TRƯỜNG ĐẠI HỌC THỦY LỢI



BÁO CÁO Bài Tập lớn môn xử lý ngôn ngữ

tự nhiên

XÂY DỰNG ỨNG DỤNG HỖ TRỢ HỌC SINH học trung học cơ sở học toán sử dụng Phobert, Faiss và Fuzzy Matching

Sinh viên thực hiện : Nguyễn Văn Quân

MSV : 2251262627

Lớp : 64TTNT1

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | 1. PGS.TS. NGUYỄN QUANG HOAN |

Hà Nội, năm 2025

MỤC LỤC

**[MỞ ĐẦU](#_Toc10127)** [4](#_Toc10127)

[C](#_Toc25475)**[HƯƠNG 1 GIỚI THIỆU BÀI TOÁN](#_Toc25475)** [5](#_Toc25475)

[1. Giới thiệu bài toán 5](#_Toc10930)

[2. Mục tiêu 6](#_Toc15153)

**[CHƯƠNG 2 GIỚI THIỆU BỘ DỮ LIỆU](#_Toc4608)** [7](#_Toc4608)

[1. Giới thiệu tập dữ liệu 7](#_Toc1372)

**[CHƯƠNG 3 CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ](#_Toc20952)** [8](#_Toc20952)

[3.1 PhoBERT – Mô hình ngôn ngữ tiếng Việt dựa trên BERT 8](#_Toc3813)

[3.2 FAISS 9](#_Toc28078)

[3.3 Fuzzy Matching – Tìm kiếm gần đúng theo ký tự 10](#_Toc22329)

**[CHƯƠNG 4 PHÂN TÍCH HỆ THỐNG](#_Toc29144)** [11](#_Toc29144)

[4.1 Kiến trúc tổng thể của hệ thống 11](#_Toc8556)

[4.2 Luồng xử lý của học sinh 12](#_Toc17611)

[4.2.1 Hỏi lý thuyết 12](#_Toc29449)

[4.2.2 Yêu cầu bài tập 12](#_Toc21292)

[C](#_Toc5730)**[HƯƠNG 5 THỰC NGHIỆM VÀ KẾT QUẢ](#_Toc5730)** [13](#_Toc5730)

[5.1 Trực quan hóa và tiền xử lý dữ liệu 13](#_Toc30037)

[5.1.1 Số lượng câu hỏi theo từng lớp 13](#_Toc9321)

[5.1.2 Số lượng câu hỏi theo từng chủ đề 14](#_Toc8786)

[5.1.3 Số lượng bài tập theo từng thể loại 15](#_Toc31326)

[5.1.3 Tỷ lệ câu hỏi có hướng dẫn giải 16](#_Toc25427)

[5.2 Tiêu chí đánh giá 16](#_Toc25042)

[5.3 Kết quả thực nghiệm 17](#_Toc21548)

**[CHƯƠNG 6 DEMO GIAO DIỆN](#_Toc9675)** [18](#_Toc9675)

[Demo online: https://hecwdkaavo9t2pdu49nuan.streamlit.app/ 18](#_Toc8284)

[6.1 Giao diện lý thuyết 18](#_Toc26129)

[6.2 Giao diện bài tập 19](#_Toc10663)

**[CHƯƠNG 7 KẾT LUẬN](#_Toc16883)** [20](#_Toc16883)

[TÀI LIỆU THAM KHẢO 21](#_Toc11352)

**MỞ ĐẦU**

Trong bối cảnh công nghệ thông tin ngày càng phát triển, trí tuệ nhân tạo (AI) đang dần khẳng định vai trò quan trọng trong nhiều lĩnh vực của đời sống, đặc biệt là giáo dục. Việc ứng dụng các hệ thống thông minh vào quá trình giảng dạy và học tập không chỉ góp phần nâng cao chất lượng truyền đạt kiến thức mà còn mở ra cơ hội học tập linh hoạt, hiệu quả và cá nhân hóa cho từng học sinh.

Một trong những hướng ứng dụng nổi bật của AI trong giáo dục là việc phát triển các **ứng dụng hỗ trợ học tập**, đặc biệt dưới hình thức **trợ lý ảo** có khả năng giao tiếp bằng ngôn ngữ tự nhiên. Các ứng dụng này không chỉ giúp học sinh giải đáp thắc mắc, hướng dẫn làm bài tập mà còn có thể tương tác trực tiếp để hỗ trợ học tập theo sát chương trình giảng dạy chính thống.

Từ nhu cầu thực tế trong việc nâng cao hiệu quả học tập môn Toán – một môn học cốt lõi trong chương trình trung học cơ sở – đề tài “**Xây dựng ứng dụng hỗ trợ học sinh trung học cơ sở học Toán**” được đề xuất. Ứng dụng được thiết kế như một hệ thống chatbot thông minh có khả năng tương tác bằng tiếng Việt, hỗ trợ học sinh ôn tập lý thuyết, luyện tập bài tập và củng cố kiến thức Toán học theo chương trình phổ thông cơ bản.

