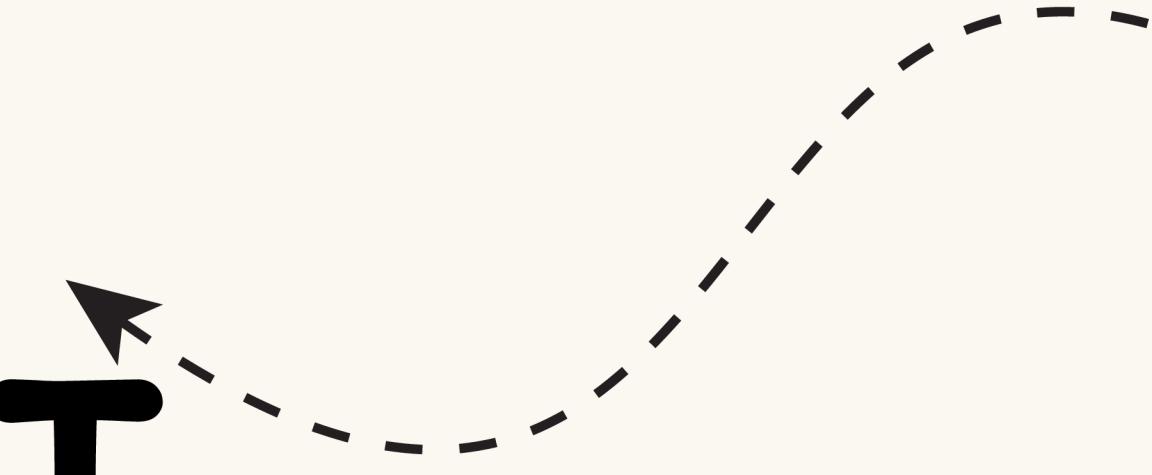
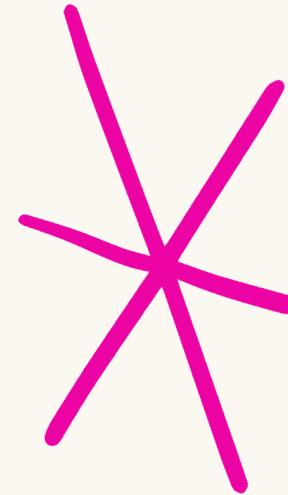


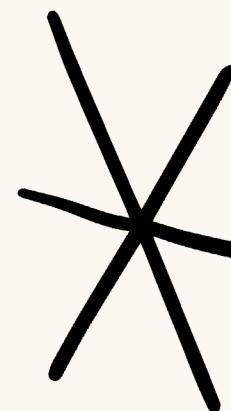


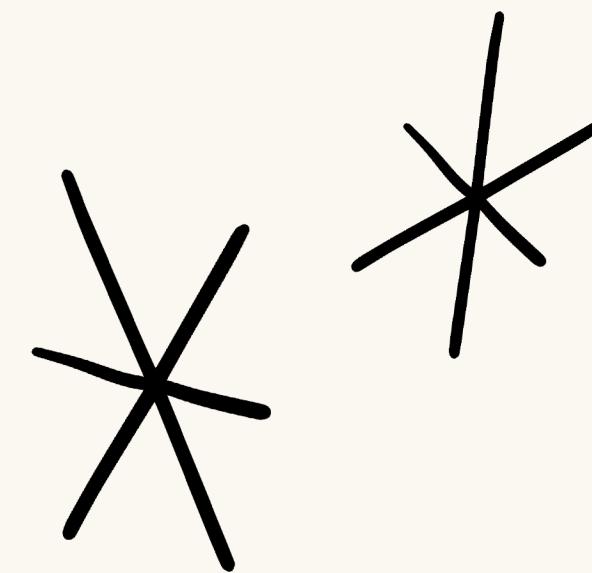
Vietnamese Text Readability



Mục lục

1. Giới thiệu thành viên
2. Bối cảnh và mục tiêu
3. Giới thiệu mô hình và bộ dữ liệu
4. Quy trình tổng quát
5. Kết quả huấn luyện và đánh giá
6. Ứng dụng thực tế





Giới thiệu thành viên

Trần Mạnh Hùng
Trưởng nhóm

Mai Nhựt Huy
Thành viên

Võ Nguyễn Song Huy
Thành viên

Vy Quốc Huy
Thành viên

Nguyễn Hùng Việt
Thành viên

Bối cảnh và mục tiêu



Bối cảnh

- Nhu cầu phổ cập tri thức → cần kiểm soát độ dễ hiểu văn bản.
- Ứng dụng: giáo dục, báo chí, chatbot, trợ lý ảo.
- Tiếng Việt ít công cụ đánh giá độ dễ hiểu, còn hạn chế so với tiếng Anh.

