình viết câu trả lời tổng hợp

3. PIPELINE CƠ BẢN

Tổng quan pipeline RAG: offline + online

OFFLINE · LẬP CHỈ MỤC



Thu thập
tài liệu

>



Chia nhỏ
(chunk)

>



Embedding>



Vector
database

ONLINE · TRUY XUẤT & SINH TRẢ LỜI



Câu hỏi →
embedding>



Tìm đoạn
gần nhất

>



Rerank
top-n

>



LLM sinh +
citation

3. PIPELINE CƠ BẢN

Document ingestion & Chunking



Document ingestion

Đưa tài liệu vào hệ thống

- Nguồn: PDF, Word, Excel, manual, report, log, website nội bộ
- PHM bổ sung: maintenance report, CMMS, historian, SOP
- Cần làm sạch & chuẩn hóa trước khi xử lý
- RAG vào thì RAG ra — chất lượng đầu vào quyết định đầu ra



Chunking

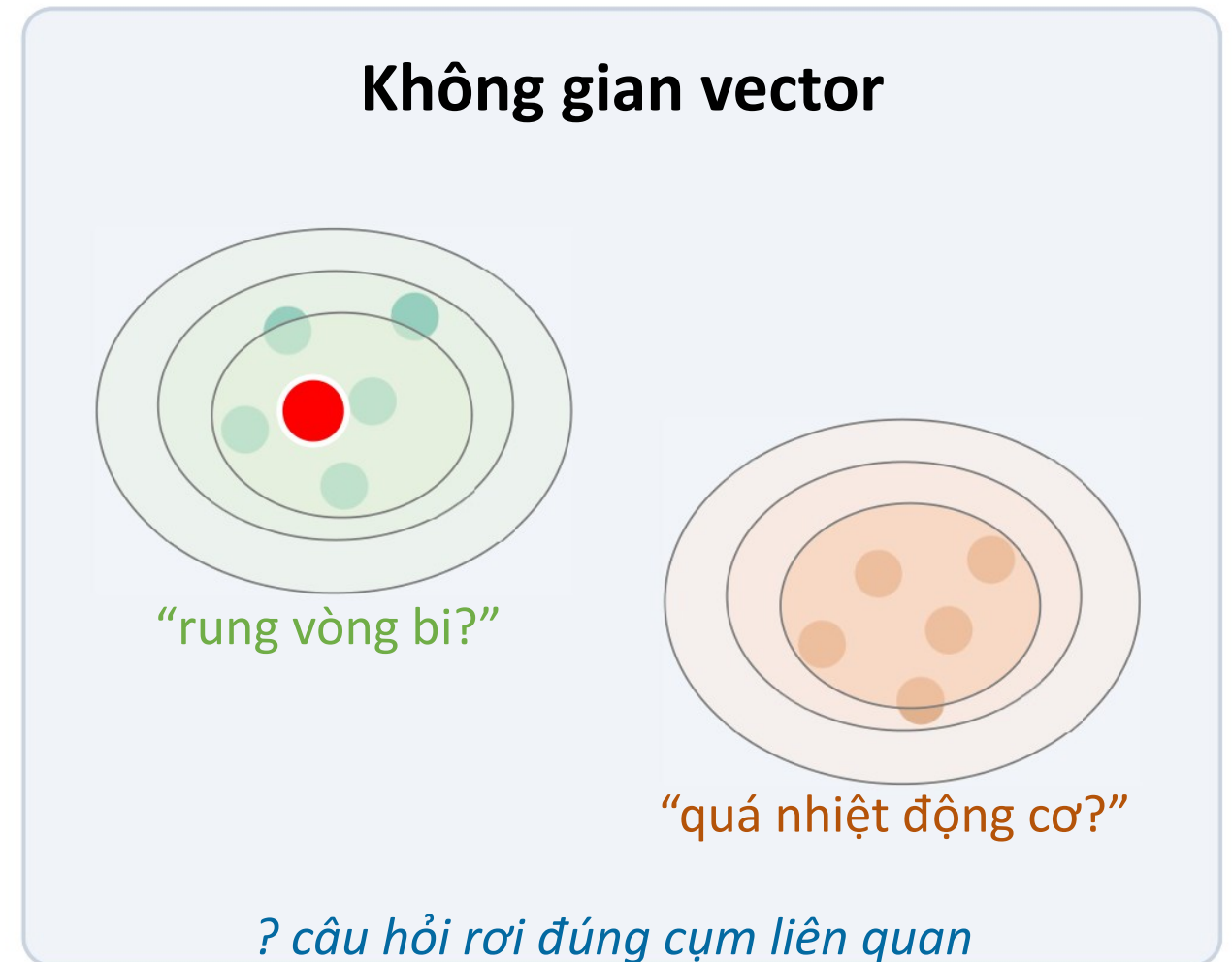
Chia tài liệu dài thành các đoạn

- Quá ngắn → mất ngữ cảnh; quá dài → giảm độ chính xác
- Kỹ thuật: sliding window, fine-grained, Small-2-Big
- Giữ metadata: tên máy, loại lỗi, thời gian, nguồn
- Metadata giúp lọc đúng thiết bị khi truy xuất