Hệ thống tích hợp nhiều công nghệ hiện đại như xử lý ngôn ngữ tự nhiên (NLP), và giao diện người dùng thân thiện nhằm mang đến một trải nghiệm học tập hiệu quả, tiện lợi và gần gũi. Đây là một giải pháp thiết thực giúp học sinh tự học tốt hơn.

# GIỚI THIỆU BÀI TOÁN

1. **Giới thiệu bài toán**

Trong chương trình giáo dục phổ thông, đặc biệt ở bậc trung học cơ sở, môn Toán đóng vai trò then chốt trong việc hình thành tư duy logic, kỹ năng giải quyết vấn đề và nền tảng cho các môn học tự nhiên khác[1]. Tuy nhiên, không ít học sinh gặp khó khăn trong quá trình tiếp thu kiến thức Toán học do nhiều nguyên nhân như: chương trình học nặng, phương pháp học chưa phù hợp, thiếu sự hỗ trợ kịp thời khi tự học tại nhà.[2]

Trong khi đó, nhu cầu học tập linh hoạt và khả năng tiếp cận nguồn tài nguyên giáo dục ngoài giờ học chính khóa ngày càng tăng cao. Đặc biệt, sau đại dịch COVID-19, mô hình học tập kết hợp và học tập tự định hướng trở nên phổ biến, đòi hỏi sự hỗ trợ từ các công cụ công nghệ thông minh nhằm tăng cường hiệu quả học tập.

Từ thực tế đó, bài toán đặt ra là: “ **Làm thế nào để xây dựng một hệ thống chatbot có khả năng giao tiếp tự nhiên bằng tiếng Việt, hỗ trợ học sinh trung học cơ sở học Toán hiệu quả hơn, cá nhân hóa hơn và chủ động hơn**? “

Việc giải quyết bài toán này đòi hỏi sự kết hợp giữa các công nghệ như **xử lý ngôn ngữ tự nhiên (NLP)**, **mô hình ngôn ngữ sâu** như PhoBERT, cùng với kỹ thuật thiết kế hệ thống phần mềm thân thiện, dễ sử dụng đối với lứa tuổi học sinh trung học cơ sở.

1. **Mục tiêu**

Đề tài hướng đến việc xây dựng một hệ thống chatbot hỗ trợ học tập dành cho học sinh trung học cơ sở, tích hợp trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên để mang lại trải nghiệm học tập chủ động và tương tác. Các mục tiêu cụ thể bao gồm:

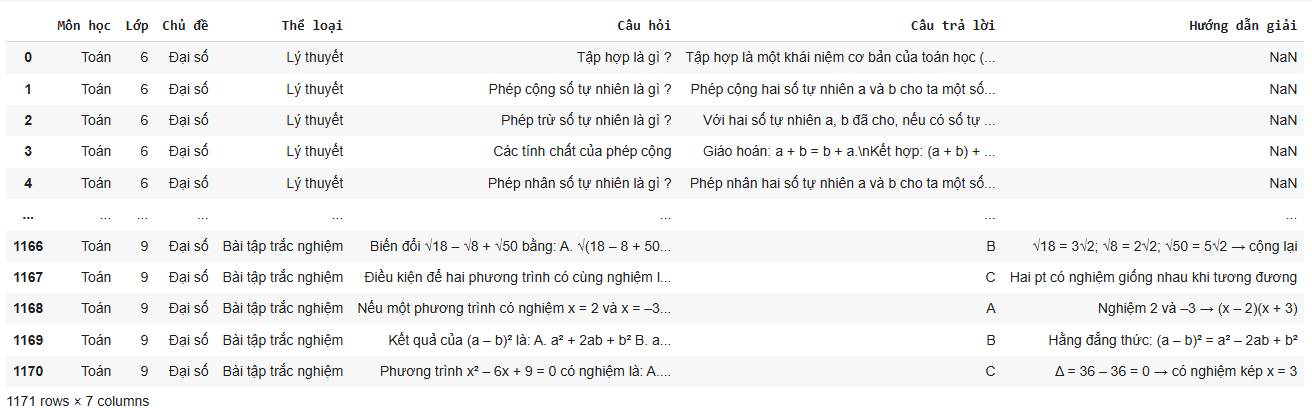
1. **Xây dựng bộ dữ liệu câu hỏi – đáp và bài tập học tập** dựa trên chương trình giáo dục trung học cơ sở, bao gồm các trường thông tin như: môn học, lớp học, thể loại, độ khó, câu hỏi, đáp án và hướng dẫn giải.
2. **Phát triển hệ thống chatbot tiếng Việt có khả năng hiểu ngữ nghĩa**, sử dụng mô hình PhoBERT để xử lý và trích xuất đặc trưng ngữ nghĩa từ câu hỏi đầu vào của người dùng.
3. **Triển khai cơ chế tìm kiếm thông minh kết hợp giữa tìm kiếm ngữ nghĩa (FAISS) và tìm kiếm mờ (fuzzy matching)** nhằm cải thiện độ chính xác và tính linh hoạt khi người dùng đặt câu hỏi với nhiều cách diễn đạt khác nhau.
4. **Tích hợp chức năng bài tập tương tác**, học sinh sẽ bảo chatbot đưa ra bài tập với các chủ đề , lớp , thể loại . Học sinh sẽ trả lời và chatbot sẽ đưa ra phản hồi và hướng dẫn học tập cụ thể.
5. **Thiết kế giao diện người dùng bằng Streamlit** đảm bảo tính thân thiện, dễ sử dụng và phù hợp với lứa tuổi học sinh trung học cơ sở

# GIỚI THIỆU BỘ DỮ LIỆU

## Giới thiệu tập dữ liệu

Bộ dữ liệu chứa nội dung giáo dục môn Toán, cho học sinh lớp 6 đến 9, bao gồm các câu hỏi lý thuyết và bài tập về đại số, hình học. Dữ liệu được phân loại theo lớp, chủ đề và thể loại.