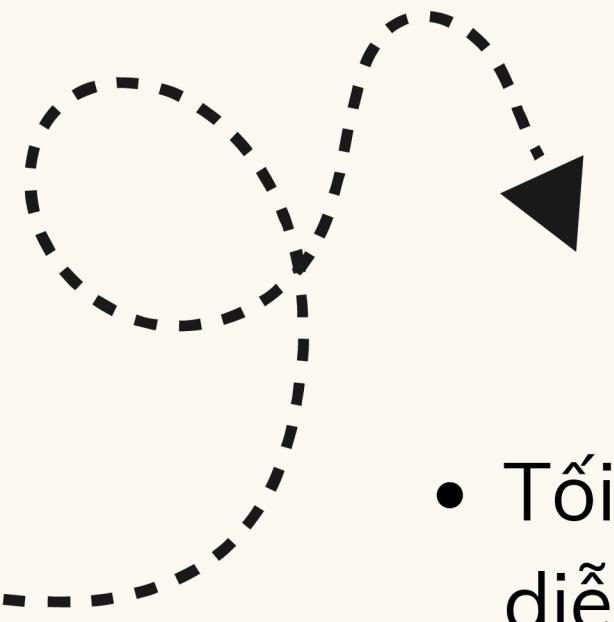
Mục tiêu

- Xây dựng mô hình đánh giá độ dễ hiểu văn bản tiếng Việt:
 - Thu thập & tiền xử lý dữ liệu: văn bản có nhãn độ dễ hiểu.
 - Tinh chỉnh mô hình Transformer (PhoBERT) → học ngữ nghĩa và đánh giá độ dễ hiểu.
 - Đánh giá mô hình: Accuracy, F1-score, Confusion Matrix.



Giới thiệu mô hình

- PhoBERT: Biến thể BERT cho tiếng Việt (Transformer Encoder).
- Tiền huấn luyện: 20GB Wikipedia, tin tức, web.
- Dùng PhoBERT-base phù hợp tài nguyên.
- Hai hướng khai thác:
 - Fine-tuning: Tinh chỉnh toàn bộ mô hình + lớp phân loại.
 - Feature Extraction: Trích embedding → phân loại ngoài.



LÝ DO CHỌN

- Tối ưu cho tiếng Việt → Biểu diễn ngữ nghĩa, cú pháp tốt.
- Hiệu quả cao nhiều tác vụ NLP tiếng Việt.
- Linh hoạt: Hỗ trợ Fine-tuning & Feature Extraction.
- Dễ mở rộng sang tóm tắt, tái viết...

Bộ dữ liệu

- Nguồn: Vietnamese Text Readability Dataset (GitHub, Lương An Vinh).
- Dạng: Văn bản ngắn, đa lĩnh vực (giáo dục, tin tức, xã hội...).
- Label: 4 mức — Dễ, Trung Bình, Khó, Rất Khó
- Tổng cộng: 1825 mẫu.
 - DỄ: 809
 - TRUNG BÌNH: 453
 - KHÓ: 242
 - RẤT KHÓ: 321
- Độ dài: 53 – 34,709 từ.

TIỀN XỬ LÝ

- Chữ thường, chuẩn Unicode.
- Loại ký tự không hợp lệ.
- Chuẩn hóa khoảng trắng.
- Tách từ: VnCoreNLP.
- Tokenizer PhoBERT → Trích đặc trưng.

QUY TRÌNH TỔNG QUÁT

- Tiền xử lý & chuẩn hóa (VnCoreNLP, Unicode, làm sạch).
- Chia tập dữ liệu: Train/Test = 80/20.
- Mã hóa: Tokenizer PhoBERT → tensor đầu vào.
- Xây dựng mô hình: PhoBERT + các lớp phân loại.
- Huấn luyện: Fine-tuning toàn bộ.
- Đánh giá: Accuracy, Confusion Matrix, F1-score.