3. PIPELINE CƠ BẢN

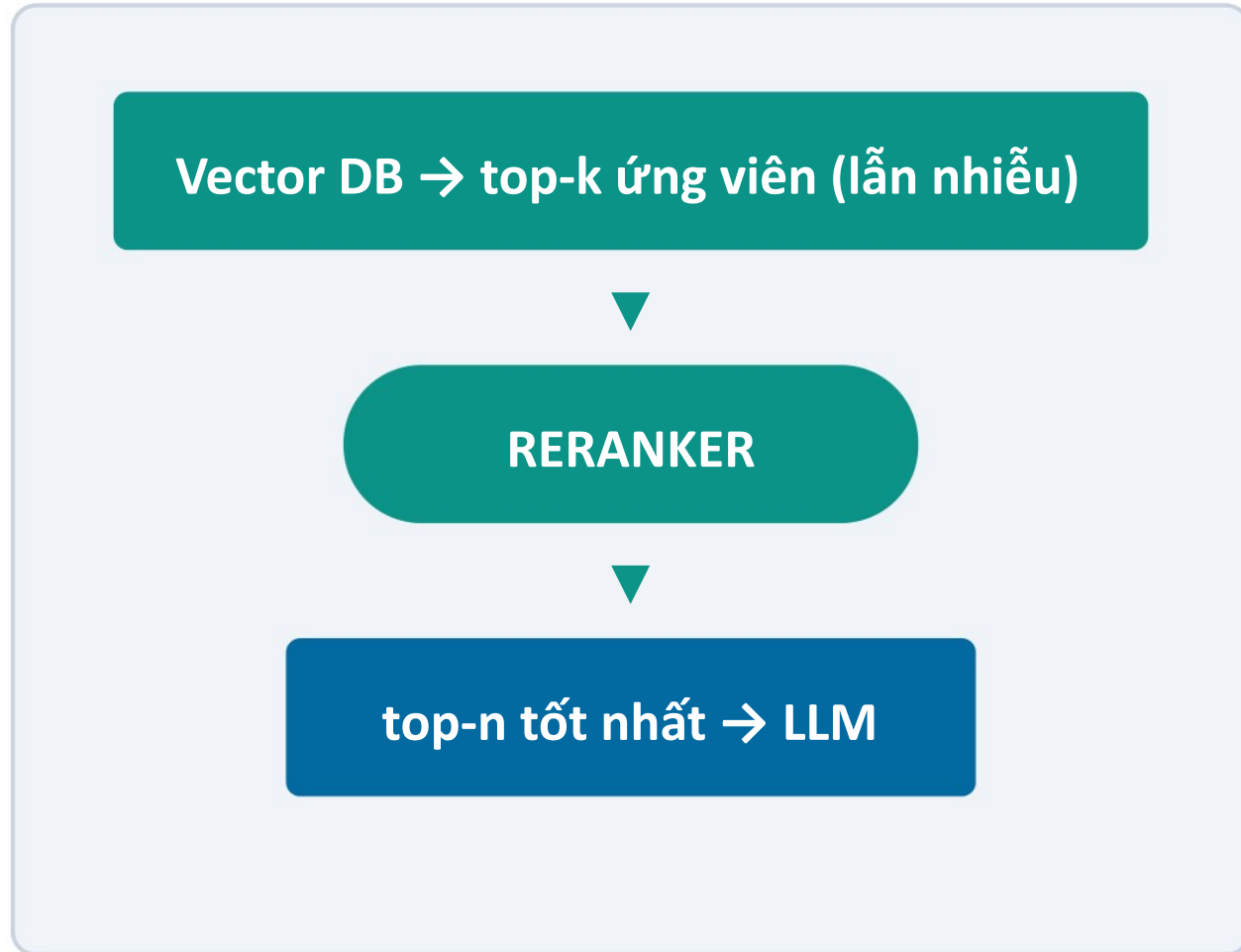
Embedding và Vector Database

- Embedding biến văn bản thành vector số
- Đoạn cùng ý nghĩa → vector nằm gần nhau
- Mô hình hiện đại (BERT, RoBERTa, BGE) cho biểu diễn theo ngữ cảnh
- Vector DB (Chroma, Pinecone, FAISS) tìm 'hàng xóm gần nhất' rất nhanh



3. PIPELINE CƠ BẢN

Retrieval và Reranking — bước then chốt



Retrieval

- Tìm top-k đoạn có khả năng liên quan

Reranking

- Chấm điểm lại để chọn top-n tốt nhất đưa vào LLM
- Reranker đọc trực tiếp (query, doc) → mất ít thông tin hơn

⚠ Truy xuất sai = trả lời sai — đây là bước cực kỳ quan trọng

3. PIPELINE CƠ BẢN

Phản hồi có trích dẫn

- LLM nhận: câu hỏi + các đoạn tài liệu liên quan
- Prompt dạng: "Trả lời câu hỏi dựa trên ngữ cảnh này"
→ bị ràng buộc, ít bịa
- Sinh câu trả lời dựa trên tài liệu đã truy xuất
- Đầu ra kèm trích dẫn nguồn: manual nào, trang nào, báo cáo nào



Vì sao citation quan trọng?

