

THỰC TẬP TỐT NGHIỆP

# XÂY DỰNG ỨNG DỤNG FIN CHATBOT



Hiệu quả, năng suất và kết quả khả quan hơn

# Giới thiệu chung

## Công ty cổ phần Goline:

Được thành lập vào tháng 02 năm 2010, là doanh nghiệp công nghệ chuyên cung cấp giải pháp phần mềm Fintech cho thị trường tài chính và chứng khoán tại Việt Nam và khu vực. Sau hơn 15 năm hoạt động và phát triển, Goline đã trở thành một trong những đơn vị uy tín hàng đầu trong lĩnh vực này.

# Goline

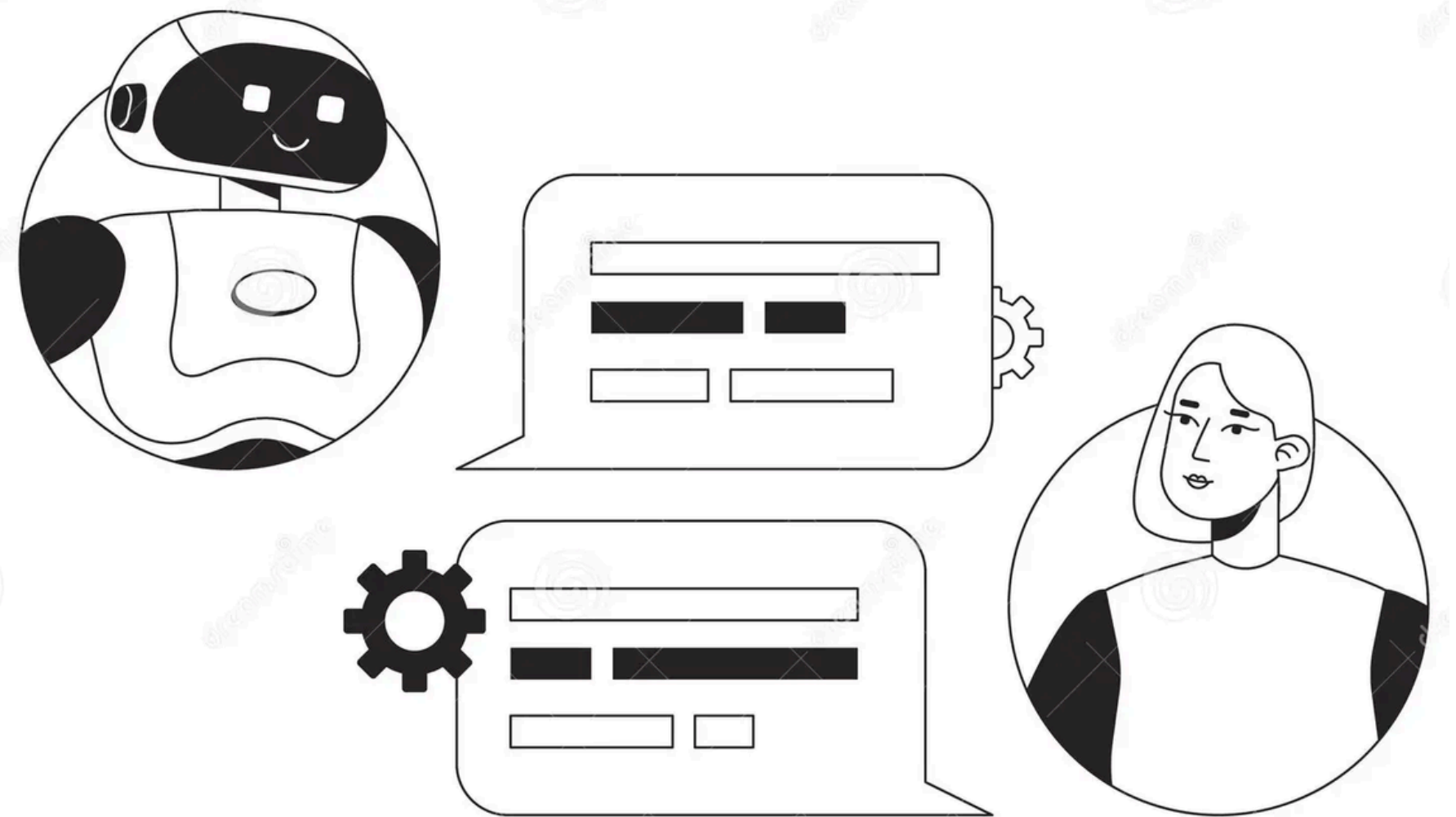
*smart but simple*

# Giới thiệu chung

## Dự án thực hiện:

Ứng dụng Fin Chatbot – một dự án trí tuệ nhân tạo (AI) với mục tiêu xây dựng một hệ thống chatbot tài chính thông minh:

