Customer Support Chatbot

Nhóm 4

GIẢNG VIÊN

- Lê Ngọc Thành
- Lê Nhựt Nam
- Trần Quốc Huy

THÔNG TIN NHÓM

- 20127206 - Vũ Đình Duy Khánh
- 20127384 - Văng Khánh Tường
- 20127443 - Nguyễn Hồ Hữu Bằng
- 20127652 - Hoàng Minh Triết

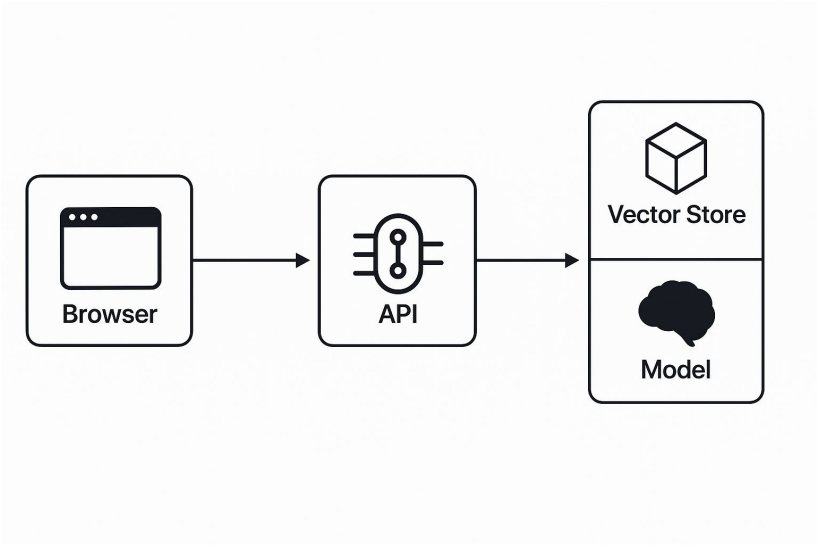
1. Tổng quan
2. Kiến trúc tổng thể
3. Luồng hoạt động
4. Các tính năng nổi bật
5. Công nghệ sử dụng
6. Demo
7. Khả năng mở rộng trong tương lai
8. Live Demo và Q&A

Tổng Quan

- **Mục tiêu dự án**
  - Hỗ trợ đa ngôn ngữ
  - Xây dựng chatbot hỗ trợ khách hàng tự động, giảm tải cho đội ngũ chăm sóc khách hàng, nâng cao trải nghiệm người dùng.
  - Giảm thời gian chờ đợi của khách.
  - Xử lý 24/7, không phụ thuộc vào thời gian làm việc
- **Hỗ trợ đa ngôn ngữ**
  - Tiếng Anh & Tiếng Việt (có thể mở rộng dễ dàng)
  - Chuyển đổi ngôn ngữ real-time tùy theo input
- **Intent Detection**
  - Tự động phân loại ý định (ví dụ: hỏi giá, khiếu nại, đặt lịch…)
  - Cơ chế học liên tục, cải thiện độ chính xác theo thời gian
- **Gợi ý câu hỏi tiếp theo**
  - Dựa trên ngữ cảnh hội thoại
  - Giúp giữ luồng tương tác liên tục, tránh dead-end

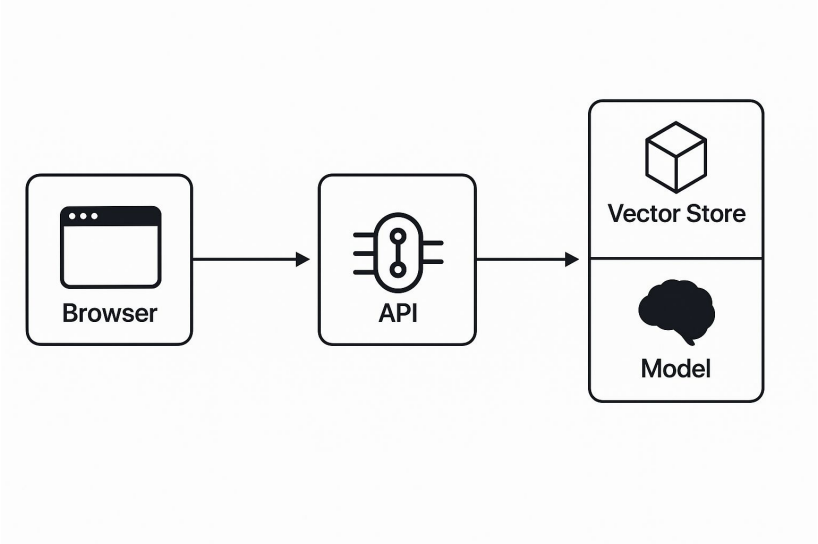
Kiến Trúc Tổng Thể

- **Frontend**
  - Sử dụng React với Hooks để quản lý state linh hoạt.
  - CSS Variables giúp theming và responsive dễ dàng.
  - Fetch API kết nối đến backend qua các endpoint REST.



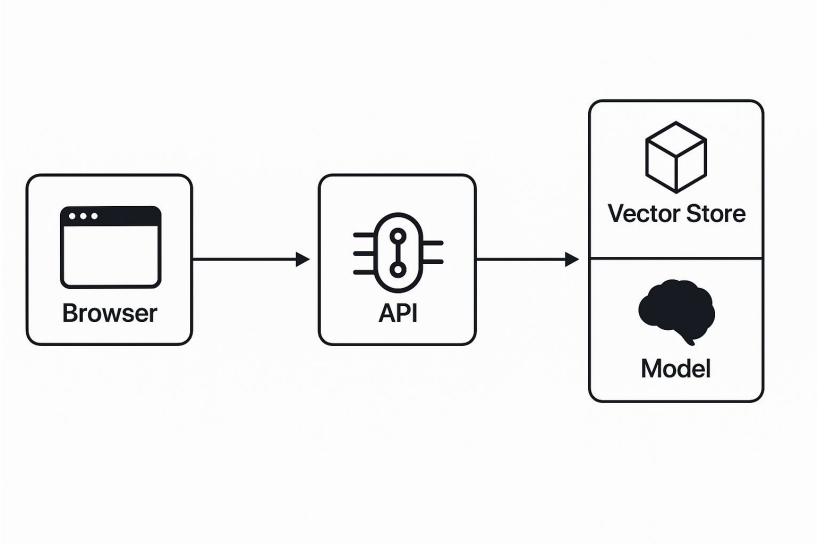
Kiến Trúc Tổng Thể

- **Backend**
  - FastAPI đảm bảo hiệu năng cao, hỗ trợ async/await.
  - Tất cả business logic (intent detection, trả lời) nằm trong ChatbotService.
  - Vector store FAISS lưu trữ embeddings, cho phép tìm kiếm gần nghĩa nhanh chóng.



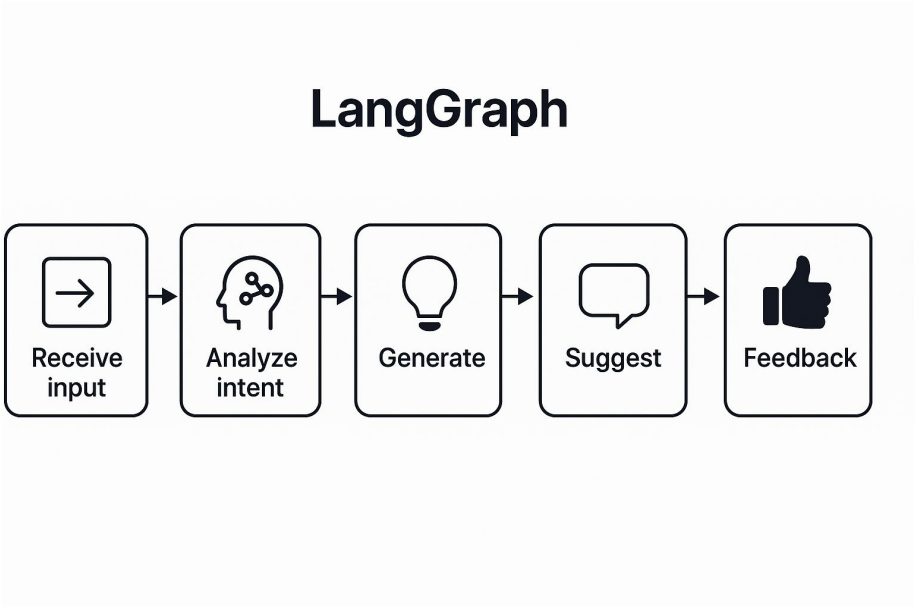
Kiến Trúc Tổng Thể

- **AI/ML Layer**
  - Embeddings lấy từ HuggingFace (sentence-transformers).
  - Mô hình LLM (Transformers) xử lý generation.
  - Intent Detection tách riêng thành module, kết hợp transformer fine-tuning.



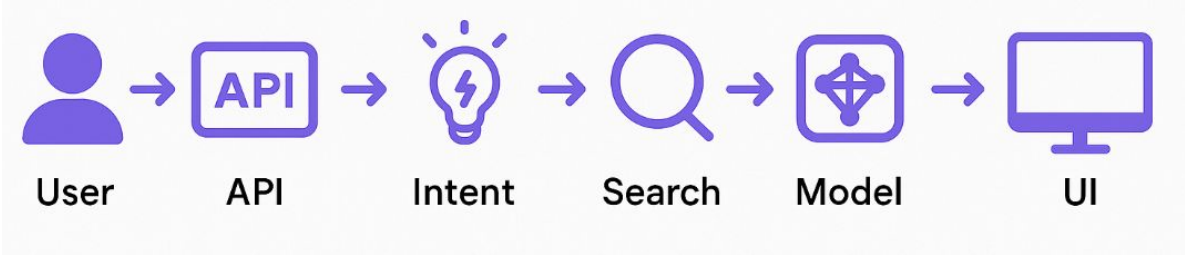
Kiến Trúc Tổng Thể

- **Workflow Orchestration**
  - LangGraph định nghĩa các bước: nhận input → phân tích intent → search → generate → gợi ý → feedback.
  - Cho phép tái sử dụng và mở rộng luồng dễ dàng.



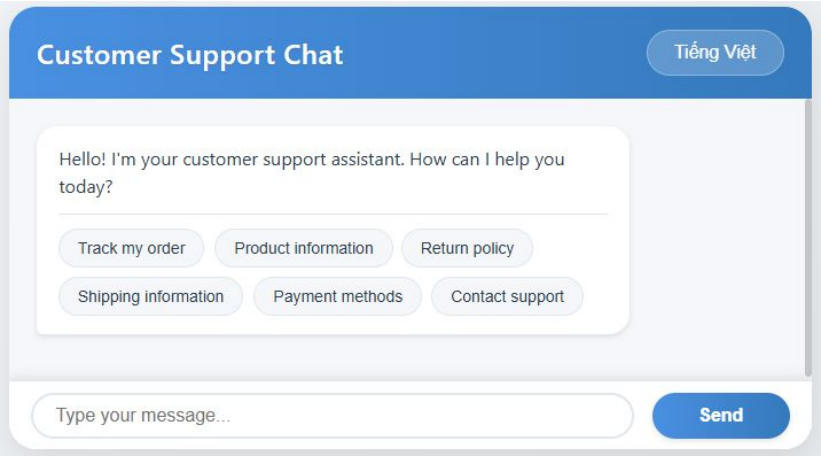
Luồng Hoạt Động

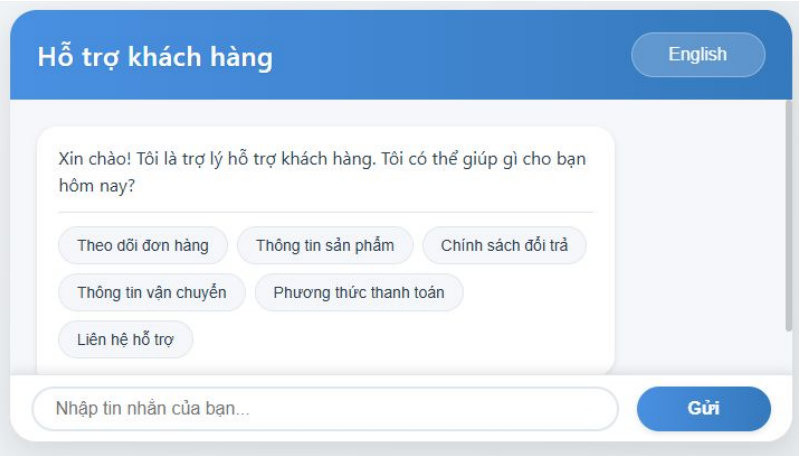
- **Tổng quan**
  - Luồng hoạt động gồm hai chiều tương tác: từ người dùng (frontend) đến hệ thống AI (backend), rồi phản hồi lại người dùng.
  - Đây là kiến trúc event-driven và context-aware.
- **Chi tiết Frontend**
  - Khi người dùng gửi câu hỏi, giao diện sẽ hiển thị trạng thái "đang xử lý" để cải thiện UX.
  - Sau đó gửi dữ liệu qua Fetch API đến endpoint FastAPI backend.
- **Chi tiết Backend**
  - Tin nhắn được xử lý qua một pipeline:
    - Trích xuất ý định (intent)
    - Semantic search trong FAISS để tìm câu trả lời tốt nhất kết hợp sử dụng language model để sinh phản hồi.
- **Suggestions & Feedback**
  - Dựa vào câu trả lời, backend gợi ý các câu tiếp theo có liên quan để giữ người dùng tiếp tục hội thoại.



Các Tính Năng Nổi Bật

- **Đa ngôn ngữ**
  - Khả năng xử lý đa ngôn ngữ là điểm mạnh đặc biệt, hỗ trợ mở rộng thị trường.
  - Sử dụng trình phiên dịch giúp dễ dàng thêm các ngôn ngữ mới cho các quốc gia khác nhau.
- **Trải nghiệm người dùng**
  - Giao diện sử dụng thiết kế tối giản, thân thiện với người dùng trên cả mobile và desktop.
  - Các hiệu ứng loading, animation giúp tạo cảm giác “đang được lắng nghe”, tăng sự tin cậy.