Kỹ sư bảo trì sẽ không tin "thay phớt sau 8000 giờ" nếu không biết con số đó đến từ đâu.

Khi câu trả lời ghi rõ "theo Manual X, mục 4.2", kỹ sư có thể kiểm chứng và yên tâm hành động.

3. PIPELINE CƠ BẢN

RAG khác gì fine-tuning? Khi nào dùng cái nào?



RAG

- Nối mô hình với tri thức ngoài
- Dễ cập nhật — chỉ đổi kho tài liệu
- Truy vết được nguồn · ít ảo giác
- Dùng khi: trả lời thủ công, tiêu chuẩn, báo cáo, dữ liệu thay đổi liên tục



Fine-tuning

- Dạy mô hình cách làm (văn phong, format, etc)
- Khó cập nhật · tốn tài nguyên
- Khó nhớ chính xác kiến thức cụ thể
- Dùng khi: cần văn phong/định dạng/suy luận đặc thù

→ Trong công nghiệp, RAG thường an toàn & rẻ hơn — và có thể kết hợp cả hai.

4. RAG CHO DOANH NGHIỆP

Dữ liệu công nghiệp & RAG giúp gì cho doanh nghiệp



Dữ liệu công nghiệp

- Đa dạng: văn bản, bảng, log, tín hiệu cảm biến, ảnh kiểm tra
- Nhiều nguồn: SOP, SCADA, CMMS, historian
- Rời rạc, khó truy xuất theo cách truyền thống