Đánh giá bằng Confusion Matrix,
Accuracy, F1-score

KẾT QUẢ HUẤN LUYỆN

Accuracy

- Train: 68.8% → 89.8% (5 epoch)
- Validation: ~73%–81%, ổn định

Loss

- Train Loss giảm mạnh
- Validation Loss ~11–13 → không overfit

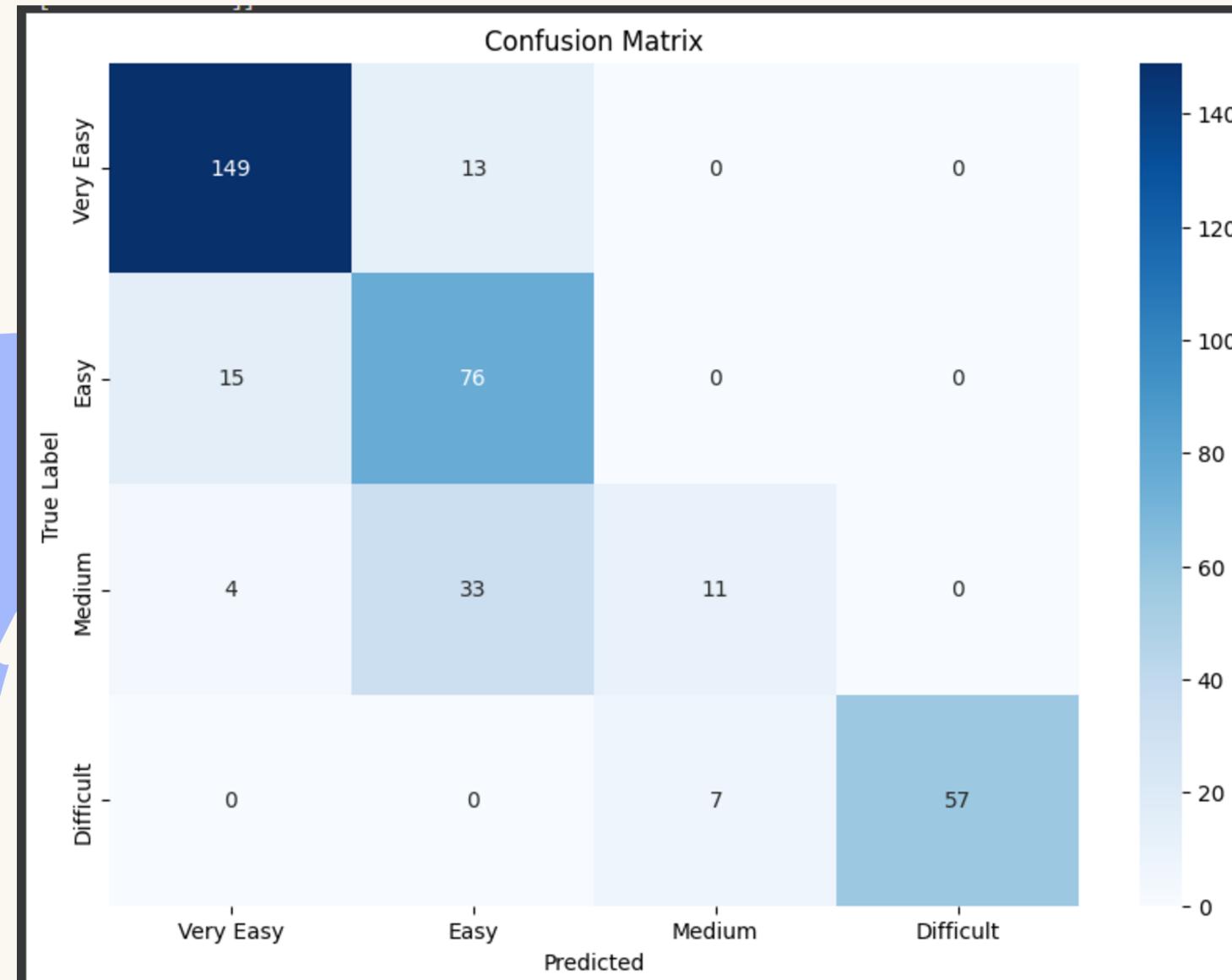
F1-score

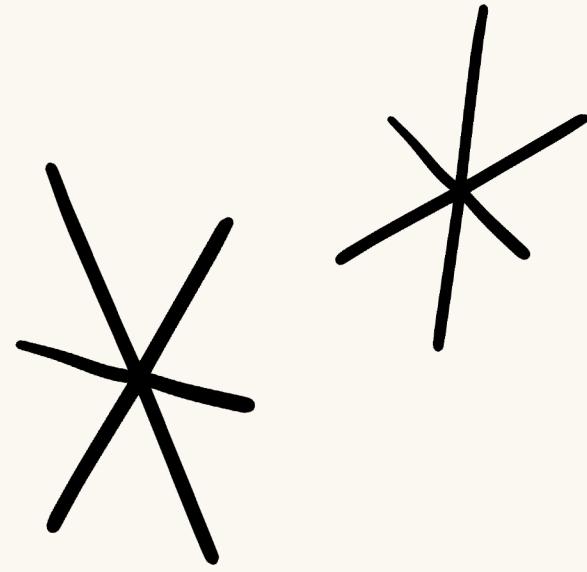
- Train: 0.55 → 0.84
- Validation: 0.63 → ~0.72–0.75

Kết luận: PhoBERT fine-tuning hoạt động tốt, Accuracy & F1-score tăng đều

Hạn chế: Dữ liệu chưa cân bằng (Medium ít), Câu trung tính khó gán nhãn

Hướng nâng cao: Thêm dữ liệu mới, fine-tuning hyperparameter chi tiết hơn hoặc cân bằng số lượng giữa các lớp





Ứng dụng thực tế

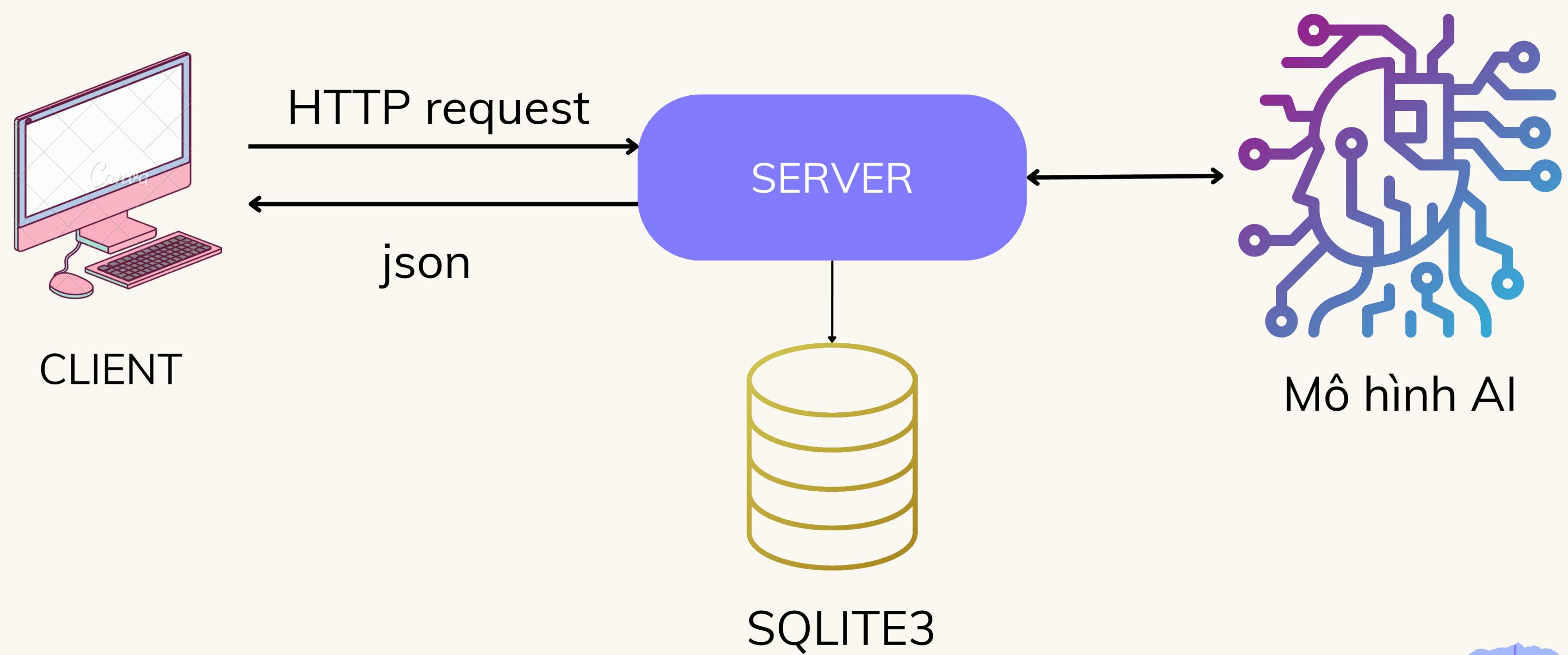
- Gợi ý tài liệu phù hợp với độ tuổi học sinh.
- Điều chỉnh bài kiểm tra, bài đọc hiểu cho đúng cấp lớp.
- Báo chí muốn tiếp cận đại chúng: dùng để kiểm tra xem bài có "dễ nuốt" không.
- Dành cho người nước ngoài học tiếng Việt: giúp chọn bài đọc theo trình độ.
- Chấm điểm độ khó khi ai đó đọc hay viết.





