BÁO CÁO ĐỒ ÁN MÔN KỸ THUẬT LẬP TRÌNH TRÍ TUỆ NHÂN TẠO PHẦN XỬ LÝ NGÔN NGỮ TỰ NHIÊN IELTS CHATBOT

Huỳnh Đặng Vĩnh Hiền[1,2] and **Dương Thị Tú Yến**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{21520029,23521846}@gm.uit.edu.vn

Tóm tắt

Đây là báo cáo đồ án cuối kỳ của môn CS311 - Kỹ thuật lập trình trí tuệ nhân tạo. Thông qua đồ án này, nhóm tiến hành xây dựng một chatbot hỗ trợ người dùng ôn tập cho bài thi IELTS. Chatbot này được xây dựng trên hướng tiếp cận *Retrieval-augmented generation (RAG)* (IBM Research, 2023) với tính năng giúp ôn luyện hai kỹ năng *Reading (Đọc)* và *Speaking (Nói)* của bài thi IELTS. Kết quả đánh giá chatbot sau khi xây dựng cho thấy hệ thống có khả năng xử lý tốt các yêu cầu của người dùng ở mức độ dễ đến trung bình, nhưng vẫn bộc lộ nhiều điểm yếu khi gặp phải những prompt phức tạp. Ngoài ra, thông qua đồ án này, nhóm cũng phân tích các điểm yếu của hệ thống vừa được xây dựng và hướng phát triển tiếp theo cho hệ thống. Mã nguồn của chatbot được công khai tại https://github.com/vinhhien323/CS311-Hien-Yen.

1 Tổng quan

Trong thời đại toàn cầu hóa, chứng chỉ Tiếng Anh nói chung và IELTS ngày càng trở nên phổ biến và đóng vai trò quan trọng. Tuy nhiên, nhiều học viên gặp khó khăn trong việc tìm kiếm tài liệu học phù hợp và xây dựng lộ trình học tập hiệu quả. Điều này đặt ra yêu cầu cấp thiết về một công cụ hỗ trợ học IELTS một cách tiện lợi, tiết kiệm thời gian và nâng cao hiệu quả học tập. Do đó chúng tôi xây dựng một chatbot hỗ trợ người học IELTS dựa trên kiến trúc RAG (IBM Research, 2023), kết hợp với các công cụ như GeminiEmbedding (Google AI, 2025a), Gemini-1.5-Flash (Google AI, 2025b), FAISS (Johnson et al., 2017). Về phần giao diện, chúng tôi sử dụng Streamlit (Streamlit, 2025) vì khả năng tích hợp trực tiếp với Python, dễ sử dụng và không yêu cầu cấu hình phức tạp.

2 Phân công công việc

Trong quá trình thực hiện đồ án, nhóm đã phân chia công việc cụ thể giữa các thành viên để đảm bảo tiến độ và chất lượng dự án. Chi tiết về việc phân công nhiệm vụ giữa các thành viên được thể hiện trong Table 1.

Công việc	Hiền	Yến
Khảo sát phương pháp	✓	✓
Chuẩn bị dữ liệu	✓	✓
Lập trình frontend		✓
Lập trình backend	✓	
Đánh giá hệ thống	✓	✓
Làm slide	✓	✓
Viết báo cáo	✓	✓

Table 1: Bảng phân công công việc giữa các thành viên thực hiện đồ án

3 Các tính năng của đồ án

Hệ thống chatbot được phát triển tập trung vào việc hỗ trợ người học IELTS với các tính năng chính như sau:

- *Trò chuyện với chatbot*: Chatbot tương tác với người dùng thông qua hội thoại, cung cấp câu trả lời phù hợp với ngữ cảnh dựa trên các câu hỏi và yêu cầu liên quan đến IELTS. Hệ thống có khả năng xử lý cả ngôn ngữ tự nhiên và các truy vấn ở mức độ dễ đến trung bình.

- *Tìm kiếm các bài học theo chủ đề*: Người dùng có thể tìm kiếm các bài học dựa trên các chủ đề cụ thể. Chatbot sẽ gợi ý các bài học liên quan kèm theo nội dung (bao gồm bài đọc, các câu hỏi) phù hợp với từng chủ đề.

- *Hiển thị tài liệu PDF*: Hệ thống tích hợp chức năng hiển thị tài liệu PDF từ nguồn uy tín ngay trong giao diện chatbot, cho phép người học truy cập nhanh các tài liệu luyện thi IELTS.