- Hỗ trợ người dùng tra cứu thông tin
- Phân tích cảm xúc thị trường và
- Trả lời các câu hỏi liên quan đến chứng khoán – tài chính.



# Đặt vấn đề & Mục tiêu dự án

## Bối cảnh:

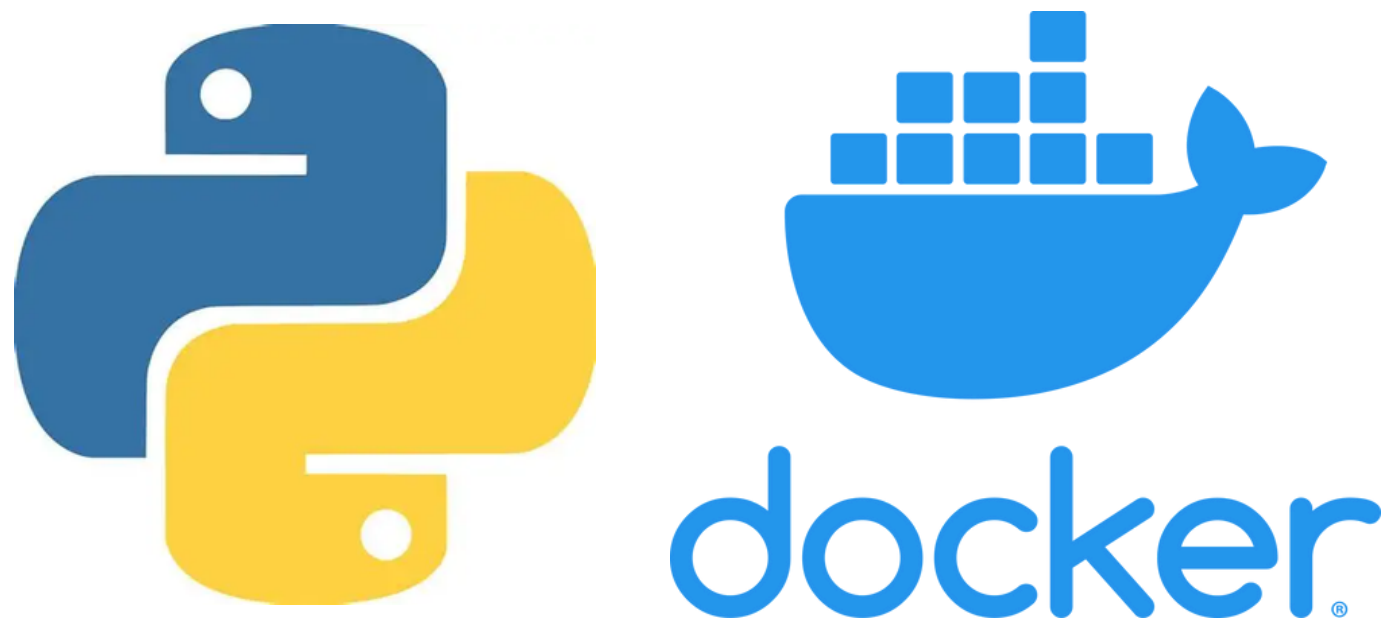
- Sự bùng nổ của AI và LLM trong lĩnh vực tài chính.
- Nhu cầu truy vấn, phân tích thông tin tài chính nhanh chóng và chính xác của nhà đầu tư.

## Mục tiêu chính:

- Xây dựng hệ thống Fin Chatbot thông minh.
- Hỗ trợ người dùng tra cứu, phân tích cảm xúc và tóm tắt tin tức tài chính.
- Áp dụng kiến thức đã học vào dự án thực tế tại doanh nghiệp (Công ty Goline).