Hỏi-đáp kỹ thuật từ manual



Tìm nhanh lỗi tương tự trong lịch sử



Tổng hợp báo cáo bảo trì



Hỗ trợ đào tạo kỹ sư mới



Chuẩn hóa tri thức từ chuyên gia

5. RAG and PHM

PHM = Prognostics and Health Management

(chẩn đoán & quản lý sức khỏe thiết bị) · RUL = Remaining Useful Life

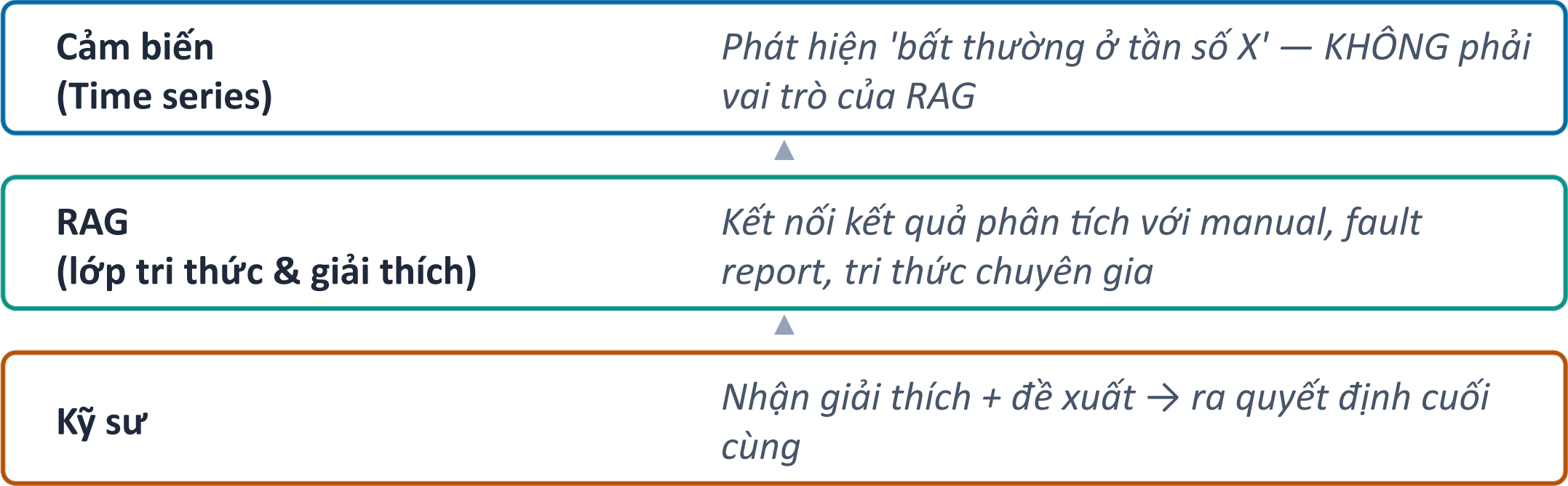


“Chuyển từ bảo trì phản ứng → bảo trì dự đoán”

5. RAG and PHM

PHM còn thiếu gì? — và RAG nằm ở đâu

Khoảng trống: Mô hình phát hiện bất thường tốt nhưng **khó giải thích "vì sao"**. Kỹ sư vẫn phải tự tra cứu thủ công và truy vết từ báo cáo cũ.



6. USE CASES 1-2

Trợ lý chẩn đoán lỗi & Trợ lý bảo trì theo manual



UC1 · Chẩn đoán lỗi

- Input: triệu chứng / cảnh báo / tín hiệu bất thường
- Retrieval: lỗi tương tự + manual + báo cáo cũ
- Output: nguyên nhân khả dĩ, bằng chứng, checklist



UC2 · Bảo trì thủ công

- Kỹ thuật viên hỏi bằng ngôn ngữ tự nhiên
- Hệ thống tìm đúng mục hướng dẫn trong manual
- LLM tóm tắt thành các bước dễ thực hiện



Bắt buộc: giữ nguyên cảnh báo an toàn & điều kiện vận hành (vd: khóa nguồn, xả áp trước khi tháo) khi LLM tóm tắt các bước.