- *Quản lý hội thoại và lưu trữ*: Chatbot lưu trữ lịch sử hội thoại và bài học mà người dùng đã thực hiện. Người học có thể truy cập lại các hội thoại trước đó để xem lại câu trả lời hay phân loại các chủ đề trò chuyện.

4 Quá trình thu thập và xử lý dữ liệu

Dữ liệu cho hệ thống được xây dựng với mục tiêu phục vụ trực tiếp hai kỹ năng trong kỳ thi IELTS: *Reading (Đọc)* và *Speaking (Nói)*. Chúng tôi đã tiến hành thu thập, xử lý và chuẩn bị dữ liệu theo các bước cụ thể như sau:

1. *Thu thập dữ liệu*: Chúng tôi thu thập dữ liệu từ bộ sách IELTS Cambridge từ quyển 16 đến 19 và dữ liệu sẵn có từ trước.

2. *Sắp xếp và phân loại dữ liệu*: Để đảm bảo tính hệ thống và dễ dàng truy xuất, dữ liệu được phân loại theo danh sách các chủ đề chuẩn bị trước, bao gồm các lĩnh vực phổ biến trong kỳ thi IELTS. Để đảm bảo các chủ đề có tính cụ thể cao, chúng tôi tham khảo việc phân loại các chủ đề theo gợi ý từ ChatGPT. Chi tiết về danh sách các chủ đề có thể được xem tại:

   https://chatgpt.com/share/674dc757-1388-8009-bb30-b480fe48(

3. *Xử lý PDF*: Đối với các bài thuộc kỹ năng *Reading*, với mỗi mẫu data được xây dựng, chúng tôi tiến hành tạo và lưu trữ một file pdf chứa bài đọc và các câu hỏi tương ứng, giúp người sử dụng có thêm nguồn tài nguyên để luyện tập khi cần thiết.

4. *Tạo dữ liệu JSON*: Sau khi xử lý, dữ liệu được chuyển đổi sang định dạng JSON để dễ dàng tích hợp vào hệ thống chatbot. Figure 1 là ví dụ cho một phần của file JSON sau khi hoàn thành.

5 Kiến trúc hệ thống

Để phù hợp với mục tiêu ứng dụng các kiến thức, nội dung được tiếp thu trong quá trình học tập môn CS311, nhóm thực hiện đề tài tiến hành sử dụng hướng tiếp cận RAG (IBM Research, 2023) cho việc xây dựng chatbot.

Kiến trúc của một hệ thống sử dụng RAG điển hình được trình bày trong Figure 2, trong đó các thành phần *Embedder*, *Vector Store & Retrieval* và *Large Language Model (LLM)* sẽ được tùy biến phụ thuộc vào yêu cầu cụ thể của bài toán được yêu


```
{
    {
        "ID": "1731",
        "Title": "The thylacine",
        "Pages": "59–60",
        "Topics": [
            "Environment"
        ],
        "Type of Questions": [
            "Sentence Completion",
            "True/False/Not Given"
        ],
        "Origin": {
            "Book": 17,
            "Test": 3
        }
    },
}
```


Figure 1: Minh họa cho một mẫu dữ liệu do nhóm thực hiện

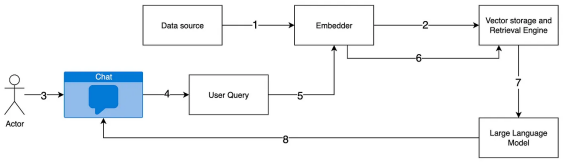

Figure 2: Kiến trúc cơ bản của một hệ thống sử dụng kiến trúc RAG

cầu xử lý. Sau đây, chúng tôi sẽ trình bày cụ thể chi tiết về các tùy biến được chọn cho ba thành phần này.

5.1 LLM: Gemini-1.5-Flash

Gemini-1.5-Flash (Google AI, 2025b) là một mô hình ngôn ngữ lớn được phát triển bởi Google, nhằm trong lớp các mô hình thuộc họ Gemini, một trong những hệ thống LLM được sử dụng phổ biến nhất hiện nay. Nhóm thực hiện chatbot này lựa chọn mô hình này nhằm tính hợp vào hệ thống vì nhiều nguyên nhân khác nhau, bao gồm nhưng không giới hạn ở:

- Mô hình được cung cấp miễn phí bởi Google. Do đặc thù đồ án này được thực hiện bởi sinh viên với chi phí hạn hẹp, việc sử dụng các mô hình không phát sinh chi phí giúp cân bằng giữa chi phí và hiệu quả thực hiện đồ án.