Bộ dữ liệu được thu thập thủ công từ Sách giáo khoa và các trang web giáo dục uy tín như hocmai.vn, vietjack.com, loigiaihay.com, kenhgiaovien.com,... Hiện tại bộ dữ liệu gồm hơn 1100 dòng, mỗi dòng là mỗi dòng là một câu hỏi, được phân loại rõ ràng theo các cột và sẽ được tiếp tục cập nhật thêm.

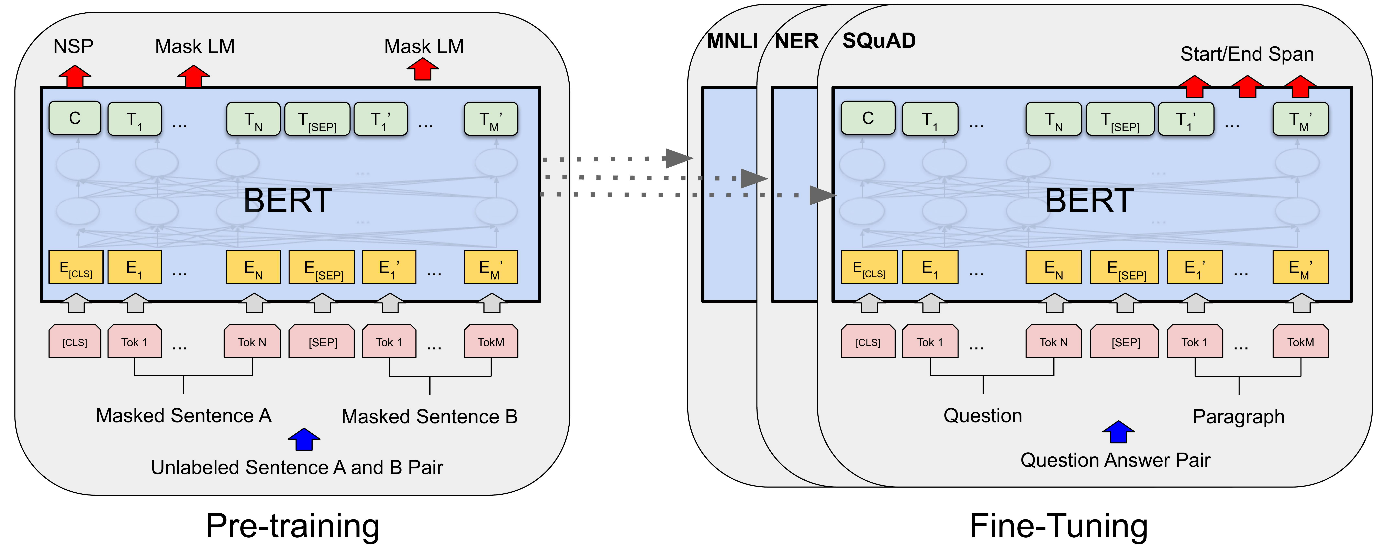


Cấu trúc dữ liệu có 7 đặc trưng gồm: Môn học, Lớp, Chủ đề, Thể loại, Câu hỏi, Câu trả lời, Hướng dẫn giải.

# CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ

## PhoBERT – Mô hình ngôn ngữ tiếng Việt dựa trên BERT

**PhoBERT** là một **mô hình ngôn ngữ tiền huấn luyện (pre-trained language model)** được phát triển dành riêng cho **tiếng Việt**, dựa trên kiến trúc **BERT (Bidirectional Encoder Representations from Transformers).** PhoBERT được xây dựng bởi nhóm nghiên cứu **VinAI Research** (Việt Nam) và đã đạt hiệu quả rất cao trong nhiều tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt. Mô hình này được ứng dụng rộng rãi trong các bài toán xử lý ngôn ngữ tự nhiên như phân loại văn bản, trích xuất thực thể, trả lời câu hỏi và xây dựng chatbot tiếng Việt.