Các Tính Năng Nổi Bật

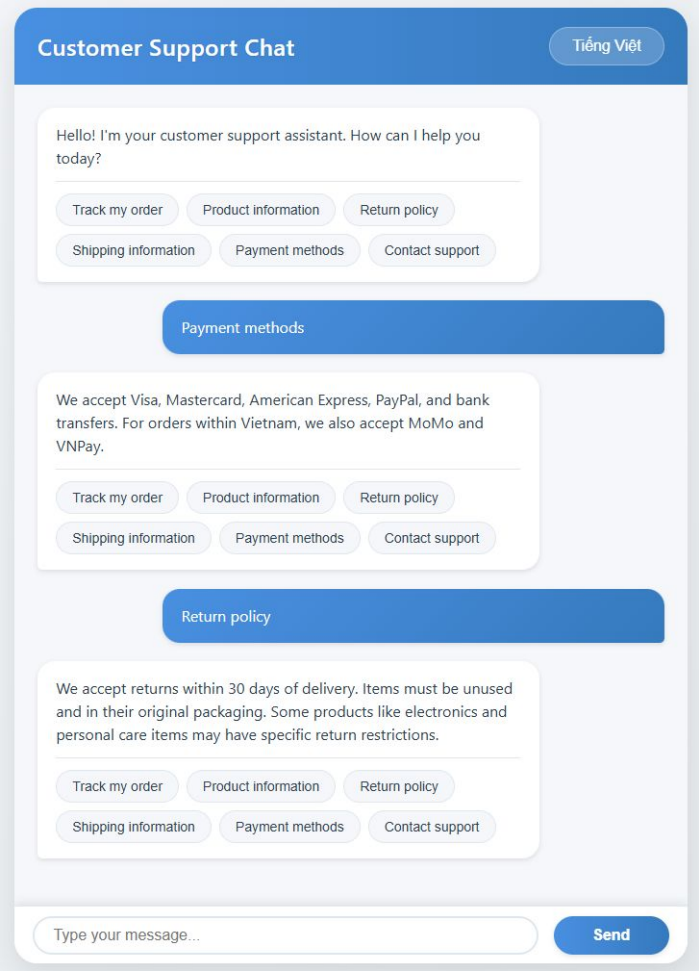
- **AI/ML Integration**
  - Semantic Search cho phép chatbot hiểu và tìm đúng thông tin dù câu hỏi không trùng khớp 100%.
  - Intent Detection giúp phân loại yêu cầu và kích hoạt logic phù hợp (ví dụ: báo mất hàng, hỏi giá, đổi trả…).

```
Question: How long will it take to receive my order?
Top 3 predictions:
- delivery_period: 90.92%
- delivery_options: 1.41%
- place_order: 1.36%

Question: Do you deliver to Hanoi?
Top 3 predictions:
- delivery_options: 63.08%
- delivery_period: 16.18%
- check_payment_methods: 2.58%
```

```
Question: How do I create a new account?
Top 3 predictions:
- create_account: 93.29%
- delete_account: 1.02%
- newsletter_subscription: 0.60%

Question: I want to update my personal information
Top 3 predictions:
- edit_account: 92.67%
- recover_password: 0.74%
- switch_account: 0.54%
```

Công nghệ sử dụng - Front End

React (Hooks)	: Sử dụng useState, useEffect, useRef để quản lý trạng thái và side effects.
CSS Variables	: Tối ưu hoá khả năng custom giao diện (dark/light mode, brand color…).
Modern JS (ES6+)	: Sử dụng destructuring, optional chaining, async/await.
Fetch API	: Gửi/nhận dữ liệu từ backend bằng HTTP request.
Error Boundaries	: Bắt và xử lý lỗi UI mà không làm sập toàn bộ ứng dụng.
Responsive Design	: Hỗ trợ giao diện mượt mà trên mobile, tablet, desktop.



Công nghệ sử dụng - Back End

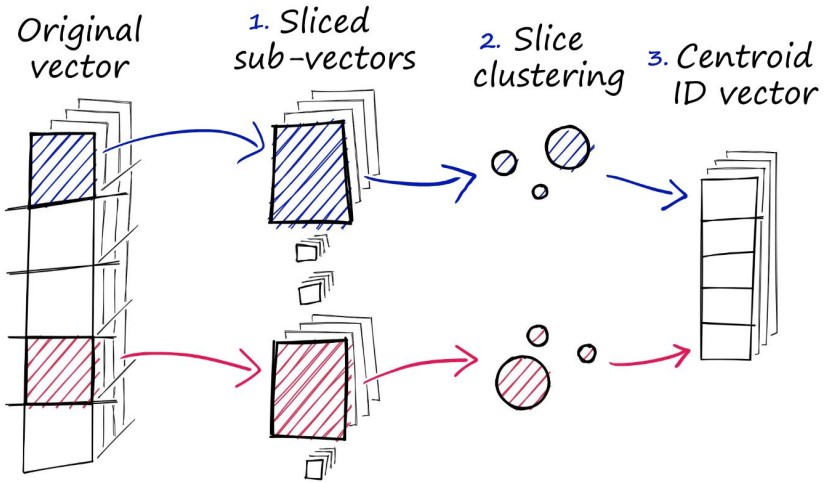
- FastAPI
  - Framework Python hiệu năng cao, hỗ trợ async I/O
  - Dễ viết & dễ maintain.

Công nghệ sử dụng - Back End

- LangChain
  - Vai trò : Tổ chức pipeline xử lý AI (retrieval → prompt → response
  - Ưu điểm :
    - Dễ modular hóa : mỗi bước như intent detection, memory, query… đều là một component.
    - Kết hợp tốt với nhiều LLMs (OpenAI, HuggingFace, …)
- LangGraph :
  - Vai trò : Định nghĩa flow hội thoại như đồ thị, thay vì pipeline tuyến tính
  - Ưu điểm :
    - Linh hoạt cho luồng nhân nhánh (nếu intent thay đổi -> xử lý khác).
    - Hỗ trợ flow phức tạp.
- Kết hợp LangChain với LangGraph
  - Dùng LangChain agents như một node trong LangGraph để xử lý linh hoạt hơn.
  - Tạo multi-step workflows như:
    - Node A: Trích xuất thông tin (IE).
    - Node B: Tóm tắt.
    - Node C: Hỏi đáp từ tóm tắt.
  - Dùng LangGraph để dễ dàng **quản lý nhánh logic** dựa trên trạng thái hoặc kết quả.