- Mô hình đi kèm với thông tin và hướng dẫn

sử dụng rõ ràng. Do được phát triển bởi một tổ chức có uy tín là Google, việc tiếp cận các thông tin chi tiết của mô hình là tương đối đơn giản, giúp hỗ trợ nhóm tùy chỉnh các thông số của mô hình, từ đó giúp tiết kiệm thời gian và tăng hiệu quả xây dựng đồ án.

5.2 Embedder: GeminiEmbedding

Để có được độ tương thích cao nhất với mô hình ngôn ngữ lớn đã chọn, chúng tôi chọn GeminiEmbedding (Google AI, 2025a) nhằm mã hóa dữ liệu đầu vào của hệ thống. Ngoài ra, bộ hóa này được cung cấp dưới dạng API, không những giúp tăng tốc độ xử lý mà còn giảm tải yêu cầu phần cứng của hệ thống, từ đó gián tiếp tối ưu hiệu suất tổng thể của hệ thống.

5.3 Vector Store & Retrieval: FAISS

FAISS *(Facebook AI Similarity Search)* (Johnson et al., 2017) là một thư viện mã nguồn mở được phát triển bởi Meta AI, được thiết kế để tìm kiếm hiệu quả trong không gian vector với kích thước lớn. FAISS hỗ trợ việc lưu trữ, tìm kiếm và so khớp các vector, đặc biệt hữu ích trong các bài toán như tìm kiếm tương tự, xếp hạng, hoặc truy vấn các hệ thống nhúng. Thư viện này được tối ưu hóa cao, hỗ trợ GPU để tăng tốc xử lý, giúp thực hiện các tác vụ tìm kiếm trong thời gian ngắn với dữ liệu hàng triệu điểm. Ngoài ra, FAISS hỗ trợ lưu trữ toàn bộ thông tin trong bộ nhớ mà không cần sử dụng đến cơ sở dữ liệu *(database)*. Do đó, chúng tôi nhận thấy việc sử dụng FAISS là hướng tiếp cận phù hợp nhất cho việc xây dựng chatbot với bối cảnh của bài toán.

6 Đánh giá hệ thống

Sau công đoạn lập trình hoàn thiện các tính năng, việc tiến hành đánh giá hiệu suất là một bước quan trọng nhằm đảm bảo chất lượng cho hệ thống vừa được xây dựng. Sau khi tham khảo các phương pháp đánh giá hiệu suất cho kiến trúc RAG (Yu et al., 2024), nhóm thực hiện đồ án nhận thấy việc sử dụng các bộ dữ liệu *(dataset)* sẵn có để đánh giá mô hình là không khả thi do các đặc trưng của bài toán. Do đó, nhóm tiến hành thiết kế các tình huống thường gặp do các thành viên tự sáng tạo để đánh giá hiệu suất mô hình xử lý.

Figure 3 và Figure 4 cho ví dụ về hai trường hợp tương ứng mà hệ thống xử lý tốt và xử lý không tốt prompt đầu vào. Trong đó, khi prompt đầu vào có độ tường minh cao, hệ thống dễ dàng đáp ứng được