*Nguồn: Phân loại về chất bài báo cáo mạng sử dụng mô hình PhoBert*

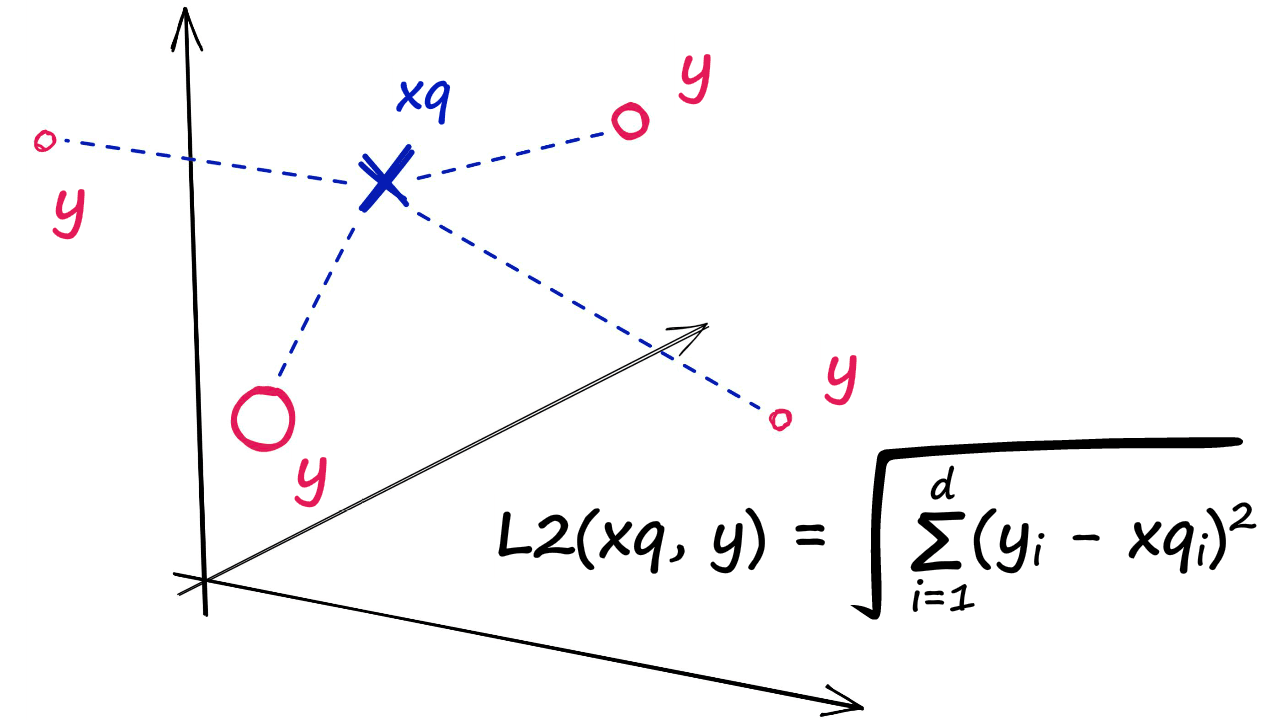
Chức năng của Phobert trong đề tài :

- **Biến câu hỏi tiếng Việt thành vector ngữ nghĩa** (dạng số) – gọi là embedding.

- Sau đó chuyển sang FAISS

## FAISS

FAISS (Facebook AI Similarity Search) là thư viện mã nguồn mở do Meta phát triển, chuyên dùng để tìm kiếm nhanh các vector có độ tương đồng cao trong không gian nhiều chiều. FAISS cho phép tìm kiếm hiệu quả ngay cả khi có hàng ngàn hoặc hàng triệu vector.



Chức năng của FAISS trong đề tài :

- Lưu toàn bộ câu hỏi trong CSDL dưới dạng vector embedding.

- Khi người dùng hỏi, FAISS **so sánh vector ngữ nghĩa** của câu hỏi mới với toàn bộ vector trong kho.

- Trả về **câu gần nghĩa nhất** (dựa vào khoảng cách cosine/L2).

- Nếu khoảng cách giữa câu hỏi người dùng và câu trong FAISS không đủ thì sẽ chuyển sang Fuzzy Matching

**3.3 Fuzzy Matching – Tìm kiếm gần đúng theo ký tự**

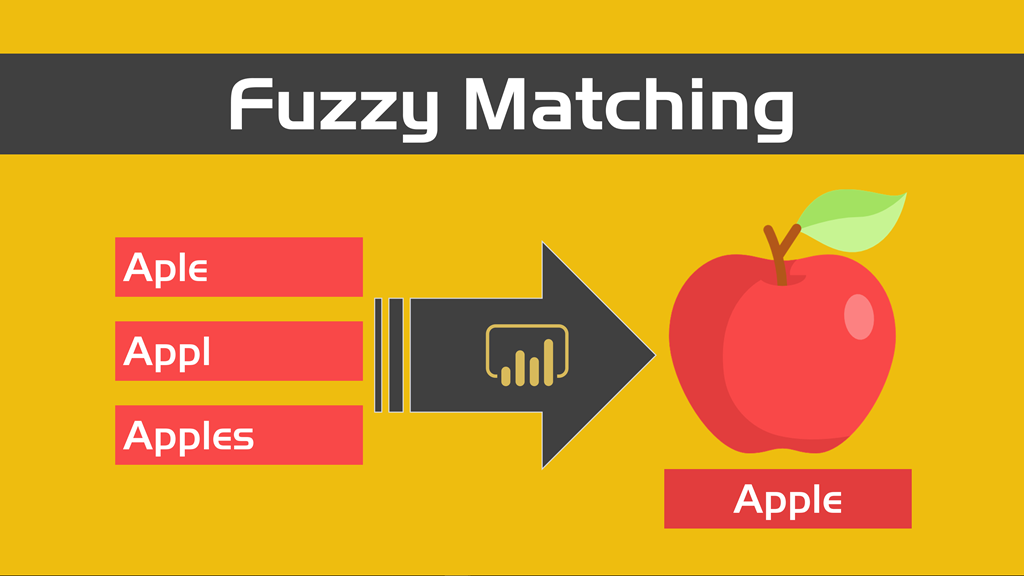
Bên cạnh tìm kiếm ngữ nghĩa bằng **FAISS**, hệ thống còn kết hợp với kỹ thuật **fuzzy matching** là phương pháp dùng để tìm kiếm hoặc so khớp hai chuỗi văn bản có **độ tương đồng gần đúng nhằm xử lý các trường hợp như :**