6. USE CASES 3-5

Phân tích báo cáo · Bảo trì dự đoán · Đào tạo



UC3 · Phân tích báo cáo lỗi

*Truy xuất lỗi tương tự, tóm tắt
nguyên nhân gốc rễ, tìm pattern lặp
→ xây knowledge base nhà máy*



UC4 · Hỗ trợ bảo trì dự đoán

*Khi mô hình báo rủi ro tăng, RAG tìm
ca lịch sử tương tự + gợi ý hành động
kiểm tra*



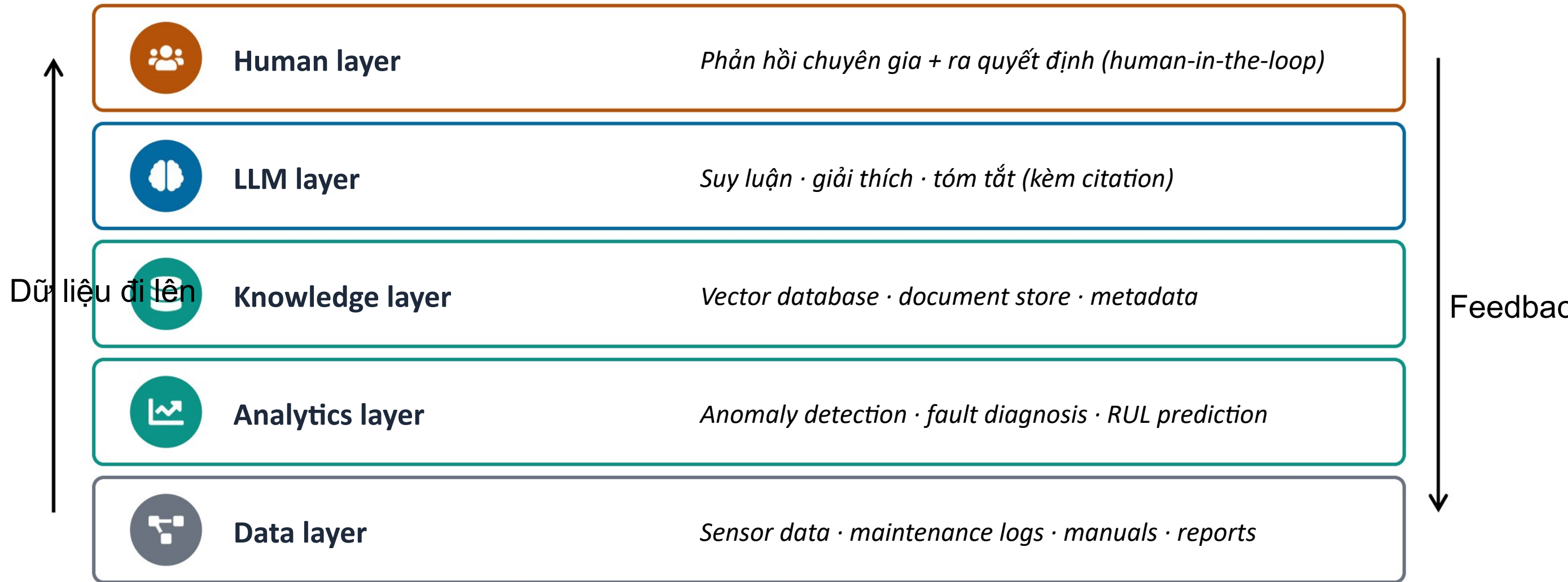
UC5 · Đào tạo & chuyển giao

*Kỹ sư mới hỏi về máy/lỗi/quy trình;
giảm phụ thuộc kinh nghiệm không
ghi chép*

Xuyên suốt mọi use cases: RAG đề xuất — kỹ sư là người quyết định cuối cùng (human-in-the-loop).