Công nghệ sử dụng - Back End

- FAISS :
  - Vai trò : Lưu và tìm kiếm vector embeddings cho truy vấn ngữ nghĩa
  - Ưu điểm :
    - Tìm kiếm semantic rất nhanh, hỗ trợ hàng triệu vectors.
    - Tích hợp dễ dàng với LangChain



Công nghệ sử dụng - Back End

- HuggingFace : Sử dụng embeddings và model NLP như Tranformer T5
  - Vai trò : cup cấp model NLP (embedding và generation).
  - Ưu điểm :
    - Linh hoạt : Chỉ cần 1 model duy nhất cho nhiều task NLP khác nhau.
    - Tối ưu prompt engineering : Dễ kết hợp context và dữ liệu đầu vào.
    - Thân thiện đa ngôn ngữ : Dễ fine-tune cho đa ngôn ngữ, giúp đúng văn phong, quy định.
- Vì sao tự training model Tranformer T5 không phải dùng API có sẵn ?
  - → Giúp tăng tính học thuật cho đồ án, học tập được nhiều hơn.
- Vì sao sử dụng T5 mà không phải model khác của Tranformer như Bert?

Tiêu chí	T5	Bert
Hướng xử lý	Generative - sinh văn bản	Encoder-only - hiểu, không sinh
Ứng dụng chính	Trả lời câu hỏi, dịch, tóm tắt, chat	Phân loại, tìm kiếm
Khả năng sinh câu trả lời	Có thể **generate full response** dựa trên context	Không thể generate câu trả lời
Fine-tuning theo task cụ thể	Rất linh hoạt, các task đều là “text -> text”	Bị giới hạn, mỗi task cần head riêng

Công nghệ sử dụng - Back End

- HuggingFace : Sử dụng embeddings và model NLP như Tranformer T5
  - Vai trò : cup cấp model NLP (embedding và generation).
  - Ưu điểm :
    - Linh hoạt : Chỉ cần 1 model duy nhất cho nhiều task NLP khác nhau.
    - Tối ưu prompt engineering : Dễ kết hợp context và dữ liệu đầu vào.
    - Thân thiện đa ngôn ngữ : Dễ fine-tune cho đa ngôn ngữ, giúp đúng văn phong, quy định.
- Vì sao tự training model Tranformer T5 không phải dùng API có sẵn ?
  - → Giúp tăng tính học thuật cho đồ án, học tập được nhiều hơn.
- Vì sao sử dụng T5 mà không phải model khác của Tranformer như Bert?

Tiêu chí	T5	Bert
Hướng xử lý	Generative - sinh văn bản	Encoder-only - hiểu, không sinh
Ứng dụng chính	Trả lời câu hỏi, dịch, tóm tắt, chat	Phân loại, tìm kiếm
Khả năng sinh câu trả lời	Có thể **generate full response** dựa trên context	Không thể generate câu trả lời
Fine-tuning theo task cụ thể	Rất linh hoạt, các task đều là “text -> text”	Bị giới hạn, mỗi task cần head riêng

Công nghệ sử dụng - Back End

- Cấu hình mô hình:
  - Sử dụng mô hình “T5-base”
    - Đủ mạnh để phân biệt 27 intent khác nhau
    - Độ chính xác cao hơn “T5-Small”
  - max_length = 80
    - Cắt chuỗi tại 80 token vì hầu hết câu lệnh intent đều khá ngắn
    - Giảm VRAM so với dùng, tránh OOM
    - Vẫn đủ giữ nguyên ngữ cảnh cho intent
  - batch_size = 16:
    - Cân bằng giữa tốc độ và bộ nhớ, tránh bị tràn ram trên 8Gb VRAM
    - Tăng throughput so với batch nhỏ hơn
  - epochs = 5 :
    - Đủ để mô hình học sâu ranh giới 27 lớp intent mà không quá lâu
    - Giảm nguy cơ overfitting so với epochs quá cao
  - Learning_rate = 3e-5
    - “T5-base” thường dùng LR trong khoảng 1e-5 -> 3e-4
    - 3e-5 là điểm an toàn vì đủ lớn để converge nhanh, đủ nhỏ để ổn định
    - Phù hợp khi fine-tune