+ Người dùng gõ sai chính tả

+ Thiếu dấu tiếng Việt

+ Khác định dạng

Fuzzy matching so sánh chuỗi văn bản dựa trên **độ tương đồng ký tự** (string similarity), cho điểm số từ 0–100, giúp chatbot linh hoạt hơn khi xử lý dữ liệu đầu vào không hoàn hảo.



*Fuzzy Matching (Source: thepoweruser)*

# PHÂN TÍCH HỆ THỐNG

## Kiến trúc tổng thể của hệ thống

Hệ thống được chia thành 3 thành phần chính như sau:

1. **Giao diện người dùng**   
   Được xây dựng bằng **Streamlit**, giao diện cho phép học sinh:

+ Đặt câu hỏi bằng tiếng Việt

+ Nhận phản hồi dưới dạng văn bản, có giải thích rõ ràng

+ Yêu cầu bài tập, nhập câu trả lời và nhận đánh giá

1. **Bộ xử lý truy vấn và sinh phản hồi** Chức năng chính:

+ Phân tích câu hỏi đầu vào

+ Kiểm tra biểu thức toán và tính toán

+ Tìm kiếm câu hỏi tương tự trong cơ sở dữ liệu

+ Sinh câu trả lời phù hợp (trả lời lý thuyết hoặc chấm bài tập)

1. **Hệ thống tìm kiếm thông minh**   
   Bao gồm:

**+ PhoBERT**: mã hóa câu hỏi thành vector ngữ nghĩa

**+ FAISS**: tìm kiếm ngữ nghĩa các câu hỏi tương đồng

**+ Fuzzy Matching**: so sánh gần đúng theo ký tự, hỗ trợ lỗi chính tả hoặc cách viết khác nhau

## Luồng xử lý của học sinh

**4.2.1 Hỏi lý thuyết**

1. Học sinh nhập câu hỏi lý thuyết qua giao diện Streamlit.
2. Hệ thống sẽ kiểm tra có phải biểu thức toán không:

+ Phải thì sẽ tính biểu thức và đưa ra kết quả

+ Nếu không thì sẽ chuyển đến phần tiếp theo

1. Câu hỏi được xử lý bởi PhoBERT để tạo vector ngữ nghĩa.
2. Hệ thống FAISS tìm kiếm những câu hỏi có vector gần nhất trong cơ sở dữ liệu. Nếu độ tương đồng cao, chatbot truy xuất ngay câu trả lời tương ứng.
3. Nếu độ tương đồng thấp (tức câu hỏi mới lạ hoặc viết khác nhiều), chatbot dùng thêm **fuzzy matching** (so khớp mờ) để tìm câu hỏi gần đúng nhất về mặt ký tự.
4. Chatbot trả về phần **Câu trả lời** đã lưu trong dữ liệu.

**4.2.2 Yêu cầu bài tập**

1. Học sinh yêu cầu bài tập qua giao diện Streamlit.

2. Chatbot sẽ dẫn dắt học sinh lần lượt qua các bước chọn lớp,chủ đề ,thể loại

3. Chatbot lọc dữ liệu theo các tiêu chí: lớp, chủ đề và thể loại.

Nếu tìm được, chatbot hiển thị một câu hỏi ngẫu nhiên từ danh sách phù hợp.

4. Học sinh nhập câu trả lời cho bài toán đã cho.

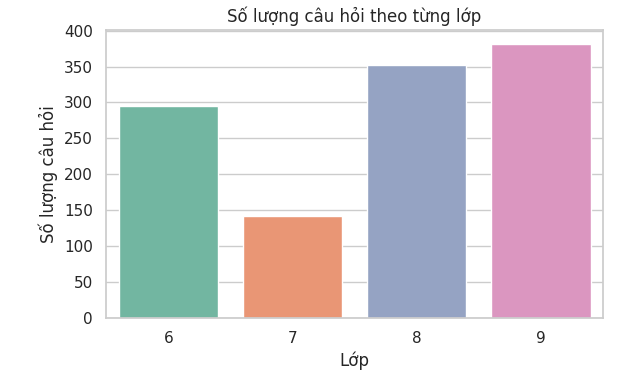
5. Nếu đúng: Chatbot trả lời **"Chính xác!"**

Nếu sai: Chatbot hiển thị **đáp án đúng** và kèm **hướng dẫn giải** (nếu có).

# CHƯƠNG 5 THỰC NGHIỆM VÀ KẾT QUẢ

## 5.1 Trực quan hóa và tiền xử lý dữ liệu

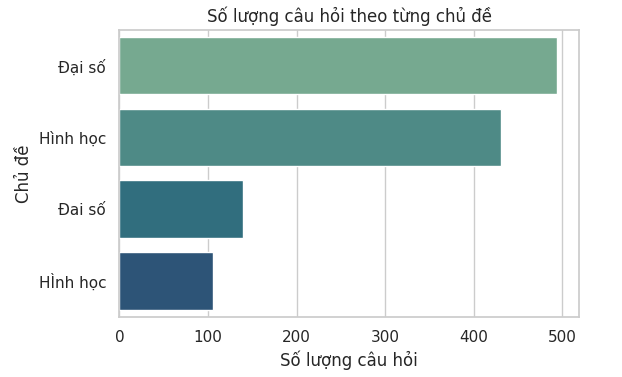
**5.1.1 Số lượng câu hỏi theo từng lớp**



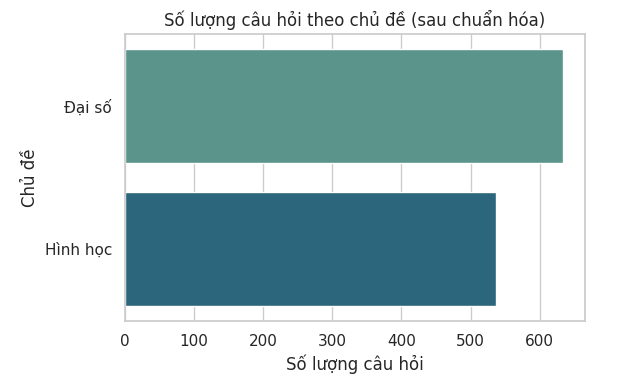
**Lớp 7** có **ít nội dung nhất ->** Cần **bổ sung dữ liệu cho lớp 7** để cân bằng nội dung, tránh tình trạng học sinh lớp 7 không được hỗ trợ đầy đủ.

Có thể xem xét **ưu tiên phát triển nội dung nâng cao cho lớp 9** (thi chuyển cấp)

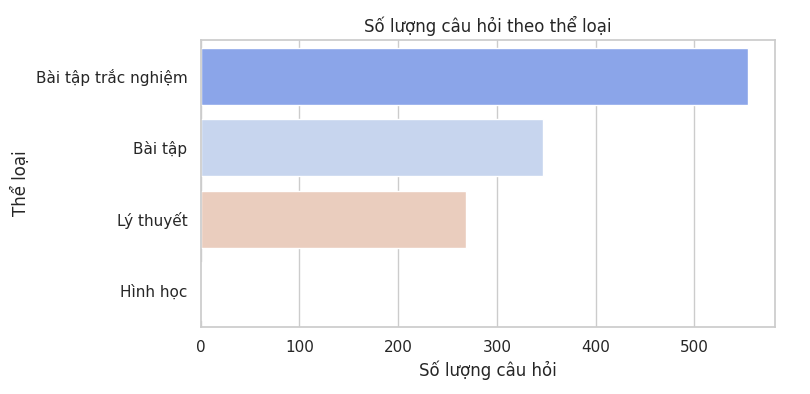
**5.1.2 Số lượng câu hỏi theo từng chủ đề**



Có thể thấy rằng chủ đề đang bị lặp lại vì chưa được chuẩn hóa . Sau khi chuẩn hóa bằng ánh xạ mapping :

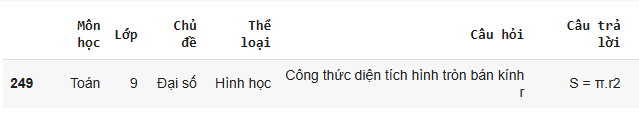


**5.1.3 Số lượng bài tập theo từng thể loại**

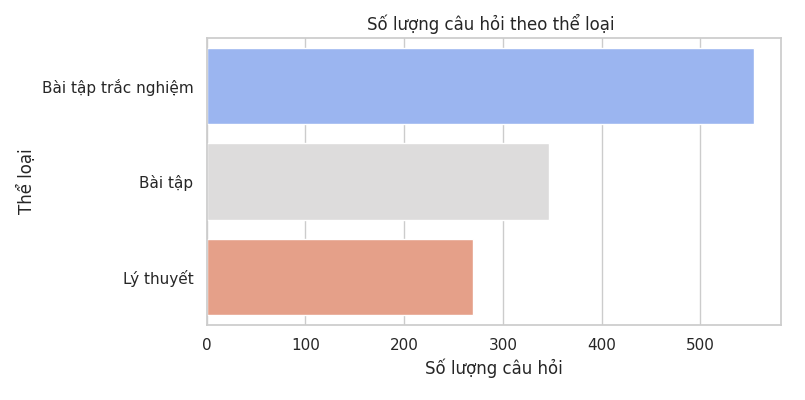


Trong cột thể loại có “ Hình học “ thuộc cột chủ đề bị gán sai vào thể loại.

Đầu tiên kiểm tra xem gán sai ở dữ liệu dòng nào



Sau đó chuyển “ Hình học “ sang cột “ Chủ đề “ và thay giá trị “ Thể loại “ thành “ Lý thuyết”



**5.1.3 Tỷ lệ câu hỏi có hướng dẫn giải**



Có tới **74.6%** câu hỏi có kèm theo **hướng dẫn giải**, chủ yếu thuộc thể loại **bài tập**. Điều này phản ánh sự đầu tư kỹ lưỡng nhằm hỗ trợ học sinh trong quá trình luyện tập và tự học.

Ngược lại, **25.4%** câu hỏi còn lại **không có hướng dẫn giải**, phần lớn rơi vào nhóm **câu hỏi lý thuyết** – những câu này thường chỉ yêu cầu học sinh ghi nhớ hoặc hiểu khái niệm, nên không nhất thiết phải có lời giải chi tiết.

## 5.2 Tiêu chí đánh giá

**Top-1 Accuracy**: tỷ lệ câu hỏi mà chatbot tìm được **đúng câu trả lời ở vị trí đầu tiên**

Top-1

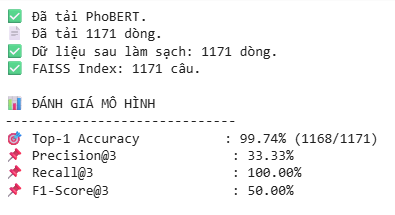
**Precision@k, Recall@k**: đo lường mức độ chính xác và khả năng bao phủ trong top-k kết quả truy xuất.

Precision@k

Recall@k

**F1-Score@k**: kết hợp **Precision@k** và **Recall@k** thành một giá trị duy nhất để đánh giá sự cân bằng giữa độ chính xác và độ bao phủ.

## 5.3 Kết quả thực nghiệm



**Top-1 Accuracy** đạt **99.74%** (1168/1171): cho thấy mô hình gần như luôn đưa ra câu trả lời đúng ở vị trí đầu tiên, đảm bảo trải nghiệm người dùng rất tốt.

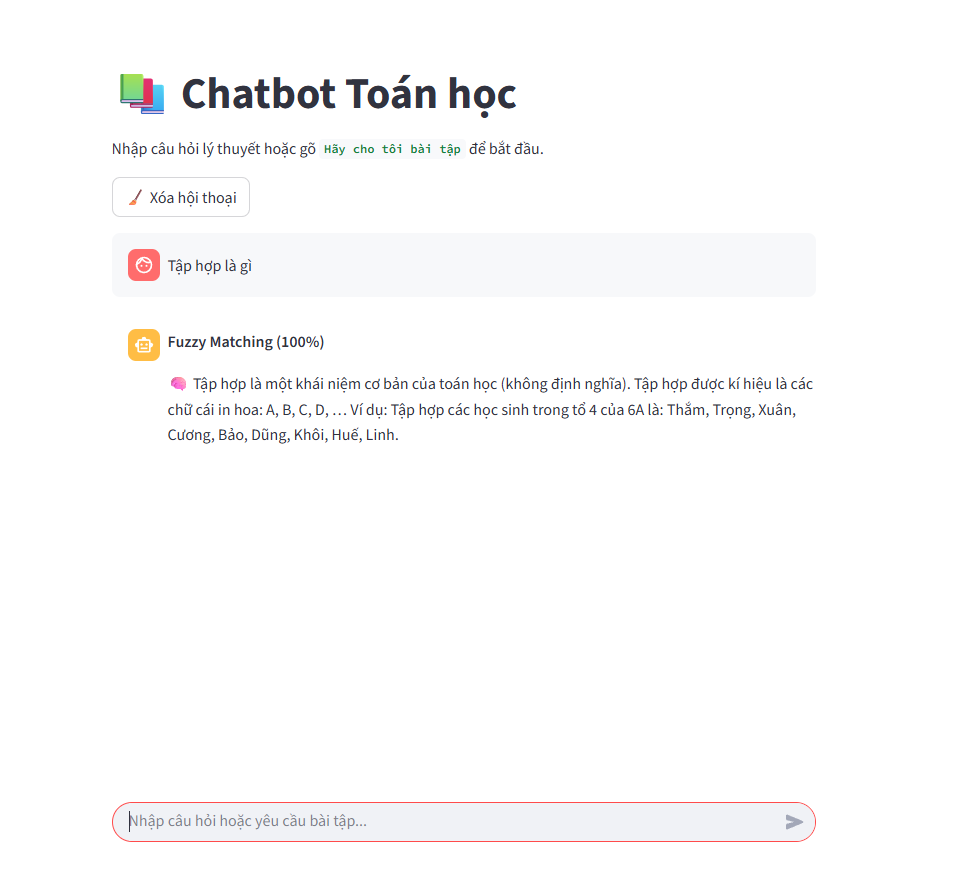
**Recall@3** đạt **100.00%**: mô hình luôn đưa câu trả lời đúng vào **top 3** kết quả truy xuất.

**Precision@3** chỉ đạt **33.33%**: chỉ 1/3 câu trả lời trong top 3 thực sự chính xác, cho thấy kết quả gợi ý còn nhiễu.

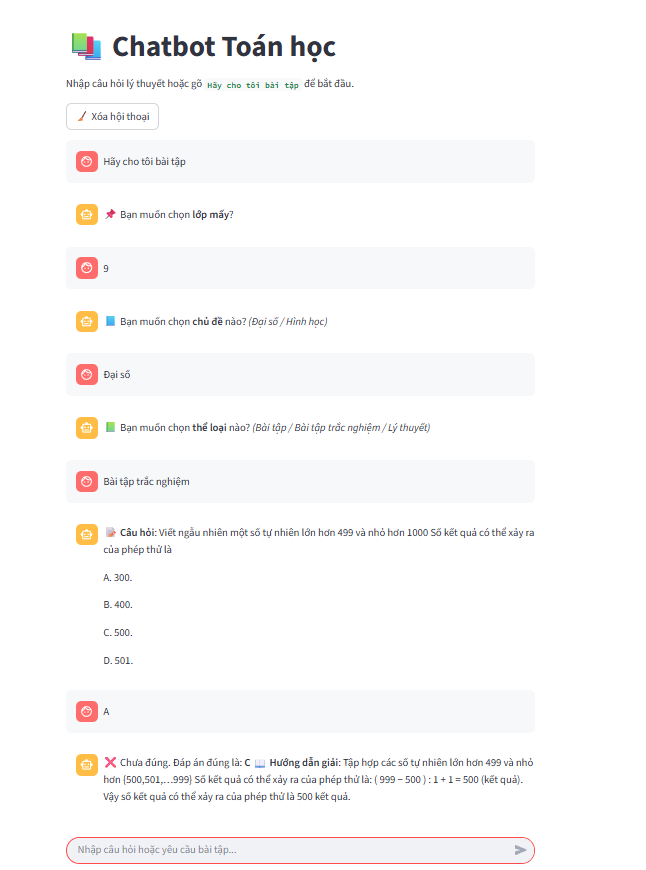
# CHƯƠNG 6 DEMO GIAO DIỆN

## Demo online: <https://hecwdkaavo9t2pdu49nuan.streamlit.app/>

## 6.1 Giao diện lý thuyết



**6.2 Giao diện bài tập**



# CHƯƠNG 7 KẾT LUẬN

Trong thời đại giáo dục 4.0, việc tích hợp các công nghệ trí tuệ nhân tạo vào quá trình dạy và học ngày càng đóng vai trò quan trọng, đặc biệt là trong việc cá nhân hóa trải nghiệm học tập và nâng cao khả năng tự học cho học sinh. Nghiên cứu này đã xây dựng một hệ thống chatbot tiếng Việt ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên hiện đại như PhoBERT [3], tìm kiếm ngữ nghĩa FAISS và so khớp mờ Fuzzy Matching để hỗ trợ học sinh trung học cơ sở học môn Toán một cách hiệu quả, chủ động và thân thiện hơn.

Hệ thống cho phép học sinh đặt câu hỏi lý thuyết, yêu cầu bài tập theo lớp, chủ đề và thể loại, đồng thời đánh giá mức độ hiểu bài thông qua tương tác đa lượt. Kết quả thử nghiệm cho thấy độ chính xác trả lời ở mức cao (Top-1 Accuracy 99,74%), khả năng mở rộng linh hoạt và phù hợp với nhu cầu học tập tự định hướng – một xu hướng nổi bật trong thời kỳ hậu đại dịch COVID-19 [4].

Trong tương lai, hệ thống có thể được mở rộng để hỗ trợ nhiều môn học khác, tích hợp giọng nói, hình ảnh và các mô hình học sâu nhằm tự động thích nghi theo năng lực cá nhân người học. Đây là một hướng đi tiềm năng góp phần đổi mới phương pháp giáo dục phổ thông theo hướng hiện đại và thông minh hơn.

TÀI LIỆU THAM KHẢO

[1]<https://moet.gov.vn/content/vanban/Lists/VBPQ/Attachments/1483/vbhn-chuong-trinh-tong-the.pdf>

1. [Đặc điểm của học sinh khó khăn trong học toán, nguyên nhân dẫn đến tình trạng học sinh khó khăn trong học toán Lớp 10 Trung học Phổ thông - Nguyễn Thuy Phương Tâm: 18 HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1075.2018-0058 Educational Sciences, 2018, Vol](https://www.thuvientailieu.vn/tai-lieu/dac-diem-cua-hoc-sinh-kho-khan-trong-hoc-toan-nguyen-nhan-dan-den-tinh-trang-hoc-sinh-kho-khan-trong-hoc-toan-lop-10-50541/)

[3][Nguyen, D. Q., & Nguyen, T. Q. (2020). PhoBERT: Pre-trained language models for Vietnamese. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.](https://arxiv.org/pdf/2003.00744)

[4] [Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. Educause Review.](https://www.scirp.org/reference/referencespapers?referenceid=3163024)