7. KIẾN TRÚC TỔNG THỂ

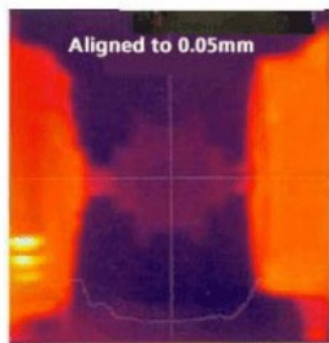
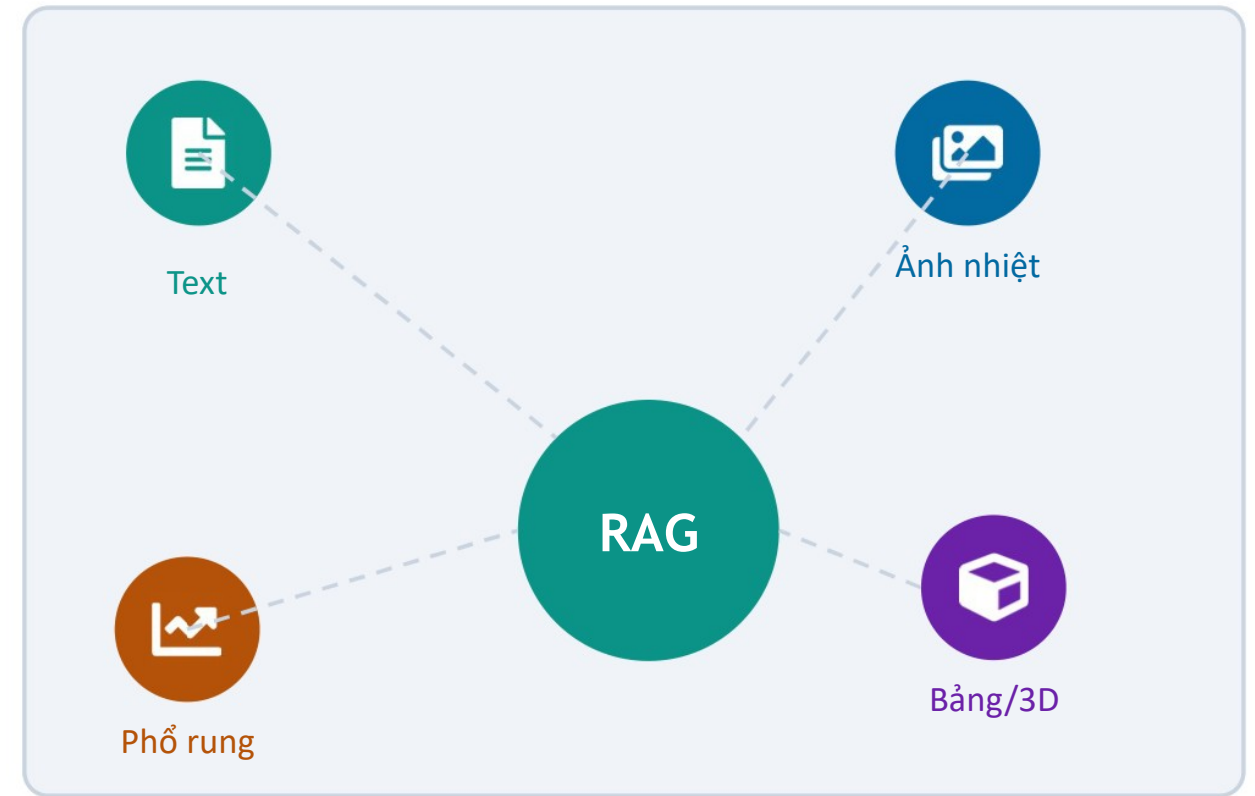
Kiến trúc PHM-RAG 5 lớp