Kiến trúc

Mô hình Client - Server



SERVER

Trung tâm xử lý logic, giao tiếp
giữa client – database – AI model

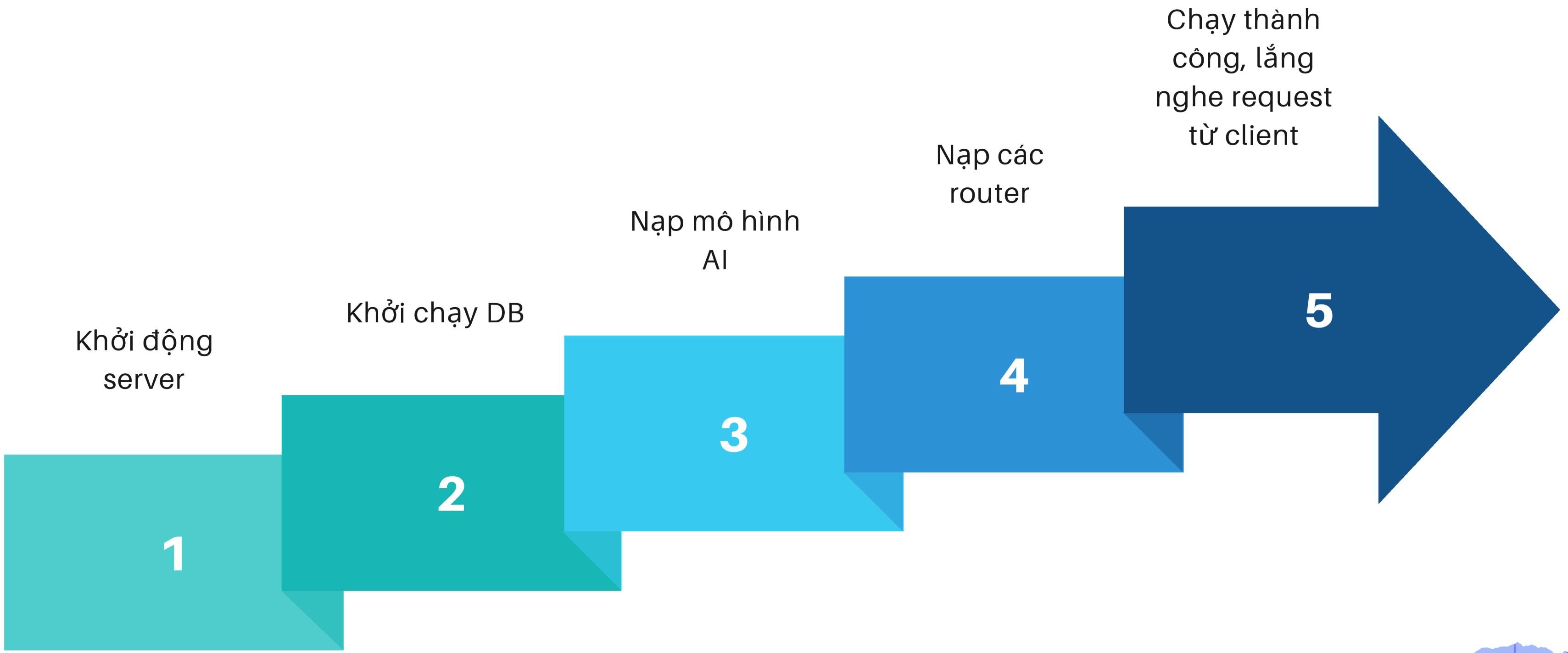
- Xử lý đăng ký và đăng nhập người dùng
- Nhận yêu cầu phân tích câu từ client
- Trả kết quả phân tích về dạng JSON
- Lưu và truy xuất lịch sử phân tích
- Đảm bảo luồng hoạt động liên tục

Ngôn ngữ: python

Framework: FastAPI

Cơ sở dữ liệu: SQLite3

KHỞI CHẠY SERVER



THANK YOU!