# Công nghệ và Công cụ sử dụng

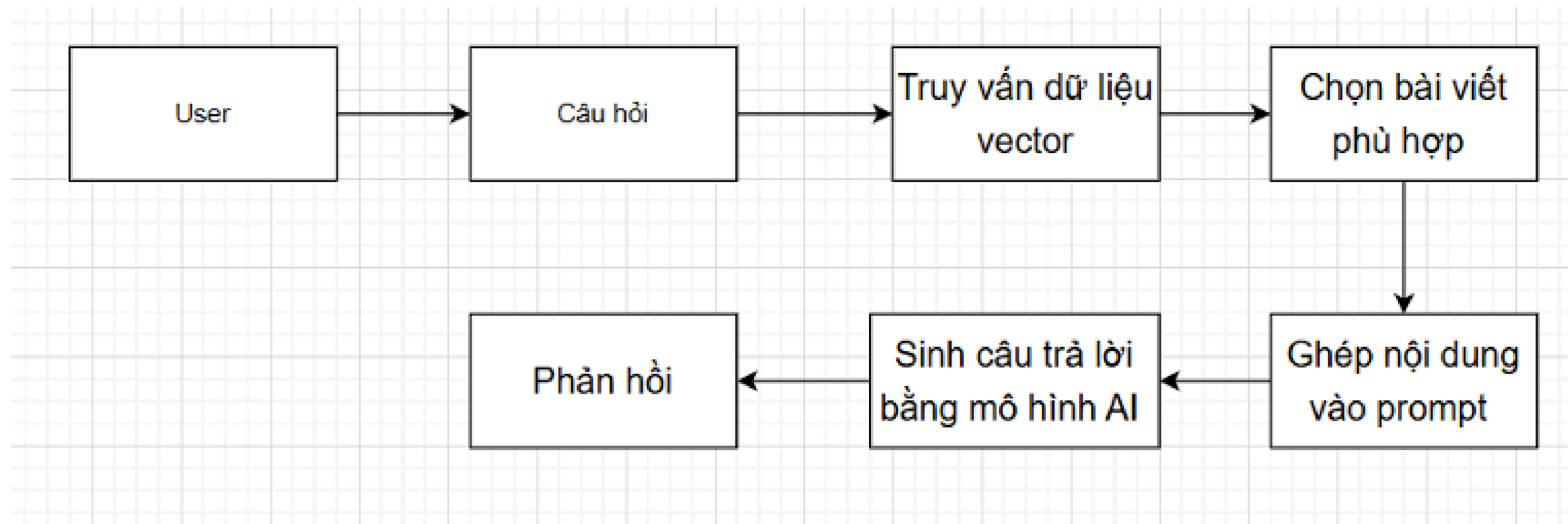
- Ngôn ngữ: Python, SQL
- Thu thập/Xử lý dữ liệu: BeautifulSoup, Pandas
- AI/NLP: HuggingFace Transformers (PhoBERT, T5), LangChain
- Cơ sở dữ liệu: PostgreSQL, Qdrant (Vector DB)
- Backend: FastAPI
- DevOps: Docker, Git



# Sơ đồ kiến trúc tổng quan hệ thống

Thu thập dữ liệu (Crawling) → Xử lý & Vector hóa → Phân tích (Tìm kiếm, Sentiment) → Sinh câu trả lời (LLM).

Tổng quan quy trình:



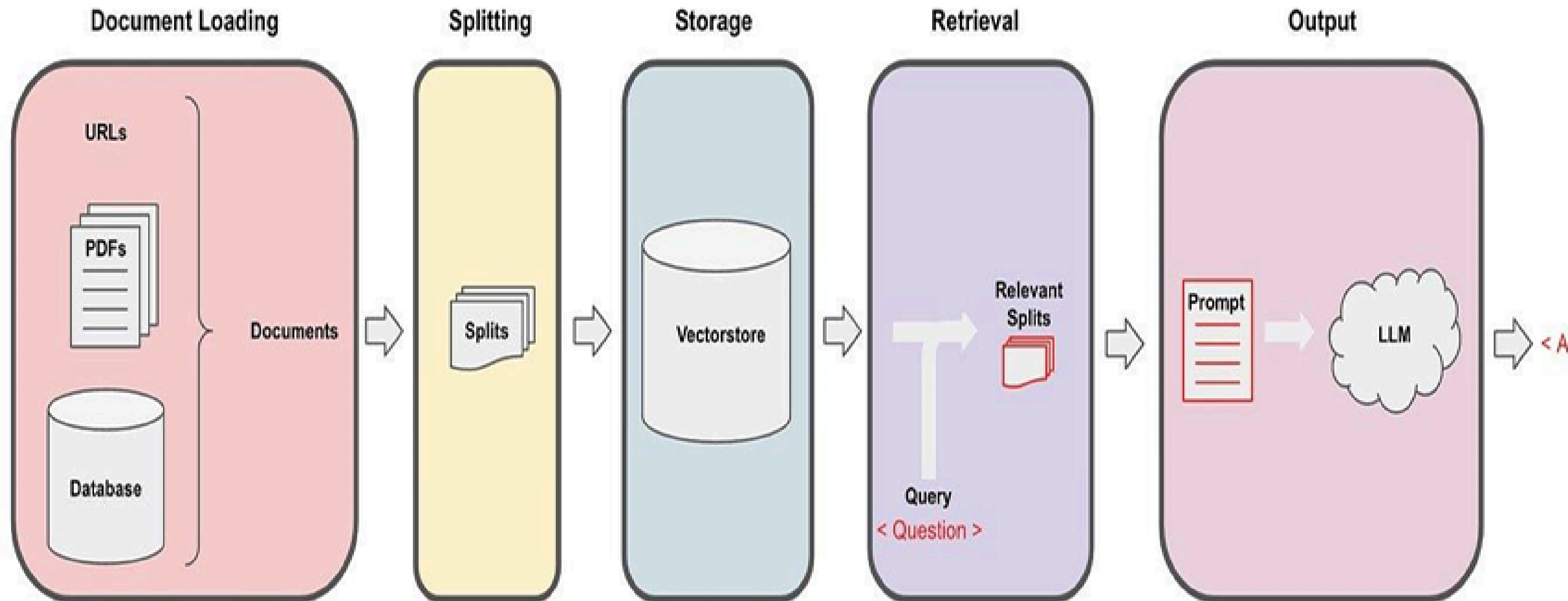
# Thu thập và xử lý dữ liệu

Thu thập: Cào dữ liệu từ CafeF, Vietstock.

Xử lý ngôn ngữ tự nhiên (NLP):

- Làm sạch dữ liệu (HTML, ký tự lạ).
- Xử lý tiếng Việt chuyên ngành (vd: "lợi nhuận gộp", mã ticker "HPG").
- Trích xuất thực thể (NER): Dùng PhoBERT-large fine-tuned để nhận diện ORG, TICKER, KPI, TIME.

# Thu thập và xử lý dữ liệu



*Hình 2. Quy trình xử lý văn bản tài chính tiếng Việt*



# Phân tích cảm xúc

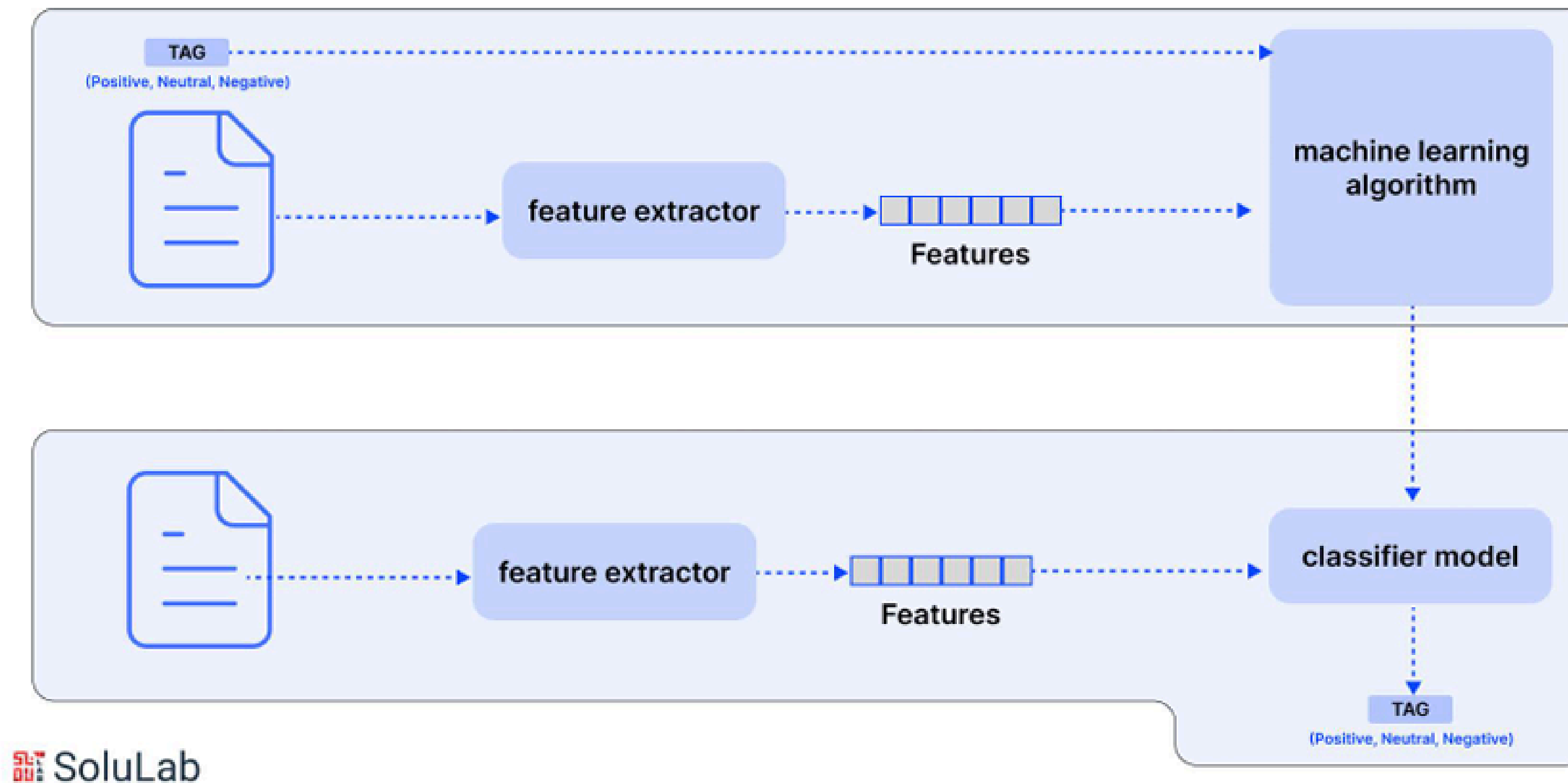
**Mục tiêu:** Đánh giá thái độ (Tích cực, Tiêu cực, Trung lập) của tin tức.

**Phương pháp:**

- Dữ liệu: 50k bài viết, gán nhãn semi-supervised.
- Mô hình: Fine-tune FinBERT-vi (biến thể từ PhoBERT).
- Kết quả: Đạt Macro F1 = 0.816.

**Ứng dụng:** Tính Sentiment Score cho từng mã cổ phiếu, vẽ biểu đồ xu hướng.

# Phân tích cảm xúc



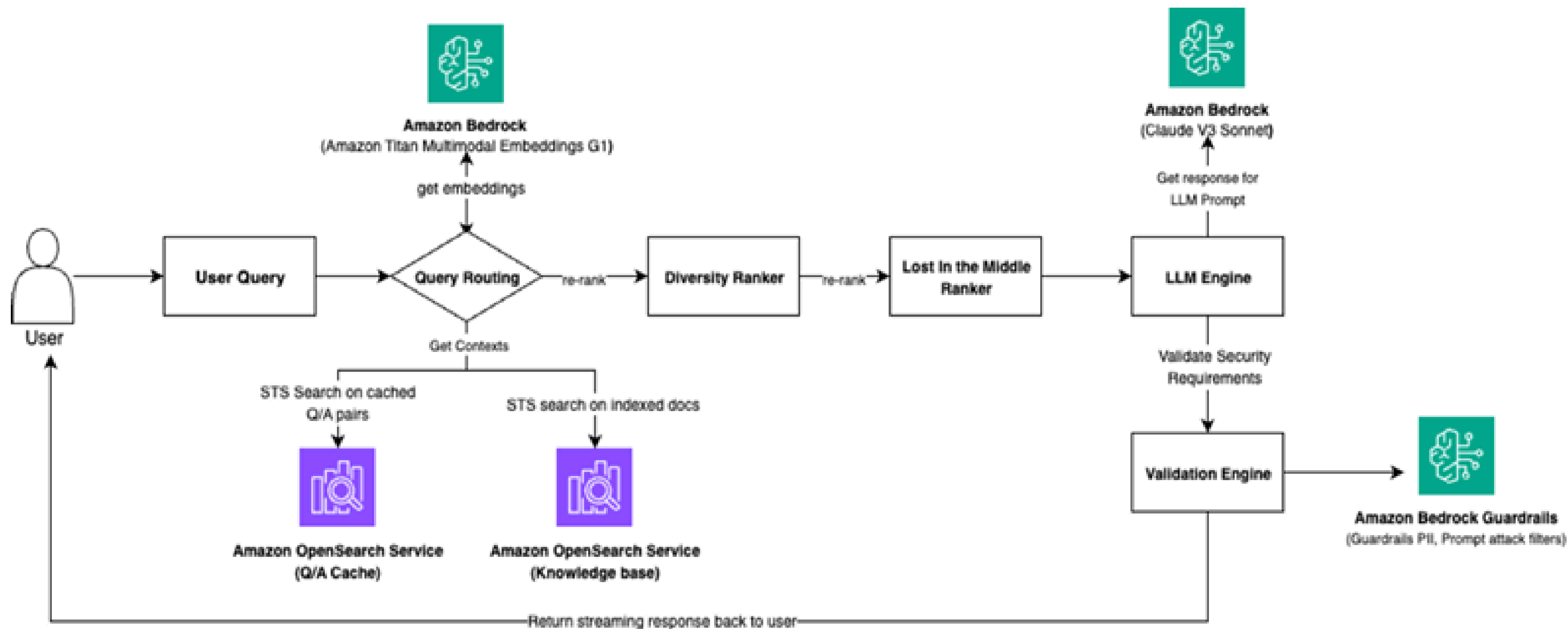
Hình 3. Quy trình phân tích cảm xúc bằng machine learning

# Tìm kiếm ngữ nghĩa và sinh câu trả lời

## Kiến trúc RAG:

- Retriever (Bộ truy xuất):
  - Vector hóa văn bản bằng PhoBERT/T5-encoder.
  - Lưu trữ và tìm kiếm trên Qdrant (Vector Database).
  - Sử dụng re-ranker để tăng độ chính xác.
- Generator (Bộ sinh):
  - Sử dụng mô hình T5-base-vi đã được fine-tune.
  - Nhận đầu vào: [Bối cảnh từ văn bản được tìm thấy] + [Câu hỏi của người dùng].

# Tìm kiếm ngữ nghĩa và sinh câu trả lời



Hình 4. Kiến trúc tổng thể của hệ thống truy vấn ngữ nghĩa kết hợp LLM

# Khó khăn trong quá trình thực hiện

**Dữ liệu:** Tiếng Việt tài chính không chuẩn hóa, thiếu dữ liệu gán nhãn.

**Mô hình:**

- Mô hình embedding còn hạn chế với ngữ cảnh chuyên ngành.
- Độ trễ của LLM.

**Triển khai:** Tối ưu pipeline RAG, kiểm soát thông tin "ảo" (hallucination).

# Hướng phát triển và Mở rộng

**Nâng cao mô hình:** Fine-tune các mô hình lớn hơn (Mistral, Qwen) trên dữ liệu chuyên ngành.

**Thêm tính năng:**

- Dự đoán xu hướng cảm xúc.
- Tích hợp giọng nói.
- Cảnh báo thông tin tiêu cực.

**Tối ưu hệ thống:** Tích hợp Guardrails để kiểm soát đầu ra, tối ưu tốc độ bằng ONNX.