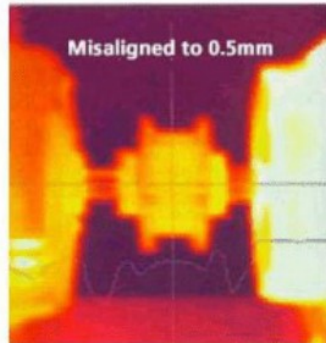
7. KIẾN TRÚC TỔNG THỂ

Multimodal RAG cho PHM

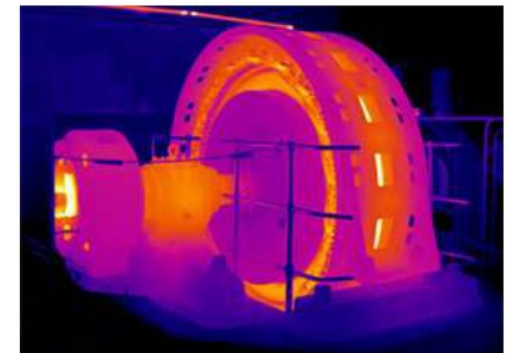
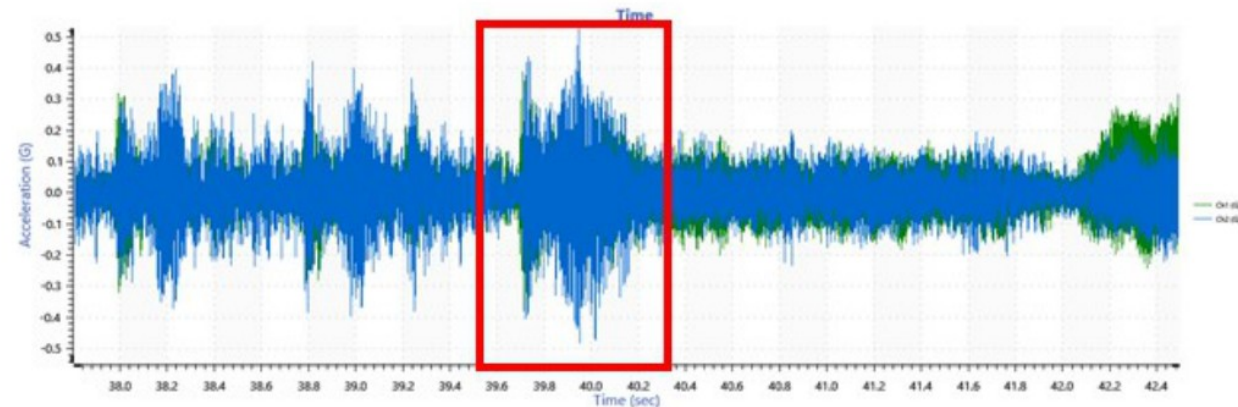
- Không chỉ văn bản: còn bảng, đồ thị rung, ảnh nhiệt, ảnh kiểm tra
- Multimodal RAG mở rộng sang ảnh, video, audio, code
- Kết nối text + image (ảnh kiểm tra) + tín hiệu cảm biến
- Rất hợp với inspection công nghiệp & báo cáo bảo trì



Aligned to +/- 2 mils



Aligned to + 20 mils



8. TRIỂN KHAI

Rủi ro, giới hạn & best practices



Rủi ro & giới hạn

- Truy xuất sai → trả lời sai (rất thuyết phục)
- Tài liệu lỗi thời làm giảm chất lượng
- Ngữ cảnh quá dài → 'lost in the middle'
- Dữ liệu công nghiệp nhạy cảm → lo bảo mật
- **KHÔNG** để LLM tự quyết trong tình huống an toàn cao



Best practices

- Bắt đầu từ use case hẹp, một loại thiết bị
- Làm sạch & chuẩn hóa tài liệu; metadata tốt
- Luôn bật citation
- Ghi log câu hỏi & phản hồi để cải tiến
- Chuyên gia kiểm tra định kỳ · human-in-the-loop

8. TRIỂN KHAI

Hướng nghiên cứu tiềm năng (PHM-RAG)



RAG cho fault diagnosis



RAG + digital twin



RAG + time-series
foundation models



Multimodal RAG cho
inspection



Expert-in-the-loop RAG



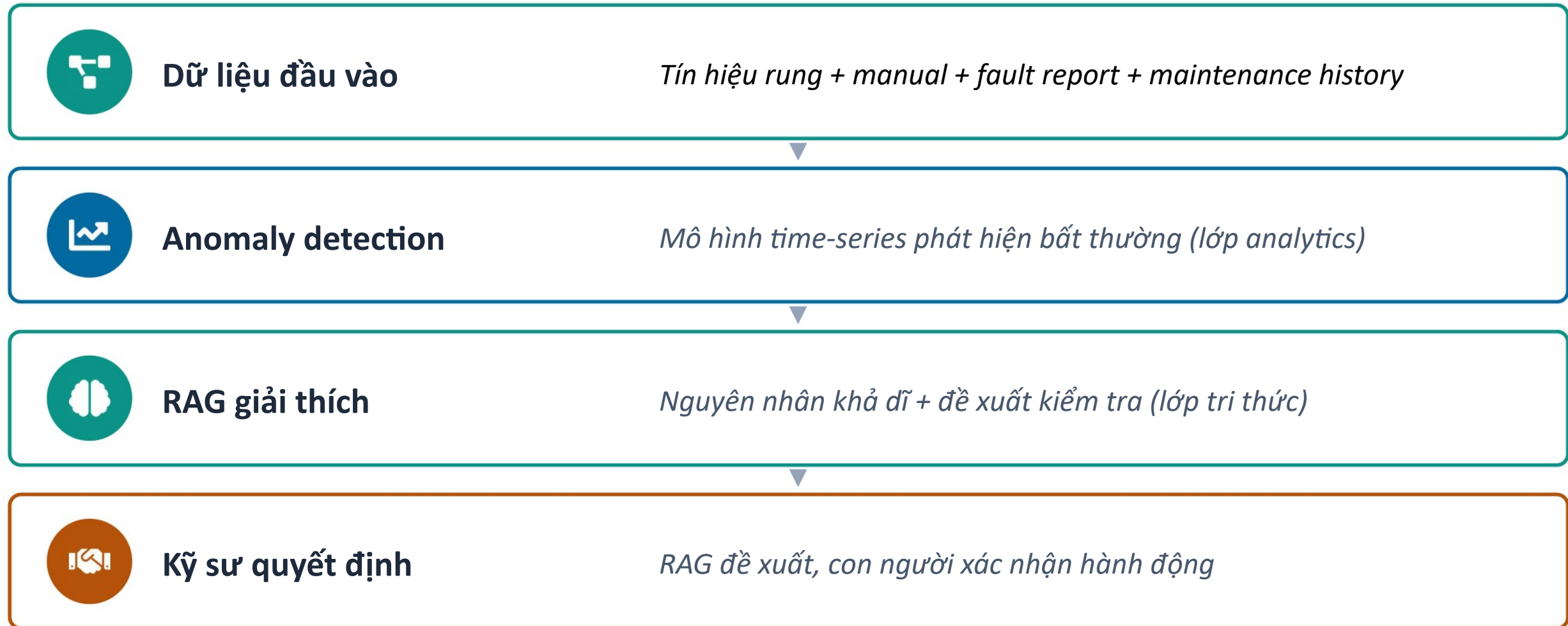
Uncertainty-aware RAG

Ba xu hướng RAG nói chung

Modularity thành dòng chính · **Scaling law** hoàn thiện · **Production-ready** & multimodal

8. CASE STUDY

Chẩn đoán lỗi vòng bi / máy quay



8. Key takeaways



RAG làm LLM đáng tin hơn

Có căn cứ, cập nhật, trích dẫn được — đúng thứ công nghiệp cần



PHM cần kết hợp 2 thứ

Dữ liệu cảm biến + tri thức kỹ thuật — RAG là cây cầu



Hỗ trợ, không thay thế

RAG giúp kỹ sư truy xuất, giải thích, ra quyết định



Công thức thành công

Có căn cứ · có trích dẫn · có human-in-the-loop

HUST

*"Từ dữ liệu cảm biến đến tri thức bảo trì có thể hành động"
(RAG bridges industrial data, engineering knowledge, and human decision-making)*



Dữ liệu → Tri thức → Quyết định

Xin chân thành cảm ơn!