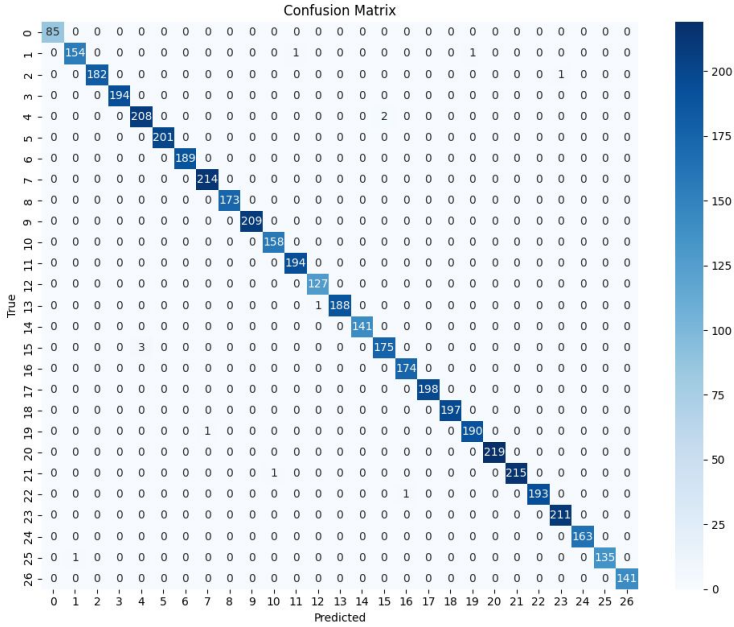
```
    # Model configuration
    self.model_name = "t5-base"
    self.max_length = 80
    self.batch_size = 16
    self.epochs = 5
    self.learning_rate = 3e-5
```

Công nghệ sử dụng - Back End

- Quá trình training:
  - Loss giảm rất tốt
    - Ban đầu : loss = 3.2981
    - Cuối cùng : loss = 0.0304
    - Mô hình học rất hiệu quả, loss giảm đều qua thời gian cho thấy nó đang học từ dữ liệu và hội tụ tốt.
  - Learning Rate giảm đều
    - Việc giảm learning rate theo thời gian là hợp lý, giúp mô hình fine-tune tốt hơn ở các epoch sau.
- Ta thấy Mô hình học rất tốt, loss đều.
- Scheduler hoạt động ổn định.
- Không có dấu hiệu quá khớp (overfitting) rõ ràng từ loss

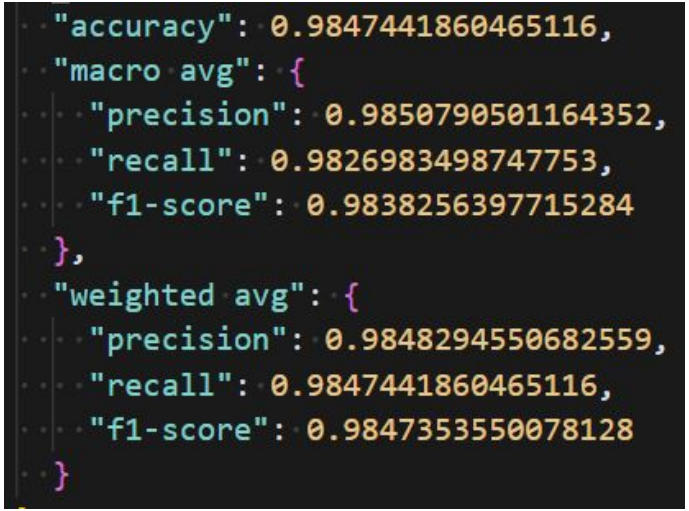
Công nghệ sử dụng - Back End

- Kết quả quá trình huấn luyện mô hình:
  - Ta thấy một số lỗi nhỏ :
    - Lớp 1: có dự đoán nhầm 1 mẫu sang lớp 11 và 24.
    - Lớp 5: có 2 mẫu nhầm sang lớp 4.
    - Lớp 16: có 3 mẫu nhầm sang lớp 1.
    - Lớp 22: có 1 mẫu nhầm sang lớp 24.
    - → Các lỗi này rất ít, nên có thể là do một số mẫu khó phân biệt.
  - Nhận định :
    - Mô hình hoạt động rất tốt.
    - Lỗi rất ít và phân bố ngẫu nhiên → không có bias mạnh
    - Các lớp có số mẫu khác nhau nhưng mô hình vẫn xử lý tốt

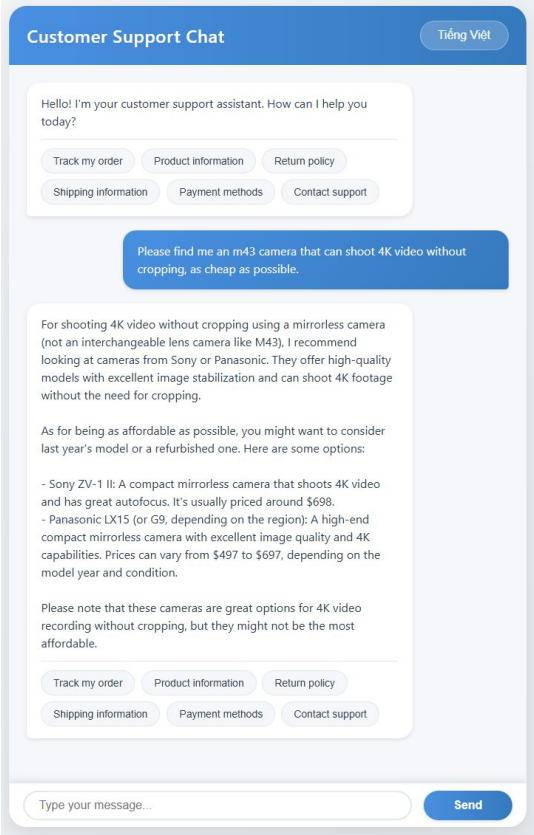


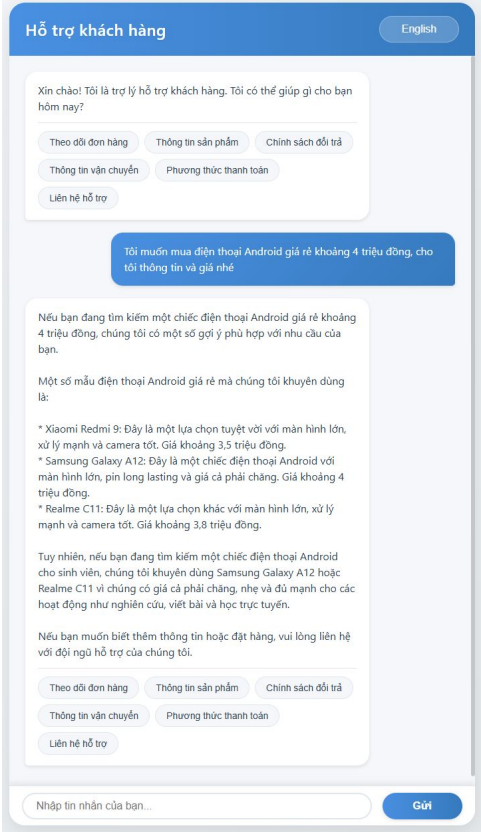
Công nghệ sử dụng - Back End

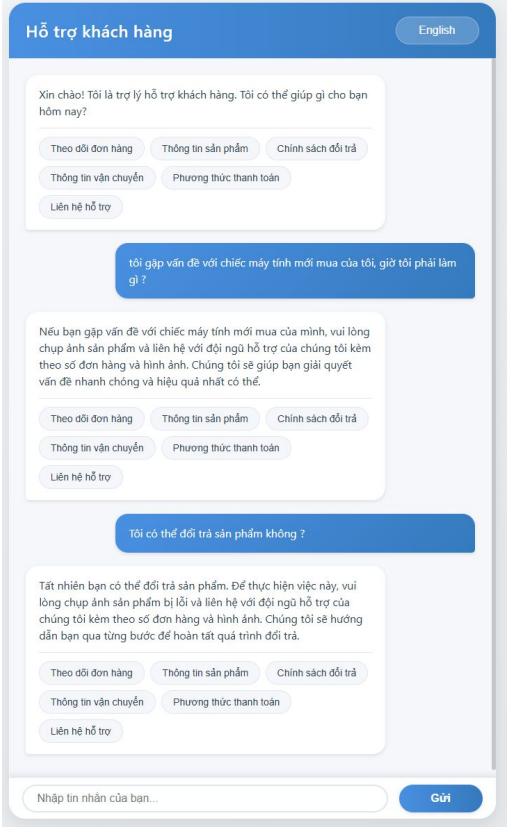
- Kết quả quá trình huấn luyện mô hình:
  - Accuracy ở mức 0.9847 (98.47%) cho thấy mô hình hoạt động rất tốt với phần lớn các dự đoán đúng. Đây là một chỉ số tốt để đánh giá tổng thể hiệu suất mô hình.
  - Macro Average
    - Precision : 0.9851 cho thấy mô hình rất chính xác trong việc dự đoán đúng các lớp với ít lỗi False Positive
    - Recall : 0.9827 cho thấy mô hình rất ít bỏ sót các lớp thực tế. Điều này chỉ ra rằng mô hình không bỏ qua nhiều mẫu thuộc các lớp quan trọng.
  - Weighted Average
    - Precision : 0.9848 cho thấy mô hình có độ chính xác rất cao trong tất cả các lớp, đặc biệt là các lớp với số lượng mẫu lớn.
    - Recall : 0.9847 phản ánh khả năng của mô hình trong việc phát hiện các mẫu từ tất cả các lớp với độ hoàn chỉnh cao.
- Đánh giá chung :
  - Mô hình hoạt động rất tốt : các chỉ số Accuracy, Precision, Recall và F1-Score đều ở mức rất cao đặc biệt trên cả Macro Average và Weighted average.
  - Độ chính xác cao, ít xảy ra lỗi False Positive và False Negative.
  - Không có dấu hiệu của mất cân bằng dữ liệu.



Demo


Customer Support Chat
Tiếng Việt
Hello! I'm your customer support assistant. How can I help you today?
Track my order
Product information
Return policy
Shipping information
Payment methods
Contact support
Please find me an m43 camera that can shoot 4K video without cropping, as cheap as possible.
For shooting 4K video without cropping using a mirrorless camera (not an interchangeable lens camera like M43), I recommend looking at cameras from Sony or Panasonic. They offer high-quality models with excellent image stabilization and can shoot 4K footage without the need for cropping.
As for being as affordable as possible, you might want to consider last year's model or a refurbished one. Here are some options:
- Sony ZV-1 II: A compact mirrorless camera that shoots 4K video and has great autofocus. It's usually priced around $698.
- Panasonic LX15 (or G9, depending on the region): A high-end compact mirrorless camera with excellent image quality and 4K capabilities. Prices can vary from $497 to $697, depending on the model year and condition.
Please note that these cameras are great options for 4K video recording without cropping, but they might not be the most affordable.
Track my order
Product information
Return policy
Shipping information
Payment methods
Contact support
Type your message...
Send


Hỗ trợ khách hàng
English
Xin chào! Tôi là trợ lý hỗ trợ khách hàng. Tôi có thể giúp gì cho bạn hôm nay?
Theo dõi đơn hàng
Thông tin sản phẩm
Chính sách đổi trả
Thông tin vận chuyển
Phương thức thanh toán
Liên hệ hỗ trợ
Tôi muốn mua điện thoại Android giá rẻ khoảng 4 triệu đồng, cho tôi thông tin và giá nhé
Nếu bạn đang tìm kiếm một chiếc điện thoại Android giá rẻ khoảng 4 triệu đồng, chúng tôi có một số gợi ý phù hợp với nhu cầu của bạn.
Một số mẫu điện thoại Android giá rẻ mà chúng tôi khuyên dùng là:
* Xiaomi Redmi 9: Đây là một lựa chọn tuyệt vời với màn hình lớn, xử lý mạnh và camera tốt. Giá khoảng 3,5 triệu đồng.
* Samsung Galaxy A12: Đây là một chiếc điện thoại Android với màn hình lớn, pin long lasting và giá cả phải chăng. Giá khoảng 4 triệu đồng.
* Realme C11: Đây là một lựa chọn khác với màn hình lớn, xử lý mạnh và camera tốt. Giá khoảng 3,8 triệu đồng.
Tuy nhiên, nếu bạn đang tìm kiếm một chiếc điện thoại Android cho sinh viên, chúng tôi khuyên dùng Samsung Galaxy A12 hoặc Realme C11 vì chúng có giá cả phải chăng, nhẹ và đủ mạnh cho các hoạt động như nghiên cứu, viết bài và học trực tuyến.
Nếu bạn muốn biết thêm thông tin hoặc đặt hàng, vui lòng liên hệ với đội ngũ hỗ trợ của chúng tôi.
Theo dõi đơn hàng
Thông tin sản phẩm
Chính sách đổi trả
Thông tin vận chuyển
Phương thức thanh toán
Liên hệ hỗ trợ
Nhập tin nhắn của bạn...
Gửi


Hỗ trợ khách hàng
English
Xin chào! Tôi là trợ lý hỗ trợ khách hàng. Tôi có thể giúp gì cho bạn hôm nay?
Theo dõi đơn hàng
Thông tin sản phẩm
Chính sách đổi trả
Thông tin vận chuyển
Phương thức thanh toán
Liên hệ hỗ trợ
tôi gặp vấn đề với chiếc máy tính mới mua của tôi, giờ tôi phải làm gì ?
Nếu bạn gặp vấn đề với chiếc máy tính mới mua của mình, vui lòng chụp ảnh sản phẩm và liên hệ với đội ngũ hỗ trợ của chúng tôi kèm theo số đơn hàng và hình ảnh. Chúng tôi sẽ giúp bạn giải quyết vấn đề nhanh chóng và hiệu quả nhất có thể.
Theo dõi đơn hàng
Thông tin sản phẩm
Chính sách đổi trả
Thông tin vận chuyển
Phương thức thanh toán
Liên hệ hỗ trợ
Tôi có thể đổi trả sản phẩm không ?
Tất nhiên bạn có thể đổi trả sản phẩm. Để thực hiện việc này, vui lòng chụp ảnh sản phẩm bị lỗi và liên hệ với đội ngũ hỗ trợ của chúng tôi kèm theo số đơn hàng và hình ảnh. Chúng tôi sẽ hướng dẫn bạn qua từng bước để hoàn tất quá trình đổi trả.
Theo dõi đơn hàng
Thông tin sản phẩm
Chính sách đổi trả
Thông tin vận chuyển
Phương thức thanh toán
Liên hệ hỗ trợ
Nhập tin nhắn của bạn...
Gửi

Khả năng mở rộng trong tương lai

- **Front end :**
  - Thêm ngôn ngữ mới vô cùng đơn giản nhờ cấu trúc component và sử dụng các tệp localization.
    - -> Hỗ trợ thêm bất kỳ ngôn ngữ nào trong vài bước.
  - Tối ưu hóa giao diện như thay đổi màu sắc, bố cục.
  - Thêm các tính năng như đánh giá sao, lịch sử chat…
- **Back end :**
  - Hỗ trợ, nâng cấp tối ưu hóa mô hình tự huấn luyện để mang lại kết quả tốt hơn.
  - Mở rộng dữ liệu huấn luyện để bổ sung thêm nhiều dữ liệu, ngữ cảnh, sản phẩm mới của thị trường.
  - Tích hợp thêm các dịch vụ như Ticketing System, gọi điện trực tiếp tổng đài…

Live Demo
và
Q&A

Cảm ơn thầy đã theo dõi