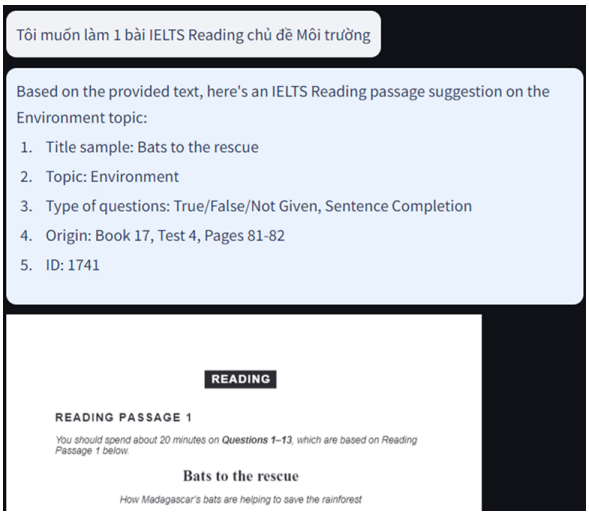

Figure 3: Một prompt cho kết quả trả về tốt từ hệ thống

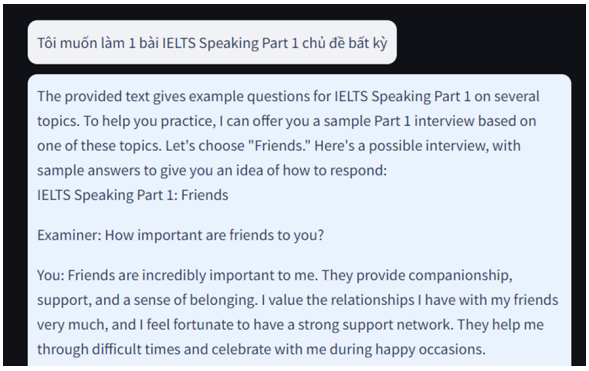

Figure 4: Một prompt cho kết quả trả về không tốt từ hệ thống

nhu cầu người dùng. Tuy nhiên, khi được yêu cầu mang tính tổng quát hóa cao hơn (như yêu cầu trả về một kết quả ngẫu nhiên), mô hình gặp lúng túng dẫn đến kết quả trả về chưa đạt kỳ vọng.

7 Phương hướng phát triển

7.1 Các vấn đề hiện hữu

Mô hình hiện tại chưa nắm bắt được đầy đủ các đặc trưng tổng thể của bài thi IELTS, dẫn đến việc hỗ trợ cho quá trình trả lời câu hỏi chưa đạt hiệu quả tối ưu. Điều này ảnh hưởng đến khả năng đưa ra những câu trả lời toàn diện và phù hợp với ngữ cảnh của bài thi. Hơn nữa, tính linh động của mô hình còn hạn chế, khi đôi lúc xảy ra hiện tượng *overfit* vào dữ liệu được cung cấp tại bước Retrieval, khiến mô hình khó thích ứng với các tình huống hoặc ngữ liệu mới. Không chỉ vậy, bản demo hiện tại vẫn còn chạy chậm, làm giảm trải nghiệm người dùng và chưa thể đáp ứng được yêu cầu về hiệu năng trong các ứng dụng thực tế. Sự thiếu linh hoạt và khả

năng cạnh tranh kém của mô hình cũng đặt ra thách thức lớn nếu muốn áp dụng rộng rãi trong các bài toán giáo dục hoặc các ứng dụng thương mại liên quan đến IELTS. Do đó, cần có những cải tiến về cả mặt kỹ thuật và hiệu suất để mô hình đáp ứng tốt hơn nhu cầu sử dụng.

7.2 Đề xuất giải pháp

Chúng tôi dự kiến phát triển thêm tính năng hỗ trợ hoàn thiện các phần còn lại của bài thi IELTS, bao gồm Listening và Writing, nhằm mang đến một giải pháp toàn diện hơn cho người học. Đồng thời, tính năng chấm bài cũng sẽ được bổ sung để giúp người dùng đánh giá kết quả chính xác và nhận được các gợi ý cải thiện phù hợp. Ngoài ra, nhóm thực hiện đồ án sẽ tập trung cải thiện chất lượng trả về của mô hình để đảm bảo các phản hồi không chỉ chính xác mà còn sát với thực tế của bài thi IELTS. Hệ thống cũng sẽ được tối ưu hóa về tốc độ khởi động và xử lý, nhằm mang lại trải nghiệm mượt mà và nhanh chóng hơn cho người dùng. Những tính năng này sẽ giúp mô hình trở thành một công cụ học tập hữu ích và đáng tin cậy trong tương lai.

8 Kết luận

Thông qua việc xây dựng chatbot hỗ trợ ôn luyện bài thi IELTS, phương pháp RAG đã chứng minh được tính hiệu quả khi được áp dụng trong bối cảnh của đồ án, nhờ khả năng kết hợp giữa truy xuất thông tin và tạo nội dung. Tuy nhiên, để có thể triển khai rộng rãi và đáp ứng yêu cầu của các ứng dụng thực tế, mô hình cần được cải thiện ở nhiều khía cạnh. Những nâng cấp này có thể bao gồm tối ưu hóa hiệu suất, nâng cao tính linh hoạt và khả năng xử lý các tình huống đa dạng hơn, nhằm đảm bảo mô hình không chỉ hoạt động tốt trong môi trường thử nghiệm mà còn đủ mạnh để đáp ứng nhu cầu thực tế.

References

Google AI. 2025a. Gemini api documentation: Embeddings. https://ai.google.dev/gemini-api/docs/embeddings. Accessed: 2025-01-27.

Google AI. 2025b. Gemini api documentation: Gemini models. https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash. Accessed: 2025-01-27.

IBM Research. 2023. Retrieval-augmented generation (rag): Enhancing ai with knowledge retrieval. https://research.ibm.com/blog/retrieval-augmented-generation-RAG. Accessed: 2025-01-27.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus.

Streamlit. 2025. Streamlit official website. https://streamlit.io/. Accessed: 2025-01-28.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey.