# TÓM TẮT

Nội dung báo cáo được tổ chức bao gồm phần mở đầu và 03 chương chính với nội dung cụ thể trong mỗi phần như sau:

Phần mở đầu sẽ trình bày bối cảnh của đề tài, lý do lựa chọn đề tài này, nói rõ về những khó khăn hiện nay trong việc học tập bộ môn Lịch sử ở Việt Nam và nguyên nhân sử dụng mô hình ngôn ngữ lớn (LLM) làm nền tảng cho giải pháp. Ngoài ra, trong phần này cũng sẽ nêu rõ các phạm vi của đề tài cũng như mục tiêu dự kiến để triển khai các nội dung chi tiết.

Chương 1 – Cơ sở lý thuyết sẽ trình bày các cơ sở lý thuyết về Học máy (ML), Học sâu (DL), Xử lý ngôn ngữ tự nhiên (NLP), mô hình Transformer và mô hình ngôn ngữ lớn (LLM). Ngoài ra, trong chương này cũng sẽ trình bày các công nghệ Trí tuệ nhân tạo (AI) và các công nghệ hỗ trợ khác được sử dụng trong hệ thống, từ đó làm rõ các cơ sở khoa học và những tiềm năng để ứng dụng trong hệ thống chatbot hỗ trợ học lịch sử Việt Nam.

Chương 2 – Phương pháp đề xuất xây dựng chatbot sẽ trình bày về phương pháp nghiên cứu, quá trình gồm các công đoạn như thu thập và xử lý dữ liệu, xây dựng các thành phần, thực hiện embedding dữ liệu, thiết kế và xây dựng thuật toán để truy xuất thông tin và thiết kế prompt; tất cả đều nhằm mục đích là tối ưu hiệu quả của chatbot.

Chương 3 – Triển khai hệ thống chatbot hỗ trợ học lịch sử Việt Nam sẽ tập trung vào mô tả về quá trình triển khai hệ thống chatbot từ đầu tới cuối, gồm có thiết kế giao diện người dùng (UI), xây dựng back-end, tích hợp các chức năng của hệ thống chatbot như trả lời câu hỏi, ra đề thi. Hiệu quả của hệ thống sẽ được đánh giá qua các thông số thu được sau khi thực nghiệm.

Phần cuối sẽ tổng kết lại những kết quả đã thu được, những đánh giá về hiệu quả của hệ thống và các đề xuất để cải thiện và phát triển hệ thống này trong tương lai.

# LỜI NÓI ĐẦU

Sự phát triển vượt bậc của công nghệ và khoa học máy tính, trí tuệ nhân tạo (AI) đã và đang đóng vai trò ngày càng quan trọng trong nhiều lĩnh vực của cuộc sống hiện đại. Đặc biệt, với sự xuất hiện và cải tiến mạnh mẽ của các mô hình ngôn ngữ lớn (LLM). Việc xây dựng các hệ thống tự động hóa, tương tác thông minh trở nên khả thi và hiệu quả hơn bao giờ hết. Các mô hình LLM đã không chỉ chứng minh được khả năng xử lý ngôn ngữ tự nhiên phức tạp mà còn giúp tạo ra những chatbot có thể giao tiếp linh hoạt, dễ dàng tùy biến để phù hợp với nhiều mục đích khác nhau. Trong bối cảnh này, xây dựng chatbot phục vụ mục tiêu tuyển sinh cho các trường đại học trở thành một giải pháp tiềm năng giúp các cơ sở giáo dục tối ưu hóa quy trình tư vấn, nâng cao trải nghiệm người dùng, và giảm thiểu chi phí nhân sự cho các nhiệm vụ lặp đi lặp lại.

Nhờ vào sự phát triển nhanh chóng của công nghệ, trí tuệ nhân tạo (AI) đã xuất hiện và thể hiện vai trò ngày càng chủ chốt trong rất nhiều lĩnh vực của cuộc sống

Tại Trường Đại học Đông Á, nhu cầu tuyển sinh và tư vấn hướng nghiệp luôn là một trong những ưu tiên hàng đầu. Hằng năm, hàng nghìn học sinh và phụ huynh có nhu cầu tìm hiểu thông tin về các chương trình học, quy chế tuyển sinh, cũng như các chính sách hỗ trợ từ phía nhà trường. Tuy nhiên, với khối lượng công việc lớn và sự biến động về thông tin tuyển sinh, các phương pháp tư vấn truyền thống dù vẫn còn đáp ứng được nhưng tốn kém lượng chi phí lớn và đôi khi chưa thể đáp ứng một cách nhanh chóng và hiệu quả. Do đó, việc áp dụng chatbot tuyển sinh dựa trên nền tảng LLM không chỉ là một bước tiến trong tự động hóa mà còn thể hiện tầm nhìn của nhà trường trong việc tiên phong ứng dụng công nghệ tiên tiến vào giáo dục. Chatbot tuyển sinh không chỉ có khả năng trả lời các câu hỏi từ phía học sinh và phụ huynh một cách chính xác, mà còn có thể cập nhật thông tin mới nhất, giúp người dùng có cái nhìn tổng quan và chính xác hơn về trường.

Mục tiêu của đồ án này là xây dựng một chatbot hỏi đáp tuyển sinh cho Trường Đại học Đông Á, dựa trên nền tảng LLM hiện đại, nhằm cung cấp một công cụ hỗ trợ hiệu quả cho quá trình tư vấn tuyển sinh. Với tính linh hoạt và khả năng xử lý ngôn ngữ tự nhiên của LLM, chatbot sẽ giúp nâng cao trải nghiệm của người dùng thông qua sự tương tác tự nhiên và liên tục. Đồng thời, chatbot cũng có khả năng mở rộng để tích hợp thêm các tính năng nâng caonhư điều hướng người dùng tới cán bộ giảng viên trong trường hợp câu hỏi mà chatbot không thể trả lời. Bài báo cáo này sẽ đi sâu vào việc nghiên cứu, phân tích, và triển khai chatbot tuyển sinh từ bước đầu đến khi hoàn thiện, góp phần khẳng định vai trò của trí tuệ nhân tạo trong lĩnh vực giáo dục.

# LỜI CAM ĐOAN

Nhóm chúng em xin cam đoan rằng tất cả các kết quả và thông tin trình bày trong đề tài này là thành quả công việc của nhóm dưới dự hướng dẫn của giảng viên TS. Nguyễn Năng Hùng Vân. Nhóm đã tuân thủ mọi nguyên tắc và quy định về nghiên cứu khoa học và đạo đức trong việc thực hiện đồ àn tốt nghiệp này.

Tất cả các nguồn tham khảo và các công trình nghiên cứu trước đây của người khác đã được nhóm trích dẫn và trình bày một cách rõ ràng và chi tiết trong danh mục tài liệu tham khảo của đồ án. Nhóm đã tuân thủ nguyên tắc không sao chép, không vi phạm bản quyền, và đã tuân thủ đúng quy định về việc ghi rõ nguồn gốc thông tin từ nguồn khác.

Các kết quả đánh giá và thống kê được trình bày trong khóa luận này đều là kết quả của những thực nghiệm mà nhóm đã tự thực hiện, và không có sự can thiệp từ bất kỳ nguồn nào khác. Nhóm cam đoan tính trung thực và chính xác trong việc trình bày kết quả và thông tin trong khóa luận này.

**Nhóm sinh viên thực hiện**

Ngô Văn Úc

Phạm Văn Quân

# LỜI CẢM ƠN

Trước tiên, nhóm chúng em xin gửi lời cảm ơn chân thành và sâu sắc đến TS. Nguyễn Năng Hùng Vân – giảng viên hướng dẫn trong suốt quá trình thực hiện đồ án tốt nghiệp, thầy đã tận tâm chỉ bảo và định hướng cho chúng em trong quá trình hoàn thiện đồ án. Những ý kiến đóng góp quý báu và sự động viên của thầy đã giúp chúng em vượt qua nhiều khó khăn và hoàn thành tốt nhiệm vụ được giao.

Chúng em cũng xin chân thành cảm ơn Ban giám hiệu nhà trường, quý thầy cô trong khoa Công nghệ thông tin và toàn thể giảng viên của Trường Đại học Đông Á, những người đã tạo điều kiện thuận lợi và cung cấp kiến thức nền tảng để chúng em có thể thực hiện đồ án tốt nghiệp cách tốt nhất.

Bên cạnh đó, nhóm cũng xin gửi lời cảm ơn đến các bạn thành viên trong nhóm, những người đã cùng trao đổi, thảo luận và đóng góp ý kiến hữu ích trong suốt quá trình nghiên cứu và triển khai đề tài.

Cuối cùng, chúng em xin dành lời cảm ơn đặc biệt đến gia đình, những người luôn bên cạnh, động viên và tạo mọi điều kiện tốt nhất để chúng em hoàn thành chặng đường học tập và thực hiện đồ án này.

Dù đã rất cố gắng, nhưng chắc chắn đồ án này vẫn còn những thiếu sót nhất định, nhóm chúng em rất mong nhận được sự đóng góp và chỉ dẫn từ quý thầy cô để có thể cải thiện và phát triển hơn trong tương lai.

MỤC LỤC

[TÓM TẮT I](#_Toc184647858)

[LỜI NÓI ĐẦU II](#_Toc184647859)

[LỜI CAM ĐOAN III](#_Toc184647860)

[LỜI CẢM ƠN IV](#_Toc184647861)

[MỤC LỤC V](#_Toc184647862)

[DANH SÁCH CÁC HÌNH ẢNH VIII](#_Toc184647863)

[DANH SÁCH CÁC KÝ HIỆU VIẾT TẮT X](#_Toc184647864)

[MỞ ĐẦU 1](#_Toc184647865)

[CHƯƠNG 1. CƠ SỞ LÝ THUYẾT 5](#_Toc184647866)

[1.1. Giới thiệu đề tài 5](#_Toc184647867)

[1.1.1. Tổng quan đề tài 5](#_Toc184647868)

[1.1.2. Nội dung nghiên cứu 5](#_Toc184647869)

[1.2. Tổng quan về Học máy 6](#_Toc184647870)

[1.3. Tổng quan về Học sâu 7](#_Toc184647871)

[1.4. Mạng Artificial Neural Network (ANN) 9](#_Toc184647872)

[1.4.1. Giới thiệu mạng ANN 9](#_Toc184647873)

[1.4.2. Các thành phần trong mạng ANN 9](#_Toc184647874)

[1.5. Tổng quan về Xử lý ngôn ngữ tự nhiên (NLP) 17](#_Toc184647875)

[1.5.1. Giới thiệu về NLP 17](#_Toc184647876)

[1.5.2. Học sâu trong NLP 17](#_Toc184647877)

[1.5.3. Các ứng dụng của NLP 18](#_Toc184647878)

[1.6. Mô hình Transformer 18](#_Toc184647879)

[1.6.1. Giới thiệu mô hình Transformer 18](#_Toc184647880)

[1.6.2. Tổng quan về kiến trúc mô hình Transformer 18](#_Toc184647881)

[1.6.3. Các thành phần kiến trúc mô hình Transformer 19](#_Toc184647882)

[1.6.4. Ứng dụng Transformer trong Xử lý ngôn ngữ tự nhiên 28](#_Toc184647883)

[1.7. Tổng quan về Mô hình ngôn ngữ lớn (LLM) 28](#_Toc184647884)

[1.7.1. Giới thiệu về LLM 28](#_Toc184647885)

[1.7.2. Quy trình huấn luyện một LLM 29](#_Toc184647886)

[1.7.3. Các kỹ thuật NLP trong LLM 30](#_Toc184647887)

[1.7.4. Ứng dụng LLM để xây dựng chatbot 37](#_Toc184647888)

[1.8. Các công cụ hỗ trợ 38](#_Toc184647889)

[1.8.1. Ngôn ngữ lập trình Python 38](#_Toc184647890)

[1.8.2. Các Framework sử dụng 39](#_Toc184647891)

[CHƯƠNG 2. PHƯƠNG PHÁP ĐỀ XUẤT XÂY DỰNG CHATBOT 41](#_Toc184647892)

[2.1. Mô tả và phân tích bài toán 41](#_Toc184647893)

[2.2. Phương pháp đề xuất 41](#_Toc184647894)

[2.3. Thiết kế hệ thống thông tin 44](#_Toc184647895)

[2.3.1. Chức năng hỏi đáp người dùng 44](#_Toc184647896)

[2.3.2. Biểu đồ use case chức năng hỏi đáp người dùng 45](#_Toc184647897)

[2.3.3. Biểu đồ tuần tự chức năng hỏi đáp người dùng 46](#_Toc184647898)

[2.3.4. Cơ sở dữ liệu người dùng 47](#_Toc184647899)

[2.4. Thu thập và tạo dữ liệu 47](#_Toc184647900)

[2.5. Tiền xử lý và phân đoạn dữ liệu 50](#_Toc184647901)

[2.6. Đánh giá các mô hình Embedding đề xuất 51](#_Toc184647902)

[2.7. Phương pháp đánh giá hiệu quả retrieval dữ liệu 54](#_Toc184647903)

[2.8. Các mô hình LLM được sử dụng 56](#_Toc184647904)

[2.9. Prompt engineering trong hệ thống chatbot 57](#_Toc184647905)

[2.10. Tác nhân (Agents) 59](#_Toc184647906)

[2.10.1. Tác nhân hỏi đáp (Q&A Agent) 59](#_Toc184647907)

[2.10.2. Tác nhân ngoài phạm vi (Out-of-Domain Agent) 60](#_Toc184647908)

[2.10.3. Quản lý tác nhân (Agent Management) 61](#_Toc184647909)

[2.11. Cá nhân hóa người dùng 61](#_Toc184647910)

[CHƯƠNG 3. TRIỂN KHAI HỆ THỐNG CHATBOT TUYỂN SINH 63](#_Toc184647911)

[3.1. Triển khai embedding dữ liệu và truy vấn người dùng 63](#_Toc184647912)

[3.2. Triển khai lưu trữ Vector embedding vào Vector database 63](#_Toc184647913)

[3.3. Retrieval cho truy vấn người dùng 64](#_Toc184647914)

[3.4. Triển khai agent cho hệ thống chatbot 65](#_Toc184647915)

[3.4.1. Xây dựng cấu trúc Prompt 66](#_Toc184647916)

[3.4.2. Triển khai prompt hỏi đáp 66](#_Toc184647917)

[3.4.3. Triển khai prompt điều hướng người dùng 68](#_Toc184647918)

[3.4.4. Trình quản lý agent và function calling 69](#_Toc184647919)

[3.5. Triển khai trình quản lý bộ nhớ chatbot 70](#_Toc184647920)

[3.6. Cơ sở dữ liệu người dùng 72](#_Toc184647921)

[3.7. Triển khai hệ thống chatbot hoàn chỉnh 72](#_Toc184647922)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 77](#_Toc184647923)

[CÁC TÀI LIỆU THAM KHẢO I](#_Toc184647924)

**DANH SÁCH CÁC BẢNG BIỂU**

[Bảng 1. Bảng so sánh sự khác nhau giữa Học máy và Học sâu [21]. 8](#_Toc184647925)

[Bảng 2. Bảng so sánh tổng quan trọng số trong mạng ANN [32]. 12](#_Toc184647926)

[Bảng 3. Bảng mô tả cơ sở dữ liệu người dùng trong hệ thống chatbot. 47](#_Toc184647927)

[Bảng 4. Minh hoạ một số mẫu dữ liệu được thu thập và tạo. 49](#_Toc184647928)

[Bảng 5. So sánh và đánh giá hiệu suất của các mô hình embedding đề xuất. 53](#_Toc184647929)

[Bảng 6. Bảng so sánh hiệu suất các phương pháp retrieval. 56](#_Toc184647930)

# DANH SÁCH CÁC HÌNH ẢNH

[Hình 1. Các thành phần cơ bản trong mạng ANN. 9](#_Toc184684590)

[Hình 2. Trực quan hóa ma trận trọng số cho một lớp của mạng ANN . 11](#_Toc184684591)

[Hình 3. Kiến trúc mô hình Transforme [39]. 19](#_Toc184684592)

[Hình 4. Minh hoạt các từ được áp dụng thuật toán Embedding. 20](#_Toc184684593)

[Hình 5. Vector tham số của mô hình Transformer [41]. 21](#_Toc184684594)

[Hình 6. Tính toán chỉ số Score cho các từ trong câu [41]. 22](#_Toc184684595)

[Hình 7. Ma trận tính toán self-attention [41]. 22](#_Toc184684596)

[Hình 8. Minh hoạt multi-head attention và ma trận trọng số Q, K, V cho mỗi từ [41]. 23](#_Toc184684597)

[Hình 9. Minh hoạt biểu diễn các attention cho mỗi từ [41]. 24](#_Toc184684598)

[Hình 10. Nhân ma trận trọng số để định hình lại kích thước của Z [41]. 24](#_Toc184684599)

[Hình 11. Minh họa tóm tắt quá trình tính toán multi-heads attention [41]. 25](#_Toc184684600)

[Hình 12. Minh hoạ một kết nối dư trong mạng nơ-ron [43]. 25](#_Toc184684601)

[Hình 13. Mạng nơ-ron thường (bên trái) và mạng nơ-ron có kết nối dư (bên phải) [43]. 26](#_Toc184684602)

[Hình 14. Mã hóa vị trí từ embedding [45]. 28](#_Toc184684603)

[Hình 15. Minh họa xử lý đa tác vụ của một LLM. 29](#_Toc184684604)

[Hình 16. Quá trình huấn luyện một mô hình LLM. 29](#_Toc184684605)

[Hình 17. Mô hình luồng hoạt động của kỹ thuật RAG. 34](#_Toc184684606)

[Hình 18. Minh hoạt hệ thống tác nhân (Agent) tự động được hỗ trợ bởi LLM [59]. 36](#_Toc184684607)

[Hình 19. Mô họa việc các câu lệnh kiến thức được tạo ra [63]. 37](#_Toc184684608)

[Hình 20. Tổng quan hệ thống chatbot sử dụng LLM. 38](#_Toc184684609)

[Hình 21. Tổng quan bài toán xây dựng hệ thống chatbot hỏi đáp tuyển sinh. 41](#_Toc184684610)

[Hình 22. Phương pháp đề xuất xây dựng chatbot hỏi đáp tuyển sinh sử dụng LLM. 42](#_Toc184684611)

[Hình 23. Biểu đồ chức năng hỏi đáp người dùng. 45](#_Toc184684612)

[Hình 24. Biểu đồ use case chức năng hỏi đáp tuyển sinh. 46](#_Toc184684613)

[Hình 25. Biểu đồ tuần tự chức năng hỏi đáp người dùng. 47](#_Toc184684614)

[Hình 26. Tạo dữ liệu tuyển sinh bằng cách thu thập và tạo kịch bản. 48](#_Toc184684615)

[Hình 27. Quy trình tiền xử lý và phân đoạn bộ dữ liệu. 50](#_Toc184684616)

[Hình 28. Tổng quan quy trình đánh giá hiệu suất các mô hình embedding. 51](#_Toc184684617)

[Hình 29. Quy trình đánh giá hiệu suất retrieval dữ liệu. 55](#_Toc184684618)

[Hình 30. Cấu trúc prompt trong hệ thống chatbot hỏi đáp tuyển sinh. 58](#_Toc184684619)

[Hình 31. Mô tả luồng hoạt động của Q&A agent. 59](#_Toc184684620)

[Hình 32. Mô tả luồng hoạt động của OOD agent. 60](#_Toc184684621)

[Hình 33. Quy trình thu thập thông tin và cá nhân hóa người dùng. 62](#_Toc184684622)

[Hình 34. Quá trình triển khai embedding bộ dữ liệu sử dụng mô hình embedding. 63](#_Toc184684623)

[Hình 35. Quá trình lưu trữ vector embedding vào vector database. 64](#_Toc184684624)

[Hình 36. Các bước thực hiện retrieval cho truy vấn người dùng. 65](#_Toc184684625)

[Hình 37. Cách một prompt được hoàn chỉnh và sử dụng trong hệ thống chatbot. 67](#_Toc184684626)

[Hình 38. Prompt điều hướng người dùng trong hệ thống chatbot. 69](#_Toc184684627)

[Hình 39. Minh họa chức năng của agent management. 70](#_Toc184684628)

[Hình 40. Mô hình kiến trúc hệ thống MemGPT [91]. 70](#_Toc184684629)

[Hình 41. Dữ liệu người dùng được thu thập và sử dụng cho mục đích cá nhân hóa. 72](#_Toc184684630)

[Hình 42. Minh họa bắt đầu trò chuyện sau skhi triển khai chatbot hoàn chỉnh. 73](#_Toc184684631)

[Hình 43. Minh họa kết quả triển khai hệ thống hoàn chỉnh. 73](#_Toc184684632)

[Hình 44. . Minh họa kết quả triển khai hệ thống hoàn chỉnh.s 74](#_Toc184684633)

[Hình 45. Minh họa kết quả triển khai hệ thống hoàn chỉnh. 74](#_Toc184684634)

[Hình 46. Minh họa kết quả triển khai hệ thống hoàn chỉnh. 75](#_Toc184684635)

[Hình 47. Minh họa kết quả triển khai hệ thống hoàn chỉnh. 76](#_Toc184684636)

# DANH SÁCH CÁC KÝ HIỆU VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Chữ viết tắt** | **Diễn giải** | **Ý nghĩa** |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| ANN | Artificial Neural Network | Mạng Nơ-ron nhân tạo |
| API | Application Programming Interface | Giao diện lập trình ứng dụng |
| DL | Deep learning | Học sâu |
| FFNN | Feed Forward Neural Network | Mạng nơ-ron truyền thẳng |
| LLM | Large Language Model | Mô hình ngôn ngữ lớn |
| LSTM | Long Short-Term Memory | Bộ nhớ dài-ngắn hạn |
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| ML | Machine learning | Học máy |
| SL | Self-supervised learning | Học tự giám sát |
| RAG | Retrieval-Augmented Generation | Tạo sinh tăng cường truy xuất |
| RNN | Recurrent Neural Network | Mạng nơ-ron hồi tiếp |

# MỞ ĐẦU

1. **Bối cảnh đề tài**

Trong những năm gần đây, sự phát triển của AI đã làm thay đổi sâu sắc đến nhiều khía cạnh của đời sống, bao gồm cả lĩnh vực giáo dục. Đặc biệt, với sự ra đời của các LLM làm tăng khả năng hiểu và xử lý ngôn ngữ tự nhiên của máy tính một cách đáng kể, mở ra những tiềm năng mới trong việc xây dựng các hệ thống tự động hóa, trong đó có hệ thống chatbot trong lĩnh vực chăm sóc khách hàng. Những công nghệ này có thể xử lý khối lượng dữ liệu khổng lồ, học hỏi từ những mẫu dữ liệu phức tạp, và tự động hóa các cuộc hội thoại một cách tự nhiên, mang lại trải nghiệm người dùng thân thiện và hiệu quả hơn.

Tại Việt Nam, việc tuyển sinh ở các trường đại học và cao đẳng hàng năm vẫn gặp nhiều khó khăn, đặc biệt là trong việc cung cấp thông tin đầy đủ, chính xác và kịp thời cho thí sinh và phụ huynh. Truyền thống tư vấn tuyển sinh thông qua các cuộc hội thảo, ngày hội tuyển sinh, hay hỗ trợ qua các kênh tư vấn trực tiếp đòi hỏi nhiều thời gian, chi phí nhân sự và khó mở rộng quy mô trong thời gian ngắn. Trong bối cảnh đó, việc áp dụng các công nghệ tự động hóa như chatbot tuyển sinh trở thành một giải pháp tiềm năng để tối ưu hóa quy trình tuyển sinh.

Trường Đại học Đông Á, với mong muốn tiếp cận nhanh chóng và hiệu quả hơn với các thí sinh tiềm năng, chúng tôi đã đề xuất ý tưởng xây dựng một chatbot hỏi đáp tuyển sinh dựa trên nền tảng LLM. Chatbot này sẽ đóng vai trò là một công cụ hỗ trợ tự động, có khả năng trả lời các câu hỏi từ phía học sinh và phụ huynh, từ thông tin về chương trình đào tạo, điều kiện tuyển sinh, đến chính sách học bổng và hỗ trợ sinh viên. Với khả năng xử lý ngôn ngữ tự nhiên của LLM, chatbot có thể mang lại trải nghiệm tư vấn tuyển sinh chân thực, giảm tải công việc cho đội ngũ tư vấn của nhà trường, đồng thời tạo ra một kênh giao tiếp hiện đại và hiệu quả hơn.

Vì vậy, bối cảnh của đề tài này không chỉ là nhu cầu ứng dụng công nghệ vào giáo dục mà còn là một bước tiến trong việc nâng cao trải nghiệm người dùng trong công tác tuyển sinh, giúp nhà trường hiện đại hóa quá trình tuyển sinh và đáp ứng tốt hơn nhu cầu thông tin của thí sinh.

1. **Lý do chọn đề tài**

Lý do chọn đề tài “Xây dựng chatbot tuyển sinh” xuất phát từ nhu cầu cấp thiết trong việc ứng dụng công nghệ hiện đại để cải thiện quy trình tuyển sinh. Hiện nay, với sự phát triển nhanh chóng của các mô hình ngôn ngữ lớn và trí tuệ nhân tạo, việc sử dụng AI để tối ưu hóa các dịch vụ chăm sóc và hỗ trợ người dùng trở nên khả thi và hiệu quả hơn bao giờ hết. Trong bối cảnh cuộc cách mạng công nghiệp 4.0, việc áp dụng AI không chỉ giúp tiết kiệm nguồn lực mà còn nâng cao trải nghiệm người dùng thông qua sự tương tác thông minh và cá nhân hóa.

Đặc biệt, trong lĩnh vực giáo dục, tuyển sinh là một quy trình quan trọng, ảnh hưởng trực tiếp đến hình ảnh và uy tín của nhà trường. Đối với Trường Đại học Đông Á, mỗi năm có hàng nghìn học sinh và phụ huynh có nhu cầu tìm hiểu thông tin về các chương trình đào tạo, quy chế tuyển sinh, học phí và học bổng. Tuy nhiên, với các phương pháp tư vấn truyền thống, việc đáp ứng kịp thời và chính xác nhu cầu thông tin của tất cả người dùng là một thách thức lớn, đặc biệt là trong thời gian cao điểm của mùa tuyển sinh. Do đó, việc xây dựng một chatbot tuyển sinh không chỉ là giải pháp tiết kiệm chi phí, thời gian mà còn đảm bảo tính liên tục, chính xác trong việc cung cấp thông tin tuyển sinh.

Lý do chọn đề tài còn xuất phát từ tiềm năng mà các mô hình LLM mang lại. Những mô hình này có khả năng xử lý ngôn ngữ tự nhiên một cách linh hoạt và sâu sắc, giúp chatbot không chỉ có thể trả lời các câu hỏi đơn giản mà còn cung cấp thông tin chi tiết, đa dạng và liên tục cập nhật. Việc ứng dụng LLM vào chatbot tuyển sinh còn là một bước đi tiên phong, mang lại lợi thế cạnh tranh và nâng cao hình ảnh hiện đại, năng động của nhà trường trong mắt các thí sinh và phụ huynh.

Tóm lại, đề tài này vừa đáp ứng nhu cầu thực tiễn trong quy trình tuyển sinh của nhà trường, vừa là cơ hội để khám phá, nghiên cứu và áp dụng một trong những công nghệ tiên tiến nhất hiện nay – mô hình ngôn ngữ lớn, từ đó góp phần nâng cao trải nghiệm người dùng và tối ưu hóa hoạt động tuyển sinh trong môi trường giáo dục.

1. **Đối tượng nghiên cứu**

Đối tượng nghiên cứu của đề tài này là tập trung nghiên cứu các LLM phục vụ cho việc xây dựng chatbot tuyển sinh, cụ thể là việc ứng dụng các mô hình như Llama, Gemini, Grok hoặc các mô hình ngôn ngữ có khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ khác vào lĩnh vực tư vấn giáo dục. Đối tượng nghiên cứu bao gồm các yếu tố kỹ thuật và thuật toán của mô hình ngôn ngữ, phương pháp nạp dữ liệu và tối ưu hóa mô hình, cũng như khả năng tương tác của mô hình với người dùng trong ngữ cảnh cung cấp thông tin tuyển sinh.

Ngoài ra, nghiên cứu cũng xem xét các phương pháp đánh giá chất lượng và hiệu quả hoạt động của chatbot, bao gồm độ chính xác của câu trả lời, khả năng hiểu ngữ cảnh, mức độ hài lòng của người dùng, và khả năng mở rộng của chatbot để đáp ứng nhu cầu người dùng thực tế.

1. **Phạm vi nghiên cứu**

Việc xây dựng một chatbot tuyển sinh là một nhiệm vụ đầy thách thức và phạm vi rộng rãi. Chính vì thế nhóm quyết định phát triển đề tài này trong phạm vi Trường Đại học Đông Á, cụ thể phạm vi nghiên cứu của đề tài bao gồm:

Phạm vi ứng dụng: Chatbot được xây dựng nhằm phục vụ công tác tuyển sinh cho Trường Đại học Đông Á. Do đó, chatbot sẽ tập trung vào các nội dung xoay quanh thông tin tuyển sinh của trường như chương trình đào tạo, quy chế tuyển sinh, học phí, chính sách học bổng, cũng như các thông tin liên quan đến cơ hội việc làm và các hoạt động hỗ trợ sinh viên. Chatbot cần đảm bảo trả lời chính xác các câu hỏi phổ biến, đồng thời có khả năng mở rộng để cập nhật thông tin mới theo thời gian.

Phạm vi công nghệ: Đề tài sẽ tập trung vào các mô hình ngôn ngữ lớn với khả năng xử lý ngôn ngữ tự nhiên cao, giúp chatbot có thể hiểu và phản hồi các câu hỏi từ người dùng một cách tự nhiên và chính xác. Các nền tảng và thư viện hỗ trợ chatbot như Chainlit, Milvus, các API dịch vụ ngôn ngữ có sẵn cũng sẽ được nghiên cứu và ứng dụng để tối ưu hóa hiệu suất của chatbot.

Phạm vi người dùng: Chatbot hướng đến nhóm người dùng bao gồm học sinh, sinh viên có nhu cầu tìm hiểu thông tin về Trường Đại học Đông Á, phụ huynh học sinh và các cá nhân quan tâm đến việc nhập học tại trường. Đối tượng người dùng có thể truy cập chatbot qua các kênh truyền thông chính thức của nhà trường, giúp đảm bảo tính thuận tiện và dễ dàng cho người dùng khi tìm kiếm thông tin tuyển sinh. Với các đối tượng và phạm vi nghiên cứu rõ ràng, đề tài không chỉ tập trung vào việc phát triển chatbot tuyển sinh mà còn mở ra cơ hội nâng cao trải nghiệm tư vấn giáo dục thông qua ứng dụng AI.

1. **Nội dung nghiên cứu**

Đề tài này tập trung vào những nội dung nghiên cứu bao gồm: Đầu tiên, đề tài sẽ tập trung tìm hiểu và đánh giá các LLM, các mô hình embedding và các biến thể phổ biến khác. Mục tiêu là phân tích các ưu nhược điểm và lựa chọn mô hình phù hợp nhất để áp dụng vào chatbot tuyển sinh, đồng thời nghiên cứu kỹ thuật tinh chỉnh nhằm đáp ứng các yêu cầu đặc thù trong bối cảnh của trường. Tiếp theo, đề tài sẽ triển khai hệ thống chatbot thông qua thiết kế kiến trúc tổng thể, bao gồm mô hình ngôn ngữ và giao diện người dùng. Một pipeline dữ liệu sẽ được xây dựng từ quá trình thu thập, tiền xử lý dữ liệu và nạp dữ liệu vào các LLM, sử dụng các thư viện và tích hợp qua API dịch vụ ngôn ngữ để tối ưu hóa hiệu suất.

Để đảm bảo chất lượng của chatbot, nhóm sẽ thiết lập các tiêu chí đánh giá như độ chính xác, thời gian phản hồi, khả năng hiểu ngữ cảnh và mức độ hài lòng của người dùng. Sau đó, tiến hành thử nghiệm trên các kịch bản tuyển sinh thực tế nhằm đánh giá hiệu quả của chatbot và từ đó thực hiện các bước tối ưu hóa để đảm bảo chatbot hoạt động ổn định, chính xác. Cuối cùng, chatbot sẽ được triển khai trên các kênh truyền thông của nhà trường, như website tuyển sinh và các nền tảng mạng xã hội, nhằm dễ dàng tiếp cận với học sinh và phụ huynh. Sau khi triển khai, đề tài cũng đánh giá khả năng mở rộng các tính năng nâng cao cho chatbot trong tương lai, ví dụ như tự động cập nhật dữ liệu tuyển sinh, hỗ trợ hướng nghiệp hoặc tích hợp các dịch vụ sinh viên khác. Nội dung nghiên cứu này giúp đảm bảo rằng chatbot tuyển sinh được xây dựng một cách bài bản, đáp ứng tốt nhu cầu người dùng và góp phần nâng cao hiệu quả quy trình tuyển sinh của trường.

1. **Cấu trúc bài báo cáo**

Báo cáo này được chia thành ba chương, tập trung vào việc xây dựng một hệ thống chatbot hỗ trợ tuyển sinh dựa trên mô hình ngôn ngữ lớn. Cấu trúc báo cáo được thiết kế để dẫn dắt người đọc từ cơ sở lý thuyết đến phương pháp triển khai và kết quả thực nghiệm, nhằm đảm bảo sự mạch lạc và tính ứng dụng của nghiên cứu.

Phần mở đầu giới thiệu bối cảnh và động lực thực hiện đề tài, nêu bật những thách thức trong việc tự động hóa tư vấn tuyển sinh và lý do lựa chọn mô hình LLM làm nền tảng. Phần này làm rõ mục tiêu nghiên cứu và kỳ vọng đóng góp của đề tài, qua đó tạo tiền đề để tiếp cận nội dung chi tiết.

Chương 1 cung cấp cơ sở lý thuyết, tập trung vào các công nghệ và phương pháp liên quan như học sâu, xử lý ngôn ngữ tự nhiên, mô hình Transformer và các đặc điểm nổi bật của mô hình ngôn ngữ lớn, các công cụ và kỹ thuật được sử dụng. Việc trình bày lý thuyết không chỉ giúp làm rõ cơ sở khoa học của nghiên cứu mà còn minh họa sự phù hợp và tiềm năng của các công nghệ này trong bối cảnh ứng dụng chatbot tuyển sinh.

Chương 2 tập trung vào phương pháp nghiên cứu và triển khai hệ thống. Nội dung chương này bao gồm các bước từ thu thập, xử lý dữ liệu đến thiết kế và tích hợp các thành phần như embedding dữ liệu, thuật toán truy xuất thông tin, xây dựng prompt. Phương pháp được xây dựng nhằm tối ưu hóa hiệu quả chatbot trong việc cung cấp thông tin và tương tác với người dùng, đồng thời đảm bảo tính linh hoạt để mở rộng cho các bài toán tương tự.

Chương 3 trình bày quá trình triển khai hệ thống chatbot hỏi đáp tuyển sinh. Phần này mô tả việc chuyển đổi lý thuyết và phương pháp đề xuất thành sản phẩm thực tế, bao gồm thiết kế giao diện người dùng, xây dựng backend xử lý dữ liệu và tích hợp các chức năng cốt lõi như hỏi đáp và điều hướng người dùng. Hiệu quả của chatbot được đánh giá thông qua các thử nghiệm thực tế, minh chứng cho khả năng áp dụng của giải pháp trong bối cảnh thực tiễn.

Cuối cùng, phần kết luận và hướng phát triển tóm tắt những kết quả đạt được, đánh giá hiệu quả hệ thống và đề xuất những cải tiến nhằm nâng cao chất lượng và khả năng mở rộng của chatbot trong tương lai. Báo cáo hướng đến mục tiêu không chỉ giải quyết bài toán tự động hóa tư vấn tuyển sinh mà còn góp phần ứng dụng các công nghệ tiên tiến trong việc nâng cao hiệu quả của các hệ thống hỗ trợ thông minh.

# CƠ SỞ LÝ THUYẾT

## Giới thiệu đề tài

### Tổng quan đề tài

Trí tuệ nhân tạo (Aritificial Intenlligence - AI) ngày càng khẳng định vai trò quan trọng trong nhiều lĩnh vực, từ kinh doanh, chăm sóc sức khỏe đến giáo dục [1], [2]. Đặc biệt, các mô hình ngôn ngữ lớn đã mở ra nhiều tiềm năng mới trong việc xử lý và hiểu ngôn ngữ tự nhiên, hỗ trợ tự động hóa các tác vụ phức tạp, đặc biệt nhất là các hệ thống hội thoại tự động (chatbot) [3][4].

Trong bối cảnh đó, việc ứng dụng AI vào quy trình tuyển sinh của các trường đại học giúp tối ưu hóa quy trình cung cấp thông tin, nâng cao trải nghiệm người dùng và giảm tải công việc cho nhân viên tư vấn. Đề tài “Xây dựng chatbot tuyển sinh” dựa trên mô hình ngôn ngữ lớn ra đời nhằm mục đích thiết kế một chatbot có khả năng trả lời nhanh chóng các câu hỏi tuyển sinh phổ biến của học sinh và phụ huynh.

Chatbot tuyển sinh không chỉ giúp tự động hóa các câu hỏi thường gặp về chương trình đào tạo, học phí, chính sách học bổng mà còn có khả năng cập nhật và cung cấp thông tin chính xác, liên tục [5]. Điều này mang lại lợi ích lớn trong bối cảnh quy trình tuyển sinh ngày càng phức tạp và nhu cầu thông tin của người dùng ngày càng tăng cao. Việc ứng dụng các mô hình ngôn ngữ lớn vào chatbot sẽ giúp hệ thống này trở nên thông minh, có khả năng hiểu và phản hồi đa dạng câu hỏi từ người dùng, từ đó đáp ứng tốt yêu cầu thực tế của quá trình tuyển sinh [6].

### Nội dung nghiên cứu

Nội dung nghiên cứu của đề tài bao gồm việc tìm hiểu và đánh giá các mô hình ngôn ngữ lớn, các mô hình embedding phù hợp với chatbot tuyển sinh, phân tích ưu nhược điểm của từng mô hình để lựa chọn mô hình tối ưu. Đề tài cũng sẽ tập trung vào việc xây dựng kiến trúc hệ thống chatbot, bao gồm mô hình ngôn ngữ và giao diện người dùng, cùng với việc thiết lập pipeline dữ liệu phục vụ cho quá trình nạp dữ liệu và thử nghiệm chatbot.

Ngoài ra, nội dung nghiên cứu còn bao gồm việc đánh giá hiệu suất của chatbot qua các tiêu chí như độ chính xác, thời gian phản hồi, và mức độ hài lòng của người dùng. Đề tài sẽ thử nghiệm chatbot trên các kịch bản tuyển sinh thực tế, sau đó tối ưu hóa hệ thống để đáp ứng tốt nhất nhu cầu của người dùng.

Cuối cùng, nghiên cứu sẽ triển khai chatbot lên các nền tảng chính thức của nhà trường và xây dựng tài liệu hướng dẫn, nhằm đảm bảo chatbot được vận hành hiệu quả và dễ dàng mở rộng tính năng trong tương lai.

## Tổng quan về Học máy

#### Giới thiệu về Học máy

Học máy (Machine Learning - ML) là lĩnh vực con của AI, một nhánh đang phát triển của các thuật toán tính toán được thiết kế để mô phỏng trí thông minh của con người bằng cách học hỏi từ môi trường xung quanh [7]. Thay vì được lập trình với các quy tắc cứng nhắc, các mô hình học máy được huấn luyện để nhận diện mẫu và đưa ra dự đoán dựa trên dữ liệu đầu vào.

Học máy có thể được chia thành ba loại chính: học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning). Ngày nay, học máy ngày càng trở nên phổ biến trong các ứng dụng thực tiễn, từ nhận diện hình ảnh, phân tích văn bản, đến dự đoán hành vi người dùng [2].

#### Học máy có giám sát

Học máy có giám sát (SL) là một mô hình học máy, trong đó các đối tượng đầu vào (là một vector đầu vào hoặc các biến dự báo) và một giá trị đầu ra mong muốn (còn được gọi là tín hiệu giám sát được gắn nhãn bởi con người) đào tạo một mô hình [8]. Dữ liệu đào tạo được xử lý, xây dựng một hàm ánh xạ dữ liệu mới thành các giá trị đầu ra mong đợi [9]. Mục tiêu của thuật toán học có giám sát là tìm ra một hàm số

với là không gian các đặc trưng (features) và là không gian các nhãn (labels). Thuật toán học có giám sử dụng dữ liệu đầu vào , , …, , trong đó mỗi cặp chứa một đặc trưng (input) và nhãn tương ứng (output) [10]. Để đo lường mức độ học tập của một hàm với dữ liệu đào tạo, một hàm mất mát

được định nghĩa. Đối với cặp , sự mất mát của việc dự đoán giá trị là [10].

#### Học máy không giám sát

Học không giám sát là một nhánh của học máy mà trong đó các thuật toán tự khám phá cấu trúc hoặc mẫu từ dữ liệu không có nhãn. Theo Ghahramani (2003) trong bài báo "Unsupervised Learning," học không giám sát chủ yếu tập trung vào việc nhận dạng và phân nhóm các mẫu trong dữ liệu mà không cần biết trước bất kỳ nhãn hoặc hướng dẫn nào [11].

Mục đích của học không giám sát là tìm cách tối ưu hóa một hàm mục tiêu sao cho mô hình có thể biểu diễn cấu trúc ẩn trong tập dữ liệu mà không có nhãn. Quá trình này thường liên quan đến việc tối ưu một hàm mất mát để khám phá các đặc trưng quan trọng hoặc cấu trúc của dữ liệu.

Giả sử ta có tập dữ liệu , trong đó mỗi là một điểm dữ liệu trong không gian . Ta xây dựng một hàm mục tiêu

(với ) để biến đổi dữ liệu sao cho có thể biểu diễn các đặc trưng quan trọng của dữ liệu. Hàm mục tiêu này có thể được chọn sao cho phản ánh được các mối quan hệ, mẫu hoặc cấu trúc ẩn trong dữ liệu. Để tìm các biểu diễn tối ưu, ta tối thiểu hóa một hàm mất mát tổng quát [12]:

Về cơ bản học máy có giám sát và học máy không giám sát là hai phương pháp học máy phổ biến nhất được sử dụng trong nhiều tác vụ từ đơn giản đến phức tạp và đã chứng minh được hiệu quả của chúng [13]. Ngoài hai phương pháp học máy trên còn nhiều phương pháp tiên tiến khác nhằm giải quyết từng nhiệm vụ cụ thể. Một vài thuật toán có thể kể đến như là học bán giám sát, học tăng cường v.v. Tuy nhiên phương pháp học máy tự giám sát (self-supervised learning) và học đa nhiệm (multi-task learning) là hai phương pháp quan trọng được sử dụng để huấn luyện mô hình ngôn ngữ lớn liên quan tới chatbot [14].

#### Ứng dụng của học máy

Học máy đóng một vai trò thiết yếu trong sự phát triển và ứng dụng của trí tuệ nhân tạo, mang lại nhiều lợi ích đáng kể trong các lĩnh vực khác nhau. ML giúp tự động hóa và tối ưu hóa các quy trình, giảm thiểu sự can thiệp của con người trong các tác vụ phức tạp [12] và cung cấp khả năng phân tích dữ liệu lớn một cách hiệu quả. Với sự gia tăng khối lượng dữ liệu trong kỷ nguyên số, các phương pháp học máy cho phép doanh nghiệp khai thác dữ liệu để tìm ra thông tin quý giá, từ đó đưa ra quyết định chính xác hơn [15]. Một số ứng dụng của học máy có thể kể đến như [16], [17], [18]:

* Phát triển Trí tuệ nhân tạo và robot
* Nhận diện hình ảnh, khuôn mặt
* Tự động nhận diện giọng nói
* Dịch tự động
* Đề xuất sản phẩm
* Trợ lý ảo cá nhân…

## Tổng quan về Học sâu

#### Khái niệm về Học sâu

Học sâu (Deep Learning - DL) là một nhánh con của học máy, sử dụng các mạng nơ-ron nhân tạo với nhiều lớp để học các đặc trưng từ dữ liệu lớn và phức tạp [19]. Mô hình học sâu có khả năng tự động rút trích đặc trưng từ dữ liệu mà không cần phải thiết kế thủ công, giúp chúng trở thành công cụ mạnh mẽ trong nhiều ứng dụng, từ nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên, đến xe tự lái [12].

#### Tầm quan trọng và ứng dụng của Học sâu

Nhờ vào khả năng học và phát triển từ dữ liệu phức tạp mà không cần sự can thiệp thủ công, học sâu đã trở thành một công nghệ quan trọng trong thế giới trí tuệ nhân tạo. Các mô hình học sâu, đặc biệt là mạng nơ-ron nhân tạo với nhiều lớp, có thể thực hiện nhiều tác vụ như nhận diện và phân tích dữ liệu không cấu trúc như hình ảnh, âm thanh và văn bản, giúp tăng cường khả năng dự đoán và phân loại [2], [5], [12], [15], [20].

#### Sự khác biệt giữa Học máy và Học sâu

Học máy và học sâu là hai nhánh chính trong trí tuệ nhân tạo, nhưng chúng có những đặc điểm khác biệt quan trọng. ML thường sử dụng các mô hình đơn giản, ngược lại DL sử dụng các mạng nơ-ron nhân tạo với nhiều lớp (mạng nơ-ron sâu) để học từ dữ liệu phức tạp. ML có thể yêu cầu người dùng phải thiết kế các đặc trưng (features) thủ công từ dữ liệu đầu vào. Trong khi đó, DL có thể tự động trích xuất đặc trưng từ dữ liệu, điều này đặc biệt hiệu quả trong việc xử lý dữ liệu không cấu trúc, như hình ảnh, âm thanh, và văn bản, nhờ vào khả năng học các đặc trưng phức tạp và ML truyền thống ít hiệu quả hơn trong các bài toán này. DL yêu cầu phần cứng mạnh mẽ và thời gian huấn luyện dài hơn so với ML truyền thống [1]. Dưới đây là bảng so sánh chi tiết về sự khác nhau về Ml và DL [21]:

Bảng 1. Bảng so sánh sự khác nhau giữa Học máy và Học sâu [21].

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Học máy** | **Học sâu** |
| Mô hình | ML là một phương pháp trí tuệ nhân tạo (AI). Không phải tất cả ML đều là học sâu. | Học sâu là phương pháp ML tiên tiến. Tất cả học sâu là ML. |
| Trường hợp sử dụng phù hợp nhất | ML là tùy chọn tốt nhất cho các tác vụ được xác định rõ ràng với dữ liệu có cấu trúc và được dán nhãn. | Học sâu là tùy chọn tốt nhất để thực hiện các tác vụ phức tạp đòi hỏi máy móc phải hiểu được dữ liệu không có cấu trúc. |
| Cách tiếp cận giải quyết sự cố | ML giải quyết các sự cố thông qua thống kê và toán học. | Học sâu kết hợp thống kê và toán học với kiến trúc mạng nơ-ron. |
| Đào tạo | Phải tự chọn và trích xuất các đối tượng từ dữ liệu thô và gán trọng số để huấn luyện một mô hình ML. | Các mô hình học sâu có thể tự học bằng cách sử dụng phản hồi từ các lỗi đã xác định. |
| Tài nguyên tính toán | ML ít phức tạp hơn và có khối lượng dữ liệu thấp hơn. | Học sâu phức tạp hơn với khối lượng dữ liệu rất cao. |

## Mạng Artificial Neural Network (ANN)

### Giới thiệu mạng ANN

Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) là một loại mô hình tính toán lấy cảm hứng từ cấu trúc và chức năng của bộ não con người. Mạng ANN bao gồm các đơn vị gọi là "nơ-ron" được liên kết với nhau theo một mạng lưới đa tầng, mạng này hoạt động bằng cách truyền tín hiệu qua các liên kết để thực hiện các nhiệm vụ tính toán [22]. Mạng ANN là nền tảng của các phương pháp học sâu hiện đại, giúp giải quyết nhiều bài toán phức tạp như phân loại ảnh, xử lý ngôn ngữ tự nhiên [23], và dự đoán chuỗi thời gian v.v., là một trong những cột mốc quan trọng trong lĩnh vực trí tuệ nhân tạo và học máy. Trong hình 1 là ví dụ về một mạng ANN với các thành phần chính của mạng.

### Các thành phần trong mạng ANN

Hình 1 là một ví dụ tổng quát về mạng ANN đơn giản. Cấu trúc của mạng ANN bao gồm những thành phần sau [24]:

A diagram of a complex structure

Description automatically generated

Hình 1. Các thành phần cơ bản trong mạng ANN.

* Lớp đầu vào có nơ-ron, mỗi nơ-ron tương ứng với một đặc trưng trong dữ liệu đầu vào.
* Tầng ẩn có lớp, mỗi lớp có số lượng nơ-ron khác nhau, mỗi nơ-ron trong lớp ẩn có một trọng số () và hàm kích hoạt kết nối với các nơ-ron ở lớp trước đó . Mỗi nơ-ron trong lớp ẩn được cộng thêm giá trị bias (sai số) .
* Lớp đầu ra có nơ-ron, với mỗi nơ-ron () đại diện cho một đầu ra.
* Các nơ-ron trong mạng liên kết với nhau, chúng là trọng số biểu thị sự quan trọng của một đặc trưng cho liên kết đó và thường được biểu diễn dưới dạng ma trận.

#### Lớp đầu vào

Giả sử ta có tập dữ liệu có mẫu dữ liệu. Với mỗi () tương ứng . Lúc này lớp đầu vào (Input layer) trong mạng ANN là lớp đầu tiên nhận dữ liệu từ và chuyển nó vào mạng để xử lý tiếp. Lớp này thường dựa trên số vector đặc trưng (feature) của dữ liệu, nơi mỗi đặc trưng của dữ liệu đầu vào được biểu diễn dưới dạng một phần tử trong vector . Mỗi nơ-ron đại diện một nút tròn nhỏ trong sơ đồ hình 1, và tương ứng với một feature của một mẫu trong tập dữ liệu [25]. Trong lớp đầu vào, các nơ-ron không thực hiện bất kỳ tính toán nào.

#### Kết nối

Các kết nối giữa các nơ-ron trong mạng ANN rất quan trọng đối với quá trình học. Mỗi nơ-ron trong một lớp được kết nối với mọi nơ-ron trong lớp liền kề. Các kết nối này được biểu diễn bằng giá trị trọng số, điều xác định tầm quan trọng của kết nối đó. Các giá trị trong trọng số là các tham số có thể đào tạo được mà mạng nơ-ron học được bằng cách lặp lại trên dữ liệu đào tạo. Việc tối ưu hóa các giá trị trọng số này rất quan trọng đối với hiệu suất chung của mạng nơ-ron và là một khía cạnh chính của quá trình học [26].

#### Độ lệch bias

Tham số bias trong mạng ANN giúp điều chỉnh đầu ra của nơ-ron, nó cho phép các mô hình học được các quan hệ phi tuyến tính và giúp quá trình tối ưu hóa mô hình nhanh hơn [12]. Trong một mạng ANN, bias sẽ được tính riêng biệt cho mỗi lớp. Mỗi lớp sẽ có một vector bias riêng và kích thước của nó phụ thuộc vào số lượng nơ-ron trong lớp đó.

Với mạng ANN có lớp ẩn, lớp thứ có nơ-ron thì bias của tầng thứ có dạng [12]:

#### Hàm kích hoạt

Hàm kích hoạt đóng vai trò quan trọng trong việc đào tạo và hiệu suất của Mạng nơ-ron nhân tạo. Chúng cung cấp các thuộc tính phi tuyến tính cần thiết cho bất kỳ mạng nơ-ron nhân tạo nào [27]. Hàm kích hoạt của một nút trong [mạng nơ-ron nhân tạo](https://en.wikipedia.org/wiki/Artificial_neural_network) tính toán đầu ra của nơ-ron dựa trên các đầu vào riêng lẻ và trọng số của chúng [28].

Một hàm kích hoạt bão hòa nếu:

và hàm kích hoạt không bão hòa nếu:

Có nhiều hàm kích hoạt khác nhau, các hàm kích hoạt không bão hòa thì có thể tốt hơn các hàm kích hoạt bão hòa, vì chúng ít có khả năng gặp phải vấn đề về biến mất độ dốc (gradient vanishing) [29].

Trường hợp tổng quát, trong một layer bất kỳ có số lượng nơ-ron trong lớp là , hàm kích hoạt có thể được biểu diễn bằng cách áp dụng một hàm số lên tổng có trọng số của các đầu vào (với trọng số ​) và cộng thêm một giá trị bias ​. Cụ thể, đối với mỗi nơ-ron trong lớp , đầu ra có thể được tính như sau [12]:

Trong đó:

* là trọng số kết nối giữa đầu vào ​ và nơ-ron trong lớp ,
* ​ là đầu vào của lớp (có thể là đầu ra của lớp trước hoặc dữ liệu gốc đối với lớp đầu tiên),
* là bias của nơ-ron trong lớp .
* là hàm kích hoạt.

Ở lớp đầu tiên, các đầu vào ​ sẽ là dữ liệu huấn luyện. Với các lớp kế tiếp, đầu vào của mỗi lớp sẽ là đầu ra từ lớp trước, được tính toán theo công thức trên.

#### Trọng số

Như đã đề cập ở trên, trọng số là kết nối giữa các nơ-ron mang giá trị. Giá trị càng cao có nghĩa là trọng số càng lớn và chúng sẽ càng coi trọng nơ-ron ở phía đầu vào của trọng số. Giả sử chúng ta đang xem xét một layer bất kỳ trong tầng ẩn và có nơ-ron, layer ẩn có nơ-ron, ma trận trọng số tại lớp là được minh họa bằng một hình ảnh (hình 2) [30].

A diagram of a number of equations

Description automatically generated with medium confidence

Hình 2. Trực quan hóa ma trận trọng số cho một lớp của mạng ANN .

*Trọng số của một kết* nối được biểu thị bằng một giá trị, ví dụ trọng số kết nối của nơ-ron thứ của lớp với nơ-ron thứ của lớp là

*Vector trọng số* là toàn bộ kết nối của một nơ-ron, ví dụ vector trọng số của nơ-ron thứ của lớp là một vector .

Công thức toán biểu diễn cho ***ma trận trọng số*** tại tầng ẩn có nơ-ron với tầng ẩn có nơ-ron được tổng quát như sau [31]:

Bảng 2 so sánh sự khác nhau giữa trọng số kết nối, vector trọng số và ma trận trọng số:

Bảng 2. Bảng so sánh tổng quan trọng số trong mạng ANN [32].

|  |  |  |  |
| --- | --- | --- | --- |
| **Loại trọng số** | **Định nghĩa** | **Đặc điểm** | **Vai trò chính** |
| Trọng số kết nối | Trọng số giữa hai nơ-ron cụ thể | Là một giá trị | Điều chỉnh cường độ tín hiệu giữa hai nơ-ron |
| Trọng số nơ-ron | Toàn bộ trọng số kết nối của một nơ-ron | Là vector | Kết hợp đầu vào để tính toán giá trị nơ-ron |
| Ma trận trọng số | Toàn bộ trọng số của một nơ-ron | Là ma trận | Lan truyền tín hiệu giữa các lớp (forward propagation) |

Trọng số đóng vai "học" của mạng nơ-ron, những tham số này được điều chỉnh để tối ưu hóa đầu ra của mạng trong suốt quá trình học. Thông qua lan truyền ngược (backpropagation) và các thuật toán tối ưu hóa như gradient descent, mạng sẽ cập nhật các trọng số này để giảm thiểu sai số dự đoán trên tập huấn luyện [33].

#### Các lớp ẩn

Các lớp ẩn (hidden layers) trong mạng ANN đóng một vai trò quan trọng trong việc xử lý và trích xuất các đặc trưng của dữ liệu đầu vào. Sau khi dữ liệu đầu vào được nhận từ lớp đầu vào, nó sẽ được truyền qua các lớp ẩn, nơi các nơ-ron sẽ áp dụng các phép toán phi tuyến để chuyển đổi dữ liệu [34].

Giả sử ta có một mạng ANN có lớp ẩn, lớp thứ có nơ-ron ( là chỉ số của lớp, ). Ta định nghĩa:

* Ma trận trọng số giữa lớp và lớp là có kích thước .
* Mỗi phần tử của ma trận trọng số biểu diễn trọng số từ nơ-ron thứ của lớp đến nơ-ron thứ của lớp .
* Quy ước lớp ẩn là lớp đầu vào.

Quá trình tính toán trọng số trong lớp ẩn như sau:

* Tính toán trọng số giữa lớp đầu vào và lớp ẩn đầu tiên: Mỗi nơ-ron trong lớp ẩn đầu tiên nhận đầu vào từ lớp đầu vào (có nơ-ron). Đầu vào cho nơ-ron ( với là số lượng nơ-ron trong lớp ) trong lớp ẩn đầu tiên có thể viết dưới dạng [31]:

Trong đó:

* + ​: Giá trị đầu vào thứ từ lớp đầu vào.
  + ​: Trọng số kết nối giữa đầu vào thứ và nơ-ron thứ .
  + : Bias của nơ-ron thứ tại lớp ẩn thứ .
  + ​: Tổng có trọng số của nơ-ron .

Và sau đó qua hàm kích hoạt:

* Tính toán kết nối giữa các lớp ẩn: Các lớp ẩn tiếp theo tính toán theo cách tương tự. Đầu vào cho nơ-ron trong lớp ẩn thứ (với ) được tính bằng tổng có trọng số từ các nơ-ron trong lớp ẩn trước đó [31]:

Trong đó:

* + ​ là số lượng nơ-ron trong lớp .
  + ​ là đầu ra của nơ-ron trong lớp .
  + ​ là trọng số giữa nơ-ron của lớp và nơ-ron trong lớp .

Và sau đó qua hàm kích hoạt:

#### Lớp đầu ra

Lớp đầu ra (output layer) là lớp cuối cùng trong mạng nơ-ron nhân tạo (ANN), chịu trách nhiệm trả về kết quả của mô hình. Số lượng nơ-ron trong lớp đầu ra phụ thuộc vào bài toán cụ thể. Đối với các bài toán hồi quy, số lượng nơ-ron đầu ra là 1 để biểu diễn giá trị liên tục. Trong khi bài toán phân loại thường có 2 nơ-ron trở lên, mỗi đầu ra tương ứng với một nhãn của tập dữ liệu [35].

Đầu vào của một nơ-ron () trong lớp đầu ra là tổng có trọng số của các giá trị đầu ra từ lớp ẩn cuối cùng:

Trong đó:

* ​: Số nơ-ron trong lớp ẩn cuối cùng.
* : Trọng số giữa nơ-ron của lớp ẩn cuối và nơ-ron trong lớp đầu ra.
* : Đầu ra từ nơ-ron của lớp ẩn cuối cùng.
* ​: Bias của nơ-ron trong lớp đầu ra.

Qua một hàm kích hoạt nữa, ta được kết quả đầu ra cho nơ-ron của lớp đầu ra như sau:

#### Hàm mất mát

Hàm mất mát (loss function) đo lường mức độ sai lệch giữa giá trị dự đoán của mạng () và giá trị thực tế (). Mục tiêu của mạng là tối thiểu hóa hàm mất mát thông qua quá trình huấn luyện bằng cách điều chỉnh các trọng số và bias trong mạng.

Nhắc lại kiến thức:

* + Tập dữ liệu huấn luyện gồm n mẫu, với .
  + : Nhãn thực tế của mẫu .
  + : Giá trị dự đoán của mạng ANN với đầu vào .

Hàm mất mát cơ bản của mô hình có dạng:

Trong đó, là hàm mất mát cho từng mẫu. Dạng cụ thể của *ℓ* phụ thuộc vào bài toán [12].

#### Hàm tối ưu

Hàm tối ưu (optimize function) trong mạng ANN là một mục tiêu cần tối thiểu hóa hoặc tối đa hóa, thường liên quan trực tiếp đến hàm mất mát. Hàm này hướng dẫn quá trình học bằng cách cập nhật trọng số để giảm giá trị mất mát [12].

Biểu diễn tổng quát:

Trong đó:

* : Tập trọng số của mạng.
* : Tập bias.
* : Hàm mất mát cho mẫu .

#### Lan truyền thuận

Lan truyền thuận (forward propagation) là quá trình trong mạng nơ-ron, trong đó dữ liệu đầu vào được truyền qua các lớp của mạng để tạo ra đầu ra. Quá trình này bao gồm các bước sau:

* Lớp đầu vào: Dữ liệu đầu vào được đưa vào lớp đầu vào của mạng nơ-ron.
* Lớp ẩn: Dữ liệu đầu vào được xử lý thông qua một hoặc nhiều lớp ẩn. Mỗi nơ-ron trong lớp ẩn nhận đầu vào từ lớp trước, áp dụng hàm kích hoạt cho tổng trọng số của các đầu vào này và chuyển kết quả sang lớp tiếp theo.
* Lớp đầu ra: Dữ liệu đã xử lý di chuyển qua lớp đầu ra, tại đó đầu ra cuối cùng của mạng được tạo ra. Lớp đầu ra thường áp dụng một hàm kích hoạt. Dự đoán: Đầu ra cuối cùng của mạng là kết quả dự đoán hoặc phân loại cho dữ liệu đầu vào [36].

Thuật ngữ lan truyền thuận được dùng như một cách chung để miêu tả quá trình tính toán trọng số đầu ra cho mạng. Các bước tính toán cho mỗi lớp trong quá trình lan truyền thuận đã được miêu tả chi tiết trước đó [12].

#### Lan truyền ngược

Quá trình lan truyền ngược (backward propagation) là quá trình tính toán gradient của hàm mất mát đối với các trọng số và bias trong mạng nhằm giảm lỗi. Dựa vào các gradient này, các trọng số được cập nhật thông qua thuật toán tối ưu. Sau khi thực hiện lan truyền thuận để tính đầu ra của mạng, các đầu ra được dự đoán này sẽ được so sánh với các giá trị thực tế (nhãn) để tính toán mất mát. Sau đó, từ lớp đầu ra này tính gradient của hàm mất mát đối với từng trọng số và bias thông qua đạo hàm theo chuỗi (chain rule). Quá trình này diễn ra từ lớp cuối ngược về lớp đầu tiên bằng cách dùng thuật toán tối ưu để điều chỉnh trọng số và bias nhằm giảm lỗi [12]. Quá trình lan truyền ngược như sau:

* Tại lớp đầu ra: Tính toán gradient của hàm mất mát đối với đầu vào của lớp đầu ra:

Trong đó:

* + ​: Gradient của hàm mất mát với đầu vào ​ tại lớp đầu ra.
  + ​: Đầu ra từ hàm kích hoạt tại lớp đầu ra.
  + ​: Nhãn thực tế.
* Lan truyền về lớp ẩn thứ : Gradient của hàm mất mát đối với đầu vào :

Trong đó:

* + ​: Gradient của hàm mất mát đối với đầu vào ​ của nơ-ron tại lớp .
  + ​: Trọng số kết nối từ nơ-ron tại lớp đến nơ-ron tại lớp .
  + ​: Gradient tại nơ-ron ở lớp .
  + : Đạo hàm của hàm kích hoạt tại ​.
* Gradient của trọng số và bias:
  + Gradient của trọng số : Gradient của hàm mất mát đối với ma trận trọng số tại lớp được biểu diễn tổng quát:

Trong đó:

* + - : Ma trận trọng số kết nối từ lớp đến lớp .
    - : Vector gradient của lớp , thể hiện mức độ thay đổi của hàm mất mát theo đầu vào của hàm kích hoạt tại lớp .
    - : Vector đầu ra của các nơ-ron trong lớp .
  + Gradient của bias : Gradient của hàm mất mát đối với vector bias tại lớp .

Với .

Gradient của trọng số là tích ngoài của vector gradient và vector đầu ra của lớp trước. Điều này cho phép tính toán toàn bộ các trọng số cùng lúc bằng cách sử dụng phép nhân ma trận.

* Cập nhật trọng số: Thuật toán tối ưu được dùng để điều chỉnh trọng số và bias nhằm giảm lỗi, Ma trận trọng số cho lớp sẽ được cập nhật toàn bộ như sau:

Trong đó:

* + : Ma trận trọng số của lớp với kích thước ​.
  + ​: Ma trận gradient của hàm mất mát theo ma trận trọng số .
  + : Tốc độ học (learning rate).

Lan truyền ngược chính là cốt lõi trong việc huấn luyện các mạng nơ-ron nhân tạo, giúp tối ưu hóa các trọng số của mạng bằng cách giảm thiểu sai số giữa đầu ra dự đoán và giá trị thực tế. Đây cũng là nền tảng giúp các mạng nơ-ron sâu xử lý các bài toán phức tạp như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên và dự đoán dữ liệu [19].

## Tổng quan về Xử lý ngôn ngữ tự nhiên (NLP)

### Giới thiệu về NLP

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) thuộc lĩnh vực liên ngành của khoa học máy tính và trí tuệ nhân tạo, tập trung vào việc cho phép máy tính hiểu, phân tích và tạo ra ngôn ngữ theo cách "tự nhiên" như con người. Mục tiêu của NLP là thu hẹp khoảng cách giữa ngôn ngữ của con người và ngôn ngữ máy, qua đó giúp máy tính xử lý và phản hồi lại một cách chính xác, ngữ nghĩa và bối cảnh [37].

Các nhiệm vụ điển hình trong NLP bao gồm phân tích cảm xúc của văn bản, nhận dạng giọng nói, tạo phản hồi thông minh cho câu hỏi, dịch thuật tự động v.v. Ngày nay các nghiên cứu NLP đã giúp tạo ra kỷ nguyên AI tạo sinh, từ các kỹ năng giao tiếp của các mô hình ngôn ngữ lớn (LLM) đến khả năng của các mô hình tạo hình ảnh để hiểu các yêu cầu. Trong thực tiễn, NLP cũng đóng vai trò ngày càng tăng trong các giải pháp doanh nghiệp giúp hợp lý hóa và tự động hóa các hoạt động kinh doanh, tăng năng suất của nhân viên và đơn giản hóa các quy trình kinh doanh [38].

### Học sâu trong NLP

Với sự thành công của mạng ANN, học sâu đã định hình lại cách xử lý ngôn ngữ tự nhiên, chuyển từ các phương pháp dựa trên quy tắc và thống kê truyền thống sang các mô hình học máy mạnh mẽ, tự động học các đặc trưng từ dữ liệu ngôn ngữ. Trước đây, các mô hình NLP thường được xây dựng trên các mô hình học máy đơn giản. Tuy nhiên, sự ra đời của mạng ANN và học sâu đã giúp NLP tận dụng sức mạnh của dữ liệu lớn. Một số mô hình học sâu được thiết kế đặc biệt dành cho các tác vụ NLP như:

* Mạng nơ-ron hồi quy (Recurent Neural Network - RNN):
* Mạng nơ-ron hồi quy nhớ dài – ngắn hạn (Long-short term memory - LSTM)
* Mạng GRU (Gated Recurrent Unit)
* Transformer

Khi ứng dụng mô hình ANN vào các tác vụ NLP, đòi hỏi mạng no-ron thông thường cần có những cải tiến phù hợp để có thể xử lý dữ liệu tuần tự như văn bản. Đặc biệt trong tác vụ NLP, các từ trong một câu có liên hệ mật thiết với nhau, khi những từ tiếp theo phụ thuộc lớn vào các từ trước đó trong một câu. Lúc này mạng RNN ra đời nhằm giải quyết vấn đề trên, sau đó mạng LSTM mà một cải tiến của RNN tuy nhiên phức tạp hơn, mạng GRU chính là một phiên bản cải tiến đơn giản hơn của mạng LSTM. Tuy nhiên, các mạng như RNN, LSTM và GRU gặp phải một vấn đề đó là biến mất đạo hàm khi mô hình phải xử lý dữ liệu văn bản có độ dài lớn, điều này làm mất đi thông tin quan trọng của mô hình. Và mô hình Transformer ra đời nhằm giải quyết vấn đề biến mất đạo hàm [39]. Mô hình này cũng chính là cốt lõi của các mô hình LLM sau này, nó đã được chứng minh được hiệu quả vượt trội trong việc xử lý ngôn ngữ tự nhiên.

### Các ứng dụng của NLP

Với sự phát triển mạnh mẽ của NLP và các mô hình học sâu, NLP có thể được ứng dụng đa dạng trong nhiều lĩnh vực khác nhau. Một số ứng dụng nổi bật của NLP phải được kể đến như [40]:

* Dịch máy: Dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác.
* Tạo văn bản: Tạo ra văn bản mới dựa trên ngữ cảnh đầu vào
* Chatbot hỏi đáp thông minh v.v.: Tự động hóa trong trò chuyện và chăm sóc khách hàng, trả lời câu hỏi tự động, trợ giúp khách hàng và cung cấp thông tin v.v.

## Mô hình Transformer

### Giới thiệu mô hình Transformer

Mô hình Transformer được giới thiệu bởi Vaswani và cộng sự vào năm 2017 trong bài báo "Attention is All You Need", Transformer đã mang lại những cải tiến vượt bậc về hiệu suất trong nhiều tác vụ như dịch máy, phân loại văn bản và sinh văn bản. Khác với các mô hình truyền thống dựa trên hồi quy (RNN), Transformer sử dụng cơ chế attention, cho phép nó xử lý các dữ liệu tuần tự một cách hiệu quả và linh hoạt hơn. Thay vì sử dụng cấu trúc tuần tự, Transformer áp dụng cơ chế attention để cho phép việc xử lý các từ trong câu song song, giúp tăng tốc độ huấn luyện và cải thiện khả năng biểu diễn. Transformer đã trở thành nền tảng cho nhiều mô hình ngôn ngữ lớn giúp cải thiện khả năng hiểu ngữ nghĩa và sinh ngôn ngữ tự nhiên [39].

### Tổng quan về kiến trúc mô hình Transformer

Kiến trúc của Transformer bao gồm hai phần chính: encoder và decoder. Mỗi phần này lại bao gồm nhiều lớp chồng lên nhau. Encoder chịu trách nhiệm tiếp nhận đầu vào và tạo ra các biểu diễn (representations) của nó, trong khi decoder sử dụng các biểu diễn này để sinh ra đầu ra (output) mong muốn. Tức là encoder sẽ ánh xạ một chuỗi đầu vào của các biểu diễn ký hiệu là đến một chuỗi các biểu diễn liên tục . Với cho trước, decoder sau đó tạo ra một chuỗi đầu ra của các ký hiệu một phần tử tại một thời điểm. Ở mỗi bước, mô hình tự động hồi quy [39]. Hình 3 là kiến trúc của mô hình Transformer.

A diagram of a diagram

Description automatically generated

Hình 3. Kiến trúc mô hình Transforme [39].

Nguyên lý hoạt động chủ yếu của Transformer dựa vào cơ chế attention, cho phép nó xác định mức độ quan trọng của các từ trong câu đối với nhau. Cơ chế này giúp mô hình chú ý đến các từ khác trong câu khi tạo ra các biểu diễn mới, từ đó cải thiện độ chính xác trong việc nắm bắt các mối quan hệ ngữ nghĩa và cú pháp. Mô hình này có khả năng xử lý các câu dài mà không gặp phải vấn đề về độ dài của chuỗi như trong các mô hình RNN truyền thống [39].

### Các thành phần kiến trúc mô hình Transformer

#### Encoder

Mỗi Encoder được tạo thành từ một chồng lớp giống hệt nhau. Mỗi lớp sẽ có hai lớp con (sub-layer). Lớp con đầu tiên là một lớp cơ chế tự chú ý đa đầu (multi-head self-attention). Lớp con thứ hai bao gồm mạng truyền thẳng được kết nối đầy đủ theo vị trí (position-wise fully connected feed-forward network). Các kết nối dư thừa (residual connection) của mỗi lớp con sẽ được sử dụng lại được tạo bởi lớp chuẩn hóa (layer normalization). Điều đó có nghĩa là đầu ra của mỗi lớp con là , tại là hàm thực hiện bởi chính lớp con đó. Để tạo điều kiện cho những kết nối thuận lợi này, tất cả các lớp con trong mô hình cũng như các lớp embedding (embedding layers), tạo ra các đầu ra có kích thước [39].

#### Decoder

Giống như encoder, decoder cũng bao gồm một chuỗi các lớp chồng giống hệt nhau với hai lớp con tương tự. Tuy nhiên decoder có thêm một lớp phụ thứ ba đó là (masked multi-head attention), có nhiệm vụ thực hiện sự chú ý đầu ra của ngăn xếp bộ encoder. Tương tự như encoder, chúng sử dụng các kết nối dư xung quanh mỗi lớp con, sau đó là lớp chuẩn hóa. Lớp con tự chú ý trong decoder cũng được giấu (mask). Việc che giấu này, kết hợp với thực tế là các embedding đầu ra bị lệch đi một vị trí, đảm bảo rằng các dự đoán cho vị trí chỉ có thể phụ thuộc vào các đầu ra đã biết ở các vị trí nhỏ hơn . Tức là đảm bảo rằng trong quá trình dự đoán từ tiếp theo, mô hình chỉ có thể “thấy” các từ trước đó, chứ không được dùng các từ sau [39].

#### Cơ chế Attention

Cơ chế attention là thành phần trung tâm của Transformer, cho phép mô hình tập trung vào các phần khác nhau của đầu vào khi tạo ra đầu ra. Cơ chế này hoạt động bằng cách tính toán một ma trận attention score, thể hiện mức độ quan trọng của từng từ trong đầu vào đối với từ hiện tại mà mô hình đang xử lý. Các scores này được chuyển đổi thành xác suất thông qua hàm softmax, từ đó tạo ra một trọng số cho mỗi từ. Kết quả là mô hình có thể tạo ra các biểu diễn mà nhấn mạnh các từ quan trọng hơn trong ngữ cảnh cụ thể [41].

Trong các nhiệm vụ ứng dụng NLP nói chung, cần phải thực hiện biến đổi mỗi từ ở đầu vào thành một vector bằng thuật toán embedding (embedding algorithm). Cụ thể mỗi từ sẽ được embedding thành vector có kích thước là [41]. Hình 4 là minh họa đơn giản cho việc áp dụng thuật toán embedding embedding cho mỗi từ trong câu. Mỗi từ đầu ra sẽ có kích thước .



Hình 4. Minh hoạt các từ được áp dụng thuật toán Embedding.

Việc embedding này xảy ra ở bộ mã hóa (encoder). Và kết quả cho tất cả bộ encoder là nhận được một vector có kích thước . Đối với các bộ encoder ở dưới cùng, chúng là các từ embedding, nhưng trong các bộ encoder khác, đó sẽ là đầu ra của bộ mã hóa nằm ngay bên dưới. Kích thước danh sách này là các siêu tham số mà có thể thiết lập tùy chỉnh. Về cơ bản thì nó sẽ là độ dài của câu dài nhất trong tập dữ liệu đào tạo [41].

Bước đầu tiên trong việc tính toán self-attention là tạo ra ba vector từ mỗi vector đầu vào của bộ encoder (trong trường hợp này là embedding của mỗi từ). Vì vậy đối với mỗi từ chúng tạo ra một vector truy vấn (query vector - ), một vector khóa (key vector - và một vector giá trị (value vector - ). Các vector này được tạo ra bằng cách nhân chúng với ba ma trận mà chúng đã đào tạo trong quá trình đào tạo (lần lượt là ). Các vector này có kích thước nhỏ hơn vector embedding và có kích thước là (64 dimention), trong khi vector đầu vào hoặc đầu ra của encoder và embedding có kích thước là . Phép nhân với ma trận trọng số tạo ra , vector truy vấn liên kết với từ đó. Sau đó chúng tạo ra một phép chiếu truy vấn, khóa và giá trị của mỗi từ trong câu đầu vào [41].

A screenshot of a computer

Description automatically generated

Hình 5. Vector tham số của mô hình Transformer [41].

Bước tiếp theo trong việc tính toán self-attention là tính chỉ số score. Giả sử đang tính toán self-attention cho từ đầu tiên x1, ta cần tính điểm (score) cho từng từ của câu đầu vào so với từ này. Score xác định mức độ vào các phần khác của câu đầu vào khi chúng ta mã hóa một từ ở một vị trí nhất định và được tính bằng cách lấy tích vô hướng của query vector với key vector của từ tương ứng đang tính score. Vì thế nếu xử lý self-attention đối với từ ở vị trí số 1, score đầu tiên sẽ là tích vô hướng của và . Và score thứ 2 sẽ là tích vô hướng của và [41].

Bước tiếp theo là chia score cho 8 (do căn bậc hai của chiều của các vector chính được sử dụng mặc định là 64, điều này dẫn đến việc có các gradiant ổn định hơn. Trên thực tế con số này có thể được thay đổi), sau đó truyền kết quả qua phép toán softmax. Softmax chuẩn hóa các score để tất cả đều dương và cộng lại bằng [41]. Công thức như sau:

Trong đó:

* là giá trị đầu vào thứ .
* là tổng giá trị đầu vào.

Kết quả của hàm softmax (hình 6) thể hiện mức độ mỗi từ sẽ được thể hiện ở vị trí này. Từ phù hợp ở vị trí này sẽ có điểm softmax cao nhất, nhưng đôi khi cũng hữu ích khi chú ý đến một từ khác có liên quan đến hiện tại [41].

A screenshot of a computer

Description automatically generated

Hình 6. Tính toán chỉ số Score cho các từ trong câu [41].

Bước tiếp theo là nhân mỗi value vector với điểm softmax. Ở bước này chúng ta cần giữ nguyên giá trị của từ hoặc các từ quan trọng và loại bỏ các từ không liên quan bằng cách nhân với các số nhỏ (learning rate = 0.001 chẳng hạn) . Bước cuối cùng là cộng các value vector có trọng số. Điều này tạo ra đầu ra của lớp self-attention tại vị trí xem xét [41]. Kết thúc quá trình tính self-attention, value vector là một vector mà chúng ta có thể gửi đến mạng nơ-ron thần kinh truyền thẳng. Tuy nhiên trong quá trình triển khai thực tế, phép tính này được thực hiện dưới dạng ma trận.

**Ma trận tính toán self-attention:** Tính toán ma trận query, key và value được thực hiện bằng cách đóng gói các embedding vào một ma trận X và nhân nó với các ma trận trọng số đã được đào tạo trước ().

A group of squares with different colored squares

Description automatically generated

Hình 7. Ma trận tính toán self-attention [41].

Mỗi dòng trong ma trận X tương ứng với một từ trong câu đầu vào và sau đó đưa vào hàm softmax để tính toán đầu ra của lớp self-attention.

#### Multi-head Attention

Multi-head attention là một kỹ thuật mở rộng của cơ chế attention, cho phép mô hình học nhiều mặt khác nhau của dữ liệu cùng một lúc. Thay vì chỉ sử dụng một cơ chế attention duy nhất, multi-head attention tạo ra nhiều head attention, mỗi head có thể tập trung vào các phần khác nhau của đầu vào. Kết quả từ các head này sau đó được kết hợp lại, giúp mô hình có cái nhìn toàn diện hơn về ngữ cảnh và cải thiện khả năng hiểu ngôn ngữ tự nhiên [42]. Việc sử dụng multi-head attention cải thiện hiệu suất của lớp chú ý theo hai cách [41]:

* Tăng khả năng tập trung vào các vị trí khác nhau của mô hình. Thật vậy, trong hình 6, z1 chứa thông tin của tất cả mã hóa khác kể cả chính nó và có thể sẽ bị chi phối bởi điều này. Trên thực tế, nếu chúng ta đang dịch một câu như: “Con vật không băng qua đường vì nó quá mệt”, việc chứa thông tin mã hóa sẽ rất có ích khi biết từ “nó” đang ám chỉ từ nào [41].
* Cung cấp cho lớp attention nhiều “không gian con biểu diễn” (representation subspaces). Với multi-head attention, mô hình không chỉ có một mà là nhiều bộ ma trận trọng số Query, Key và Value (Transformer sử dụng tám multi-head attention, do đó chúng kết thúc với tám bộ cho mỗi bộ encoder và decoder). Mỗi bộ trong số bộ này được khởi tạo ngẫu nhiên. Sau khi đào tạo, mỗi bộ được sử dụng để chiếu các embedding input (hoặc các vector từ bộ encoder và decoder trước đó) vào một không gian con biểu diễn khác nhau [41].

A chart of different colored squares

Description automatically generated

Hình 8. Minh hoạt multi-head attention và ma trận trọng số Q, K, V cho mỗi từ [41].

Với multi-head attention, các ma trận trọng số , , được duy trì riêng biệt cho mỗi head, từ đó dẫn đến mỗi , , đều khác nhau. Như đã đề cập ở trên, X được nhân với các ma trận , , để tạo ra các ma trận , , . Nếu cùng thực hiện một phép tính attention tám lần khác nhau với các ma trận trọng số khác nhau. Chúng ta sẽ có tám ma trận khác nhau [41].

A diagram of a diagram

Description automatically generated with medium confidence

Hình 9. Minh hoạt biểu diễn các attention cho mỗi từ [41].

Tuy nhiên trước khi đưa vào lớp feed-forward, ta cần đưa về thành một vector duy nhất (cũng cần lưu ý rằng với mỗi là một từ) bằng cách nối các ma trận này với nhau và sau đó nhân với một ma trận trọng số bổ sung [41].

A diagram of a test

Description automatically generated with medium confidence

Hình 10. Nhân ma trận trọng số để định hình lại kích thước của Z [41].

Ma trận **​** được áp dụng lên vector để đưa kích thước của nó về lại kích thước ban đầu của model ​. ​Với là một ma trận trọng số huấn luyện được thêm vào để kết hợp các thông tin từ các heads và giảm kích thước vector về ban đầu. Ma trận ​ này là một tham số của mô hình, tức là nó được khởi tạo ngẫu nhiên ban đầu và sau đó sẽ được huấn luyện cùng với các tham số khác trong quá trình tối ưu hóa. Nó không phải là một ma trận cố định hay được tính toán từ đầu vào mà sẽ tự động điều chỉnh để tối ưu hóa kết quả của mô hình. Hình 11 dưới đây là bản tóm tắt ngắn gọn về tất cả những gì cho multi-head attention.

A group of colorful squares

Description automatically generated

Hình 11. Minh họa tóm tắt quá trình tính toán multi-heads attention [41].

#### Lớp Add

Kết nối Residual còn được gọi là kết nối bỏ qua (skip connection)[[1]](#footnote-1), là một kỹ thuật được sử dụng trong mạng nơ-ron sâu để tạo điều kiện thuận lợi cho việc đào tạo các kiến ​​trúc rất sâu [43]. Kỹ thuật này được giới thiệu trong ResNet (Mạng lưới dư thừa) bởi Kaiming He và cộng sự vào năm 2015 [44]. Mạng lưới này dựa trên một khái niệm được gọi là Học dư thừa (Residual learning). Ý tưởng chính của khái niệm này là cho phép đầu vào của một lớp bỏ qua các hoạt động của lớp đó và được thêm trực tiếp vào đầu ra của lớp đó. Các hoạt động có thể là bất cứ thứ gì như Biến đổi tuyến tính, Biến đổi phi tuyến tính, Chuẩn hóa, Bỏ qua, Gộp, v.v. [43].

A diagram of a layer

Description automatically generated

Hình 12. Minh hoạ một kết nối dư trong mạng nơ-ron [43].

Cơ chế hoạt động của residual connection như sau. Giả sử ta có một mạng nơ-ron một (hoặc nhiều) lớp và mục tiêu là tìm một hàm để biểu diễn một lớp mạng nhất định, trong trường hợp này là hàm mà chúng ta muốn mô hình hoá. Với residual learning, thay vì trực tiếp huấn luyện mạng để mô phỏng hàm , chúng ta sẽ huấn luyện để mạng mô phỏng một hàm . Phần residual này được định nghĩa sao cho khi cộng với , chúng có thể khớp với . Hay nói cách khác [43]. Hình 13 so sánh sự khác nhau khi huấn luyện một mạng nơ-ron truyền thống và sử dụng kĩ thuật residual learning.

A diagram of a layer

Description automatically generated

Hình 13. Mạng nơ-ron thường (bên trái) và mạng nơ-ron có kết nối dư (bên phải) [43].

Ưu điểm của Residual Learning trong Transformers là tạo điều kiện cho việc truyền tín hiệu theo cả hai hướng lan truyền thuận (forward propagation) và lan truyền ngược (backward propagation), giải quyết vấn đề biến mất đạo hàm (Vanishing gradient)[[2]](#footnote-2). Các mạng phức tạp với các lớp sâu hơn như transformer gặp phải vấn đề về biến mất của đạo hàm (vanishing gradient) và skip connection giúp khôi phục gradient bằng cách thực hiện ánh xạ [43].

#### Lan truyền thuận

Như đã đề cập trong các phần trước, làn tuyền thuận là quá trình tính toán của mạng học sâu để cho kết quả đầu ra mô hình. Tương tự đối với mô hình Transformer, tuy nhiên từ hình 13 chúng ta biết rằng, mối quan hệ cho quá trình lan truyền ngược có thể được mô tả bằng do tồn tại kết nối dư , sẽ là đầu vào cho một tập hợp các lớp khác trong mạng, nhưng để đơn giản hóa, chúng ta chỉ xem xét một đơn vị kết nối dư cho trường hợp này [41].

#### Lan truyền ngược

Đây là nơi kết nối dư giải quyết vấn đề biến mất đạo hàm. Để tính toán gradient, chúng ta xem xét riêng đạo hàm riêng của hàm mất mát đối với đầu vào . Sử dụng công thức trên, phương trình lan truyền ngược sẽ là [41]:

#### Lớp Norm

Lớp chuẩn hóa (Normalization) sẽ chuẩn hóa đầu ra của những bước tính toán trước đó trên các features. Trong quá trình chuẩn hóa lớp, mỗi mục trong lô được chuẩn hóa trong phạm vi từ đến . Trong mô hình transformer, nó có thể là embedding vectơ có 512 chiều của một từ. Chúng ta tính toán giá trị trung bình và phương sai của các mục này và thay thế từng giá trị trong ma trận bằng:

Trong quá trình đào tạo mô hình transformer, các đầu ra do các nơ-ron trong một lớp tạo ra sau khi áp dụng hàm kích hoạt cho tổng trọng số của các đầu vào được gọi là các kích hoạt. Sự phân bố của các kích hoạt này thay đổi theo thời gian do các tham số mạng thay đổi. Chuẩn hóa làm giảm sự thay đổi này bằng cách duy trì sự phân bố ổn định của các kích hoạt, giúp quá trình đào tạo trở nên trơn tru [41].

Khi xử lý dữ liệu tuần tự, chuẩn hóa theo lô thường không được khuyến khích vì độ dài của các câu thay đổi, thay vào đó sẽ sử dụng chuẩn hóa theo lớp (Layer Normalization). Chuẩn hóa theo lô được áp dụng trên một lô nhỏ. Nếu các chuỗi có độ dài thay đổi, điều này có thể yêu cầu thêm phần đệm hoặc cắt bớt, có thể gây nhiễu trong dữ liệu. Vấn đề đệm này ảnh hưởng đến việc tính toán giá trị trung bình và phương sai trên toàn bộ lô nhỏ và do đó đưa nhiễu vào quá trình chuẩn hóa. Chính vì thế chuẩn hóa sẽ được dùng [41].

#### Masked Multi Head Attention

Masked Multi Head Attention là multi head attention, có chức năng dùng để encode các từ câu đích trong quá trình dịch. Lúc cài đặt cần phải che đi các từ ở tương lai chưa được mô hình dịch đến, bằng cách nhân với một vector chứa các giá trị 0,1. Trong decoder còn có một multi head attention khác có chức năng chú ý các từ ở mô hình encoder, layer này nhận vector key và value từ mô hình encoder, và output từ layer phía dưới. Do muốn so sánh sự tương quan giữa từ đang được dịch với các từ nguồn [45].

#### Positional Encoding

Word embeddings giúp biểu diễn ngữ nghĩa của một từ, tuy nhiên cùng một từ ở vị trí khác nhau của câu lại mang ý nghĩa khác nhau. Do đó Transformers có thêm một phần Positional Encoding để đưa thêm thông tin về vị trí của một từ [45].

Trong đó pos là vị trí của từ trong câu, là giá trị phần tử thứ trong embeddings có độ dài . Sau đó cộng vector và Embedding vector.

A diagram of a mathematical equation

Description automatically generated

Hình 14. Mã hóa vị trí từ embedding [45].

### Ứng dụng Transformer trong Xử lý ngôn ngữ tự nhiên

Mô hình Transformer đã mang lại những cải tiến vượt bậc trong lĩnh vực NLP nhờ cơ chế self-attention và khả năng xử lý song song dữ liệu. Một số ứng dụng quan trọng của Transformer trong NLP phải kể đến như:

* Dịch máy
* Trả lời câu hỏi (Question Answering - QA)
* Phân tích cảm xúc
* Sinh văn bản v.v.

Một trong những ứng dụng lớn nhất của Transformer là xây dựng các mô hình ngôn ngữ lớn, Transformer đã trở thành nền tảng của các mô hình LLM) này, với sự thành công vượt trội nhờ khả năng xử lý ngữ cảnh và dữ liệu lớn một cách hiệu quả. Các mô hình LLM như Grok, Llama, Gemini và các phiên bản tiên tiến hơn đều dựa trên kiến trúc Transformer.

## Tổng quan về Mô hình ngôn ngữ lớn (LLM)

### Giới thiệu về LLM

Mô hình ngôn ngữ lớn (Large Language Model - LLM) là một loại chương trình [AI](https://www.cloudflare.com/learning/ai/what-is-artificial-intelligence/) có thể nhận dạng và tạo văn bản, trong số các tác vụ khác. LLM được đào tạo trên [các tập dữ liệu khổng lồ](https://www.cloudflare.com/learning/ai/big-data/) do đó có tên là “lớn”. LLM được xây dựng trên [máy học](https://www.cloudflare.com/learning/ai/what-is-machine-learning/): cụ thể là một loại [mạng nơ-ron](https://www.cloudflare.com/learning/ai/what-is-neural-network/) được gọi là mô hình Transformer [46].

Nói một cách đơn giản hơn, LLM là một chương trình máy tính đã được cung cấp đủ ví dụ để có thể nhận dạng và diễn giải ngôn ngữ con người hoặc các loại dữ liệu phức tạp khác. Nhiều LLM được đào tạo trên dữ liệu đã được thu thập từ Internet hàng nghìn hoặc hàng triệu gigabyte văn bản. LLM sử dụng một loại máy học gọi là [học sâu](https://www.cloudflare.com/learning/ai/what-is-deep-learning/) để hiểu cách các ký tự, từ và câu hoạt động cùng nhau. Học sâu liên quan đến phân tích xác suất dữ liệu phi cấu trúc, cuối cùng cho phép mô hình học sâu nhận ra sự khác biệt giữa các phần nội dung mà không cần sự can thiệp của con người [46], [47].

A screenshot of a computer

Description automatically generated

Hình 15. Minh họa xử lý đa tác vụ của một LLM.

Sau đó, các LLM được đào tạo thêm thông qua quá trình điều chỉnh, mô hình được điều chỉnh chính xác hoặc điều chỉnh theo yêu cầu cho nhiệm vụ cụ thể mà lập trình viên muốn nó thực hiện (được gọi là là fine-tuning model), chẳng hạn như diễn giải câu hỏi và tạo phản hồi hoặc dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác [46].

### Quy trình huấn luyện một LLM

A diagram of a work flow

Description automatically generated

Hình 16. Quá trình huấn luyện một mô hình LLM.

Hình 16 mô tả ngắn gọn quy trình để huấn luyện một LLM, về cơ bản các bước huấn luyện một LLM tương tự với việc huấn luyện các mô hình học máy khác. Tuy nhiên các kỹ thuật phức tạp và các phương pháp học nâng cao cũng sẽ được sử dụng để huấn luyện mô hình. Bước đầu tiên (1) sẽ là chuẩn bị dữ liệu huấn luyện, các dữ liệu văn bản sẽ được lấy từ nhiều nguồn khác nhau và có thể được kiểm soát chặt chẽ, bởi vì chất lượng văn bản sẽ ảnh hưởng lớn đến việc huấn luyện một LLM. Bước thứ hai (2) sẽ là xây dựng kiến trúc mô hình. Các mô hình LLM tuy dựa trên kiến trúc của Transformer nhưng phải được điều chỉnh nhiều để có thể tạo ra một mô hình mạnh mẽ. Bước thứ ba (3) là huấn luyện mô hình, bước này sẽ huấn luyện mô hình một cách tổng quan trước, sau đó mô hình có thể sẽ được sử dụng tiếp để tinh chỉnh cho những tác vụ cụ thể gọi là fine-tuning ở bước (4). Bước thứ năm (5) sẽ là thực hiện đánh giá mô hình bằng các phương pháp đánh giá chuyên biệt dành cho LLM. Tổng thể quá trình từ bước (1) đến bước (5) có thể thực hiện lại nhiều lần và bước cuối cùng (6) là triển khai [48].

#### Đào tạo trước (Pre-training)

Trong huấn luyện một LLM, hai bước chủ yếu là đào tạo trước và tinh chỉnh mô hình. Trong giai đoạn pre-training, các bộ transformer được đào tạo trên một lượng lớn dữ liệu văn bản thô, internet là nguồn dữ liệu chính. Quá trình đào tạo được thực hiện bằng các kỹ thuật học không giám sát, một loại đào tạo sáng tạo không yêu cầu hành động của con người để dán nhãn dữ liệu. Mục tiêu của quá trình đào tạo trước là tìm hiểu các mẫu thống kê của ngôn ngữ. Chiến lược tiên tiến để đạt được độ chính xác cao hơn của kiến trúc Transformer là làm cho mô hình lớn hơn, điều này có thể đạt được bằng cách tăng số lượng tham số và tăng kích thước của dữ liệu đào tạo. Do đó, hầu hết các LLM tiên tiến đều có hàng tỷ tham số và đã được đào tạo trên một khối dữ liệu khổng lồ [49]. Với quy mô của mô hình và dữ liệu đào tạo, quá trình đào tạo trước thường tốn thời gian và tốn kém chi phí.

#### Tinh chỉnh mô hình (Fine-tuning)

Quá trình đào tạo trước cho phép kiến trúc Transformer có được hiểu biết cơ bản về ngôn ngữ, nhưng không đủ để thực hiện các tác vụ thực tế cụ thể với độ chính xác cao. Để tránh các lần lặp lại tốn thời gian và chi phí trong quá trình đào tạo, Transformer tận dụng các kỹ thuật học chuyển giao để tách giai đoạn (tiền) đào tạo khỏi giai đoạn tinh chỉnh. Điều này cho phép các nhà phát triển chọn các mô hình được đào tạo trước và tinh chỉnh chúng dựa trên cơ sở dữ liệu hẹp hơn, cụ thể theo miền. Trong nhiều trường hợp, quá trình tinh chỉnh được thực hiện với sự hỗ trợ của người đánh giá, sử dụng một kỹ thuật gọi là Học tăng cường từ phản hồi của con người. Quá trình đào tạo hai bước cho phép điều chỉnh LLM cho nhiều tác vụ hạ nguồn. Nói cách khác, tính năng này biến LLM thành mô hình nền tảng của vô số ứng dụng được xây dựng trên chúng, và đây chính là cơ sở để có thể xây dựng một chatbot tuyển sinh theo nhu cầu và thiết thế tùy chỉnh [49].

### Các kỹ thuật NLP trong LLM

#### Định nghĩa Token

Trong Xử lý Ngôn ngữ Tự nhiên, một token là đơn vị cơ bản của văn bản được xử lý bởi các thuật toán. Mỗi token có thể là một từ, một ký tự, hoặc một chuỗi ký tự có ý nghĩa trong ngữ cảnh cụ thể. Ví dụ câu: *"Tôi yêu lập trình."*:

* Token hóa theo từ: ["Tôi", "yêu", "lập", "trình", "."]
* Token hóa theo ký tự: ["T", "ô", "i", " ", "y", "ê", "u", " ", "l", "ậ", "p", " ", "t", "r", "ì", "n", "h", "."]
* Token hóa theo subword: ["Tôi", "yêu", "lập", "##trình"]

Token không thể xử lý trực tiếp bởi máy tính, nên cần được biểu diễn bằng các dạng **số hóa**. Có nhiều cách biểu diễn token trong NLP, một vài cách phổ biến nhất gồm:

* One-Hot Encoding: Gán mỗi token một vector nhị phân.
* Index Mapping: Gán mỗi token một số nguyên duy nhất.
* Embedding: Sử dụng thuật toán embedding để tạo vector biểu diễn ngữ nghĩa.

#### Kỹ thuật Chunking

Chunking là một kỹ thuật trong xử lý ngôn ngữ tự nhiên và phân tích văn bản. Nó giúp phân tích văn bản lớn thành các phân đoạn hoặc khối nhỏ, dễ quản lý, giúp xử lý và phân tích khối lượng dữ liệu lớn dễ dàng hơn. Kỹ thuật chunking thường được sử dụng trong nhiều ứng dụng khác nhau, chẳng hạn như tóm tắt nội dung, đánh giá tình cảm hoặc trích xuất thông tin chính. Trong mọi trường hợp, nó đóng vai trò quan trọng trong việc nâng cao hiệu suất của ứng dụng, giúp dễ dàng xử lý, phân tích hoặc đưa vào mô hình. Việc này đặc biệt quan trọng với các mô hình ngôn ngữ lớn (LLM) vì chúng có giới hạn về số lượng token có thể xử lý cùng lúc (context length) [50], [51]. Các kỹ thuật chunking phổ biến bao gồm:

* Theo Độ Dài Cố Định: Chia văn bản thành các đoạn có số lượng token cố định, ví dụ mỗi đoạn có 8000 token.

Giả sử ta có một văn bản với token, ta chia thành đoạn:

​ là token thứ , với là số token mỗi chunk. Mỗi chunk ​ là:

với .

* Theo Cấu Trúc Ngữ Nghĩa: Chia văn bản theo đoạn văn, câu, hoặc theo các mục logic (chương, tiêu đề).
* Sliding Window (Cửa Sổ Trượt): Dùng một cửa sổ cố định (ví dụ 8000 token) và trượt qua văn bản, với một khoảng trùng lặp (overlap) để không mất ngữ cảnh.

Với , khung token và khoảng trượt (overlap). Chunk thứ :

với chạy đến khi .

Ví dụ văn bản: "*Học máy là một lĩnh vực khoa học dữ liệu phát triển nhanh chóng. LLM có khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ.*" Token hóa: [*Học, máy, là, một, lĩnh, vực, khoa, học, ...*].

* Chia Theo Độ Dài Cố Định (L = 5):
  + Chunk 1: [Học, máy, là, một, lĩnh]
  + Chunk 2: [vực, khoa, học, dữ, liệu]
  + Chunk 3: [phát, triển, nhanh, chóng, LLM]
  + Chunk 4: [có, khả, năng, xử, lý]
  + Chunk 5: [ngôn, ngữ, tự, nhiên, mạnh]
  + Chunk 6: [mẽ].
* Sliding Window (L = 5, O = 2):
  + Chunk 1: [Học, máy, là, một, lĩnh]
  + Chunk 2: [một, lĩnh, vực, khoa, học]
  + Chunk 3: [vực, khoa, học, dữ, liệu]
  + ...

Lý do cần phải chunking cho dữ liệu vì hầu hết các mô hình LLM bị giới hạn ngữ cảnh, các LLM chỉ hỗ trợ số lượng token nhất định. Nếu văn bản quá dài, chunking giúp chia nhỏ văn bản để không vượt quá giới hạn này. Việc chunking được thiết kế để giữ cho các đoạn văn bản vẫn có ý nghĩa riêng lẻ, giúp bảo toàn ngữ cảnh. Ngoài ra, chunking giúp lưu trữ và tìm kiếm dễ dàng hơn, tăng hiệu quả tìm kiếm khi dùng embedding hoặc vector database [50], [51].

#### Kỹ thuật Embedding

Embedding là kỹ thuật chuyển đổi dữ liệu (thường là văn bản, hình ảnh, hoặc âm thanh) thành vector số học (tập hợp các số thực) để mô hình học máy hoặc thuật toán có thể xử lý. Ý tưởng chính chính là embeddings cố gắng biểu diễn ngữ nghĩa hoặc đặc trưng của dữ liệu trong không gian vector [52].

Giả sử chúng ta có một tập dữ liệu văn bản , trong đó ​ là một đoạn văn bản hoặc câu. Embedding là ánh xạ từ văn bản sang một không gian vector : , với , trong đó ​ là vector biểu diễn embedding của văn bản , và là số chiều của không gian vector (embedding dimension). Các embeddings thường dùng khoảng cách Euclidean hoặc cosine để đo độ tương tự giữa hai vector ​ và [52]:

* Khoảng cách Euclidean:
* Cosine Similarity:

Trong đó:

* + ​ là tích vô hướng (dot product) của hai vector.
  + là norm (độ dài) của các vector, được tính bằng:

Embedding sẽ tạo ra một không gian vector sao cho các điểm gần nhau có tính chất tương tự. Với ngữ nghĩa [52]:

Trong đó:

* + là các vector embedding tương ứng của các văn bản ​.
  + là khoảng cách Euclidean giữa hai vector.

Các điểm vector gần nhau trong không gian embedding thường có ý nghĩa tương tự nhau trong ngữ cảnh dữ liệu. Phương pháp này đặc biệt hữu ích với những đặc trưng hạng mục có số phần tử lớn ở đó phương pháp chủ yếu để biểu diễn mỗi giá trị thường là một vector dạng one-hot. Một cách lý tưởng, các giá trị có ý nghĩa tương tự nhau nằm gần nhau trong không gian embedding [52].

#### Kỹ thuật RAG

Tăng cường truy xuất tạo sinh (Retrieval Augmented Generation – RAG) là một kỹ thuật giúp nâng cao khả năng của mô hình LLM kết hợp với tri thức bên ngoài (external knowledge). Trong sự ra đời của các mô hình LLM rất hay gặp vấn đề ảo giác, trong từ điển chuyên ngành gọi là hallucination, khi đó các mô hình sinh ra những câu văn nghe rất trôi chảy và hợp lý, nhưng trong thực tế nó lại không chính xác. Chính vì đó là nguồn cảm hứng cho việc kỹ thuật RAG ra đời và như là một phương pháp để giải quyết vấn đã nêu [53].

Giả sử cơ sở dữ liệu có tập tài liệu , Một truy vấn được biểu diễn dưới dạng vector embedding , mỗi tài liệu ​ có embedding . Bài toán Retrieval là tìm tập sao cho:

Trong đó:

* : Hàm đo độ tương đồng giữa truy vấn và tài liệu được tính theo công thức khoảng cách Euclidean hoặc sự tương đồng Cosine Similarity.
* : Tập hợp các tài liệu phù hợp nhất.

Trong các mô hình ngôn ngữ truyền thống, phản hồi được tạo ra chỉ dựa trên các mẫu và thông tin đã học trước trong giai đoạn đào tạo. Tuy nhiên, các mô hình này vốn bị giới hạn bởi dữ liệu mà chúng được đào tạo, thường dẫn đến các phản hồi có thể thiếu chiều sâu hoặc kiến ​​thức cụ thể. RAG giải quyết hạn chế này bằng cách kéo dữ liệu bên ngoài khi cần trong quá trình tạo [54]. Về cơ bản, đây là một quy trình gồm hai phần bao gồm một thành phần truy xuất (retrieval) và một thành phần tạo (generation). Cách thức hoạt động của RAG được mô tả trong hình 17:

A diagram of a algorithm

Description automatically generated

Hình 17. Mô hình luồng hoạt động của kỹ thuật RAG.

* Thành phần Retrieval: Có chức năng tìm các tài liệu hoặc thông tin có liên quan có thể giúp trả lời câu hỏi truy vấn. Nó lấy truy vấn đầu vào và tìm kiếm trong cơ sở dữ liệu để thu thập thông tin có thể hữu ích cho việc tạo phản hồi.
* Thành phần Generator: Trình tạo là mô hình LLM tạo ra văn bản đầu ra cuối cùng. Nó lấy truy vấn đầu vào và ngữ cảnh được trình thu thập để tạo ra phản hồi mạch lạc và có liên quan. Trình tạo không hoạt động riêng lẻ, nó sử dụng ngữ cảnh do retrieval cung cấp để đưa ra phản hồi, đảm bảo rằng đầu ra không chỉ hợp lý mà còn giàu chi tiết và chính xác.

Cụ thể từng bước cách hoạt động của RAG như sau [54]:

* Xử lý truy vấn (Query processing): Tất cả bắt đầu bằng một truy vấn. Đây có thể là một câu hỏi, một lời nhắc hoặc bất kỳ đầu vào nào mà chúng ta muốn mô hình LLM phản hồi.
* Mô hình embedding (Embedding model): Sau đó truy vấn được chuyển đến mô hình embedding. Mô hình này chuyển đổi truy vấn thành vector, là biểu diễn số có thể được hệ thống hiểu và xử lý.
* Truy xuất Cơ sở dữ liệu vector: Vectơ truy vấn được sử dụng để tìm kiếm trong cơ sở dữ liệu vector database. Cơ sở dữ liệu này chứa các vector được tính toán trước của các ngữ cảnh tiềm năng mà mô hình có thể sử dụng để tạo phản hồi. Hệ thống truy xuất các ngữ cảnh có liên quan nhất dựa trên mức độ khớp chặt chẽ của các vector của chúng với vector truy vấn.
* Truy xuất bối cảnh (Retrieved Contexts): Các bối cảnh đã được truy xuất sau đó được chuyển đến mô hình LLM. Các bối cảnh này chứa thông tin mà LLM sử dụng để tạo ra phản hồi có hiểu biết và chính xác.
* Tạo phản hồi LLM: LLM tính đến cả truy vấn gốc và ngữ cảnh được truy xuất để tạo ra phản hồi toàn diện và có liên quan. Nó tổng hợp thông tin từ ngữ cảnh để đảm bảo rằng phản hồi không chỉ dựa trên kiến ​​thức đã có từ trước mà còn được bổ sung các chi tiết cụ thể từ dữ liệu được truy xuất.
* Phản hồi cuối cùng: Cuối cùng, LLM đưa ra phản hồi, phản hồi này được thông báo bởi dữ liệu bên ngoài thu thập được trong quá trình này, giúp phản hồi chính xác và chi tiết hơn.

Việc kết hợp RAG với mô hình ngôn ngữ LLM có khả năng tra cứu thông tin trực tiếp, như thể AI “đọc sách” trước khi trả lời. Điều này giúp phản hồi chính xác, sâu sắc và bám sát thực tế hơn, điều này đặc biệt hữu ích khi chúng ta muốn sử dụng dữ liệu chuyên biệt của chúng ta để có thể huấn luyện chatbot, đánh dấu một bước tiến lớn trong việc hiểu và tạo ngữ cảnh [54].

#### Thuật toán sắp xếp mức độ ưu tiên Reciprocal Rank Fusion

Trong quá trình truy xuất thông tin, nếu trình truy xuất không thể truy xuất các tài liệu có liên quan từ trình truy xuất, thì độ chính xác sẽ thấp và khả năng xảy ra ảo giác sẽ tăng lên [55]. Chính vì thế cần có một thuật toán để tính toán mức độ chính xác cho thông tin truy vấn được, trong dự án này sử dụng phương pháp Hợp nhất thứ hạng qua lại (Reciprocal Rank Fusion - RRF), đây là một kỹ thuật phổ biến trong việc kết hợp và sắp xếp lại các kết quả từ nhiều bộ máy tìm kiếm hoặc mô hình truy xuất thông tin khác nhau, đặc biệt là trong các hệ thống RAG. Phương pháp này giúp cải thiện độ chính xác và tính đa dạng của kết quả [55], [56]. Công thức toán biểu diễn của FFR như sau:

Trong đó:

* : Tập hợp các nguồn hoặc mô hình.
* : Vị trí xếp hạng của tài liệu trong nguồn .
* : Một hằng số để đảm bảo các tài liệu ở thứ hạng cao không bị ảnh hưởng quá nhiều (thường ).

Các bước hoạt động của RFF trong hệ thống RAG thường tuân theo các bước sau [55]:

* Truy vấn của người dùng (User Query): Quy trình bắt đầu khi người dùng nhập câu hỏi hoặc truy vấn.
* Nhiều bộ thu thập (Multiple Retrievers): Truy vấn được gửi đến nhiều bộ thu thập. Đây có thể là các mô hình thu thập khác nhau (ví dụ: dày đặc, thưa thớt, lai).
* Xếp hạng riêng lẻ (Individual Rankings): Mỗi bộ thu thập tạo ra thứ hạng riêng của các tài liệu có liên quan.
* Hợp nhất RRF (RRF Fusion): Thứ hạng từ tất cả các bộ thu thập được kết hợp bằng công thức RRF.
* Xếp hạng cuối cùng (Final Ranking): Một thứ hạng thống nhất được tạo ra dựa trên điểm RRF.
* Tạo (Generation): Mô hình tạo sử dụng các tài liệu được xếp hạng cao nhất để tạo ra câu trả lời cuối cùng.

Thuật toán FFR có ưu điểm đơn giản và hiệu quả mà lại dễ dàng triển khai. Trong hệ thống RAG, RRF giúp sắp xếp lại các kết quả một cách khách quan trong việc sắp xếp lại thứ hạng của các kết quả, giúp cải thiện chất lượng tổng thể của top-k kết quả bằng cách kết hợp ưu điểm từ nhiều nguồn, phù hợp với các bài toán yêu cầu tính đơn giản, hiệu quả và khả năng mở rộng [55], [56].

#### Tác nhân mô hình ngôn ngữ lớn (LLM Agents)

Các mô hình LLM có khả năng hiểu ngữ cảnh và sinh văn bản tốt, tuy nhiên để một LLM có khả năng suy nghĩ và trả lời chính xác điều này phức tạp hơn rất nhiều và cần có một quy trình cụ thể để hướng dẫn LLM có thể trả lời chính xác câu hỏi mong muốn, đó chính là các tác nhân LLM (LLM agents).

Mặc dù LLM thông thường đã cho thấy khả năng xử lý ngôn ngữ tự nhiên ấn tượng và cung cấp câu trả lời dựa trên dữ liệu đào tạo, chúng vẫn còn hạn chế trong việc thực hiện các tác vụ phức tạp và yêu cầu quản lý quy trình đa bước. LLM agents không chỉ hiểu và xử lý ngôn ngữ mà còn có khả năng lên kế hoạch, ra quyết định, và thực hiện các hành động cụ thể dựa trên ngữ cảnh và mục tiêu đề ra. Điều này cho phép LLM agents hoạt động một cách linh hoạt và hiệu quả hơn trong các tình huống thực tế, giúp tối ưu hóa quy trình làm việc và nâng cao trải nghiệm hiệu suất [57], [58]. Hình 18 là minh hoạ cách một tác nhân LLM tương tác với các thành phần khác.

A diagram of a company

Description automatically generated

Hình 18. Minh hoạt hệ thống tác nhân (Agent) tự động được hỗ trợ bởi LLM [59].

Một LLM agent hoạt động thông qua một quy trình có cấu trúc từ khi người dùng nhập thông tin đầu vào cho đến khi nhiệm vụ được hoàn thành. Đầu tiên, người dùng nhập một yêu cầu hoặc câu hỏi vào hệ thống. LLM agent sau đó sẽ phân tích và hiểu ngữ cảnh của yêu cầu này, sử dụng khả năng diễn giải của LLM để giải mã ý định của người dùng. Tiếp theo, LLM agent xây dựng một kế hoạch hành động, chia nhiệm vụ lớn thành các bước nhỏ hơn, có thể quản lý được [57].

Trong quá trình thực hiện, LLM agent có khả năng tương tác với các hệ thống khác (Ví dụ: Web search, Call API, v.v.) hoặc thu thập thêm dữ liệu cần thiết để hoàn thành nhiệm vụ. Nó thực hiện từng bước theo kế hoạch đã định. Cuối cùng, sau khi hoàn thành tất cả các bước, LLM agent tổng hợp kết quả và cung cấp câu trả lời hoặc thực hiện hành động cụ thể mà người dùng yêu cầu, đảm bảo rằng nhiệm vụ được hoàn thành một cách chính xác và hiệu quả [57].

#### Kỹ thuật Prompt

Mặc dù các mô hình LLM đã trở thành chuẩn mực mới trong lĩnh vực NLP, nhưng chúng đã được chứng minh là đôi khi tạo ra các câu trả lời không chính xác hoặc mơ hồ. Tuy nhiên, khả năng của nó có thể được giảm bớt thông qua một kỹ thuật gọi là kỹ thuật nhắc nhở (Prompt) [60].

Kỹ thuật Prompt có thể được định nghĩa là tạo ra đầu ra mong muốn từ LLM bằng cách đưa ra một bộ hướng dẫn rõ ràng về đầu ra mong muốn. Kỹ thuật Prompt giúp hiểu rõ hơn về khả năng và hạn chế của LLM. Phương pháp này không chỉ là thiết kế và phát triển các lời nhắc. Nó bao gồm nhiều kỹ năng và kỹ thuật cần thiết để tương tác, phát triển và kiểm tra LLM. Một lời nhắc đơn giản với một số hướng dẫn có thể giúp LLM có được đầu ra, tuy nhiên, để có được kết quả chất lượng tốt lời nhắc phải được xây dựng tốt. Một lời nhắc có thể chứa các hướng dẫn như câu hỏi và các chi tiết khác cung cấp nhiều ngữ cảnh hơn như đưa ra gợi ý về câu trả lời [60], [61], [62]. Hình 19 mô tả cách một lời nhắc hoạt động:



Hình 19. Mô họa việc các câu lệnh kiến thức được tạo ra [63].

Đầu tiên câu hỏi của người dùng nhập vào sẽ được mô hình LLM thực hiện phân tích các yêu cầu cụ thể để có được yêu cầu tri thức. Sau khi xác định được yêu cầu của câu hỏi, mô hình sẽ dựa vào các yêu cầu tri thức được học trước đó kết hợp với những yêu cầu tri thức từ câu hỏi đã được phân tích để trả lời câu hỏi [63]. Các lời nhắc này chính là nền tảng để LLM xây dựng câu hỏi và mở rộng cuộc trò chuyện [62].

### Ứng dụng LLM để xây dựng chatbot

Các mô hình LLM ngày nay đã được chứng minh là có hiệu quả vượt bậc trong các tác vụ NLP so với trước đây [48], [49], và do đó việc ứng dụng LLM vào nhiều lĩnh vực khác nhau đang ngày càng trở nên phổ biến [48]. Chính vì thế việc áp dụng một mô hình LLM phù hợp để xây dựng chatbot tuyển sinh là điều cần thiết và mang tính thực tiễn cao, giúp tối ưu hóa quy trình tuyển sinh và chăm sóc đối tượng khách hàng tiềm năng. Việc lựa chọn mô hình LLM phù hợp và huấn luyện trên lượng dữ liệu tuyển sinh chuyên biệt của Trường Đại học Đông Á sẽ tăng hiệu quả trả lời và cung cấp thông tin chính xác hơn. Ngoài ra tâm lý của giới trẻ ngày nay đề cao sự riêng tư, do đó việc tương tác với một chatbot sẽ không ảnh hưởng đến sự riêng tư của người dùng mà còn tạo tâm lý thoải moái nếu người dùng có nhiều câu hỏi và muốn tìm hiểu nhiều vấn đề. Hình 20 dưới đây là sơ đồ tổng quan cho việc sử dụng LLM để xây dựng chatbot.

A diagram of a model

Description automatically generated

Hình 20. Tổng quan hệ thống chatbot sử dụng LLM.

## Các công cụ hỗ trợ

### Ngôn ngữ lập trình Python

Python là một ngôn ngữ lập trình bậc cao, đa năng, được Guido van Rossum phát triển vào năm 1991. Ngôn ngữ này nổi bật nhờ cú pháp rõ ràng, tập trung vào tính hiệu quả và khả năng ứng dụng thực tế trong nhiều lĩnh vực [64], [65]. Python được sử dụng rộng rãi trong:

* Phát triển web: Cung cấp các framework như Django, Flask, hỗ trợ xây dựng ứng dụng từ cơ bản đến phức tạp.
* Khoa học dữ liệu và trí tuệ nhân tạo: Với các thư viện như NumPy, Pandas, TensorFlow, Python là công cụ quan trọng để phân tích dữ liệu và xây dựng hệ thống AI.
* Tự động hóa: Thích hợp để viết script xử lý các tác vụ lặp lại, từ quản lý tệp đến kiểm thử phần mềm.
* Hệ thống nhúng và IoT: Thường được dùng trên Raspberry Pi để phát triển các giải pháp IoT hoặc các hệ thống nhúng đơn giản.

Điểm mạnh của Python là khả năng tương thích đa nền tảng, hệ sinh thái thư viện phong phú và cộng đồng phát triển mạnh mẽ, giúp đáp ứng tốt các nhu cầu từ lập trình cơ bản đến ứng dụng chuyên sâu [64], [65]. Với sự phong phú đa dạng các thư viện hỗ trợ trong lĩnh vực AI và NLP, đây là lý do nghiên cứu này lựa chọn ngôn ngữ Python để thực hiện xây dựng chatbot.

### Các Framework sử dụng

#### Bộ công cụ thiết lập giao diện chatbot Chainlit

Bộ công cụ xây dựng website Chainlit là một framework mã nguồn mở được thiết kế để hỗ trợ các nhà phát triển trong việc xây dựng, thử nghiệm và triển khai giao diện người dùng (UI) tương tác cho các ứng dụng sử dụng LLM. Chainlit giúp đơn giản hóa quy trình phát triển, cung cấp các công cụ trực quan để tương tác với mô hình và nhanh chóng tạo các nguyên mẫu ứng dụng AI có giao diện thân thiện. Chainlit đặc biệt phù hợp cho việc phát triển chatbot, trợ lý ảo, và các ứng dụng hội thoại phức tạp, cho phép nhà phát triển dễ dàng theo dõi, kiểm tra logic, và tối ưu hóa phản hồi của mô hình AI [66], [67]. Việc sử dụng Chainlit trong dự án này không chỉ giúp tiết kiệm thời gian và giảm độ phức tạp trong việc xây dựng ứng dụng chatbot mà còn cung cấp một môi trường mạnh mẽ để kiểm tra và tối ưu hóa các mô hình ngôn ngữ lớn. Nhờ các tính năng trực quan và tích hợp tốt, Chainlit là một công cụ lý tưởng để phát triển các ứng dụng NLP hiện đại.

#### Bộ công cụ lưu trữ dữ liệu vector database Milvus

Milvus là công cụ mạnh mẽ dành cho các ứng dụng AI hiện đại yêu cầu xử lý dữ liệu vector lớn, với khả năng tìm kiếm và lưu trữ hiệu suất cao. Với tính năng dễ sử dụng, khả năng mở rộng, và tích hợp mạnh mẽ với các framework AI, Milvus giúp đơn giản hóa quy trình xây dựng các hệ thống thông minh như tìm kiếm ngữ nghĩa, phân loại và lưu trữ dữ liệu [68], [69]. Đối với việc xây dựng một hệ thống chatbot cần một cơ sở dữ liệu để lưu trữ lịch sử chatbot người dùng, việc lựa chọn Milvus trong dự án này là lựa chọn tối ưu khi được ứng dụng rộng rãi trong các hệ thống chatbot khác [69].

#### Công cụ quản trị cơ sở dữ liệu MongoDB

MongoDB [70] là một hệ quản trị cơ sở dữ liệu phi cấu trúc (Non Structured Query Languag - NoSQL) phổ biến, sử dụng cấu trúc dữ liệu dạng text thay vì bảng như các cơ sở dữ liệu quan hệ. Đặc điểm chính nổi bật của công cụ này nằm ở chỗ:

* Lưu trữ linh hoạt: Dữ liệu được lưu dưới dạng JSON (hoặc BSON), dễ dàng thay đổi cấu trúc.
* Khả năng mở rộng: Hỗ trợ phân mảnh (sharding) để xử lý dữ liệu lớn.
* Tốc độ cao: Tối ưu hóa cho truy vấn nhanh và xử lý dữ liệu phi cấu trúc.
* Dễ tích hợp: Thích hợp cho các ứng dụng web, mobile và big data nhờ tính linh hoạt và đa dạng API.

Với những ưu điểm trên, MongoDB thường được sử dụng trong các dự án cần lưu trữ dữ liệu phi cấu trúc hoặc bán cấu trúc, và phù hợp cho các dự án chatbot có quy mô vừa hoặc nhỏ.

#### Hệ thống bộ nhớ chatbot MemGPT

Mặc dù các mô hình LLM ngày nay đã có nhiều phát triển và cải tiến đáng kể, nhưng chúng vấn bị hạn chế bởi bộ nhớ ngắn hạn do cấu trúc của nó, và cần một hệ thống bộ nhớ mạnh mẽ để có thể ghi nhớ các ký ức và nhận thức được tình huống hiện tại. Hệ thống này phải có khả năng ghi lại và truy xuất các dấu vết bộ nhớ bằng ngữ cảnh, cho phép các tác nhân AI xử ly thông tin và đưa ra quyết định sáng suốt [71]. Một nhóm các nhà nghiên cứu tại Berkeley đã tạo ra memGPT, hoạt động như một công cụ quản lý bộ nhớ cho các mô hình ngôn ngữ lớn. Điều này cho phép truy xuất bộ nhớ dài hạn và khả năng viết cũng như bỏ qua giới hạn đầu vào của cửa sổ ngữ cảnh [72]. MemGPT, được giới thiệu trong bài báo “MemGPT: Hướng tới LLM như Hệ điều hành” [73], giúp các mô hình ngôn ngữ lớn (LLM) xử lý các cuộc hội thoại dài hơn bằng cách quản lý khéo léo các tầng bộ nhớ khác nhau. Nó biết khi nào cần lưu trữ thông tin quan trọng và truy xuất thông tin đó sau trong khi trò chuyện. Điều này giúp các mô hình AI có thể có các cuộc hội thoại kéo dài, cải thiện đáng kể tính hữu ích của chúng [74]. MemGPT được thiết kế để nâng cao hiệu suất của các Mô hình ngôn ngữ lớn (LLM) bằng cách giới thiệu một chương trình quản lý bộ nhớ tiên tiến hơn, giúp khắc phục những thách thức do các cửa sổ ngữ cảnh cố định đặt ra. Dưới đây là một số tính năng chính của MemGPT [75]:

* Quản lý bộ nhớ (Memory Management): MemGPT kết hợp hệ thống bộ nhớ theo tầng vào bộ xử lý LLM ngữ cảnh cố định, cấp cho nó khả năng quản lý bộ nhớ của riêng nó. Bằng cách xử lý thông minh các tầng bộ nhớ khác nhau, nó mở rộng ngữ cảnh có sẵn trong cửa sổ ngữ cảnh giới hạn của LLM, giải quyết vấn đề về cửa sổ ngữ cảnh bị hạn chế thường gặp trong các mô hình ngôn ngữ lớn.
* Quản lý ngữ cảnh ảo (Virtual Context Management): MemGPT giới thiệu một phương pháp được gọi là quản lý ngữ cảnh ảo. Đây là một tính năng chính hỗ trợ quản lý cửa sổ ngữ cảnh của LLM.
* Lấy cảm hứng từ hệ điều hành (Operating System-Inspired): Kiến trúc của MemGPT lấy cảm hứng từ các hệ điều hành truyền thống, đặc biệt là hệ thống bộ nhớ phân cấp giúp dữ liệu di chuyển giữa bộ nhớ nhanh và chậm dễ dàng hơn. Cách tiếp cận này cho phép quản lý tài nguyên bộ nhớ hiệu quả, tương tự như cách hệ điều hành cung cấp ảo giác về tài nguyên bộ nhớ lớn cho các ứng dụng thông qua phân trang bộ nhớ ảo.
* Xử lý gián đoạn (Interruption Handling): MemGPT sử dụng các ngắt để quản lý luồng điều khiển giữa nó và người dùng, đảm bảo tương tác mượt mà và quản lý bộ nhớ hiệu quả trong quá trình hoạt động.
* Bối cảnh đàm thoại mở rộng (Extended Conversational Context): Thông qua quản lý bộ nhớ hiệu quả, MemGPT tạo điều kiện cho bối cảnh đàm thoại mở rộng, cho phép tương tác dài hơn và mạch lạc hơn, vượt qua những hạn chế do cửa sổ ngữ cảnh có độ dài cố định áp đặt.

Về bản chất, MemGPT là một bước tiến đáng kể trong việc sử dụng Mô hình ngôn ngữ lớn, tạo ra con đường cho các tương tác hiệu quả và mở rộng hơn, giống với diễn ngôn của con người bằng cách quản lý tài nguyên bộ nhớ một cách thông minh [71].

# PHƯƠNG PHÁP ĐỀ XUẤT XÂY DỰNG CHATBOT

## Mô tả và phân tích bài toán

Mục tiêu của bài toán là xây dựng một hệ thống chatbot hỏi đáp tuyển sinh bằng cách sử dụng các công nghệ trí tuệ nhân tạo. Hệ thống chatbot phải có khả năng trả lời một cách tự động, liên tục và tự nhiên như con người để trả lời các câu hỏi tuyển sinh trong phạm vi tuyển sinh trường Đại học Đông Á nhằm hỗ trợ kịp thời nhu cầu tìm hiểu về chương trình học, ngành đào tạo và các chính sách học bổng… cho người dùng đang tìm hiểu về trường. Đối với các yêu cầu hoặc câu hỏi nằm ngoài phạm vi của người dùng, chatbot phải biết gửi lời xin lỗi và nhắc nhở người dùng những phạm vi mà chatbot có thể hỗ trợ, đồng thời có thể điều hướng hoặc cung cấp các thông tin cán bộ giảng viên trực thuộc để có thể giải đáp các yêu cầu nằm ngoài phạm vi của chatbot. Hình 21 là sơ đồ tổng quan bài toán xây dựng hệ thống chatbot hỏi đáp tuyển sinh.

A diagram of a chatbot

Description automatically generated

Hình 21. Tổng quan bài toán xây dựng hệ thống chatbot hỏi đáp tuyển sinh.

Đầu vào của bài toán này là trả lời các câu hỏi liên quan đến lĩnh vực tuyển sinh tại trường đại học Đông Á do người dùng nhập vào. Đầu ra của bài toán là câu trả lời của người dùng dạng văn bản, cung cấp đầy đủ thông tin chính xác cho câu trả lời, ngoài ra chatbot cũng gợi ý người dùng các câu hỏi hoặc thông tin liên quan đến câu hỏi trước mà người dùng có thể quan tâm, từ đó có thể tiếp tục dẫn dắt người dùng tìm hiểu sâu hơn vào các vấn đề mà người dùng quan tâm.

## Phương pháp đề xuất

Phương pháp đề xuất xây dựng chatbot tuyển sinh sử dụng mô hình ngôn ngữ lớn (LLM) được thiết kế nhằm tối ưu hóa quy trình tư vấn tuyển sinh, đảm bảo cung cấp thông tin nhanh chóng, chính xác và thân thiện với người dùng. Hình 20 là sơ đồ tổng quan phương pháp đề xuất xây dựng chatbot sử dụng LLM cho dự án này. Quy trình bao gồm từ việc thu thập và xử lý dữ liệu đến vận hành chatbot, sử dụng các tác nhân thông minh và đánh giá hiệu quả.

Hệ thống được chia thành 2 giai đoạn chính, cụ thể giải đoạn thứ nhất chịu trách nhiệm tiền xử lý và chuẩn bị hệ thống, các bước bao gồm (1, 2, 3, 4, 5.1, 5.2). Tại giai đoạn này chúng tôi tiến hành thử nghiệm và đánh giá các thành phần quan trọng ảnh hưởng đến hiệu suất, bao gồm đánh giá hiệu quả embedding, chất lượng trả lời câu hỏi của mô hình LLM, và thử nghiệm và đánh giá các thuật toán retrieval và độ chính xác. Giai đoạn thứ 2 (bao gồm các bước còn lại) chúng tôi tiến hành triển khai hệ thống chatbot theo quy trình thiết kế theo phương pháp đề xuất.



Hình 22. Phương pháp đề xuất xây dựng chatbot hỏi đáp tuyển sinh sử dụng LLM.

Trong sơ đồ hình 22, mỗi thành phần đóng vai trò cụ thể trong việc xử lý truy vấn và cải thiện trải nghiệm người dùng. Cụ thể mỗi bước chịu trách nhiệm và xử lý công việc như sau:

* (1): Bước đầu tiên trong phương pháp đề xuất của chúng tôi là thu thập và tạo kịch bản dữ liệu. Dữ liệu tuyển sinh được thu thập trên website chính thức của Trường Đại học Đông Á, đồng thời chúng tôi cũng tạo ra các kịch bản câu hỏi có thể xảy ra dựa trên tình huồng tuyển sinh thực thực tế được lấy từ các trang facebook tuyển sinh của trường năm 2024..
* (2): Bước thứ hai là thực tiện tiền xử lý dữ liệu, trong bước này thực hiện làm sạch dữ liệu, chuẩn hóa dữ liệu và gắn nhãn cho dữ liệu tùy theo từng danh mục hoặc pham vi chủ đề câu hỏi. Đồng thời để tăng thêm tính chính xác cho câu trả lời của chatbot sau này, chúng tôi bổ sung sung các trường hợp từ viết tắt phổ biến,
* (3): Bước thứ ba là thực hiện phân đoạn dữ liệu, các mẫu có độ dài vượt quá khả năng xử lý của các mô hình LLM sẽ được cắt ngắn và sắp xếp lại để có thể dễ dàng xử lý và phân tích. Việc chia nhỏ giúp cải thiện tìm kiếm thông tin và giảm tải cho các mô hình xử lý dữ liệu lớn.
* (4): Bước thứ tư sau khi hoàn thành tất cả các quy trình chuẩn hóa, dữ liệu sẽ được đưa vào mô hình embedding thể thực hiện embedding dữ liệu thành các vector. Việc embedding dữ liệu mục đích là để mô có thể tính toán và truy xuất ngược lại sau này. Ngoài ra, dữ liệu được chuyển đổi thành vector thông qua mô hình embedding này sẽ giúp biểu diễn ý nghĩa ngữ nghĩa của dữ liệu. Mô hình embedding đảm bảo rằng các đoạn dữ liệu tương tự sẽ gần nhau trong không gian vector.
* (5.1): Sau đó lưu trữ các vector embedded này vào một vector dabase. Điều này giúp tối ưu hóa quá trình truy xuất thông tin trong các bước sau truy xuất sau này. Ngoài ra, sau khi thực hiện embedding và lưu vector embedding và cơ sở dữ liệu, cũng cần thực hiện đánh giá kết quả retrieval bằng các phương pháp retrieval (5.2), điều này giúp đưa ra cái nhìn tổng quan về chất lượng dữ liệu, hiệu quả embedding và tốc độ truy xuất của vector databse. Đồng thời chắc chắn rằng các truy vấn nhập vào có thể tìm được mẫu trả lời đúng một cách hiệu quả.
* (6): Trong quá trình giao tiếp giữa người dùng và hệ thống chatbot, đầu vào của người dùng, đầu vào cũng phải được embedding để hệ thống có thể xử lý. Lúc này câu truy vấn đã được embedding này sẽ được dùng để truy xuất thông tin từ cơ sở dữ liệu vector database (7). Hệ thống RAG tìm kiếm thông tin phù hợp từ vector database bằng cách so sánh vector truy vấn của người dùng với các vector trong cơ sở dữ liệu (8). Từ đó các kết quả tương tự nhất sẽ được lấy ra (9) cho mục đích tạo câu trả lời.
* (10.1): Tiếp theo là thực hiện tạo ra các prompt tối ưu để định hướng đầu vào cho mô hình ngôn ngữ lớn (LLM). Các lời nhắc này được thiết kế sao cho LLM hiểu rõ yêu cầu của ngữ cảnh. Cùng với đó là việc thiết kế và tạo một hệ thống quản lý các agent. Agent giúp đảm bảo rằng mỗi yêu cầu của người dùng được chuyển đến công cụ xử lý phù hợp. Điều này bao gồm việc định tuyến yêu cầu tới mô hình LLM. Trên thực tế, sau thi hệ thông RAG thực hiện truy xuất thông tin từ câu hỏi người dùng, các thông tin sẽ được truyền qua trình tạo prompt (10.1) và trình quản lý agent (10.2). Thông tin prompt cũng được trình quản lý agent sử dụng để xử lý và quyết định sẽ dùng công cụ nào để gọi hàm (11).
* (12): Khi đã có đầy đủ thông tin, trình quản lý tác nhân sẽ thực hiện gọi tác nhân (agent call) để phân luồng truy vấn đến tác nhân phù hợp.
* (13.1, 13.2, 13.3): Đối với agent hỏi đáp, prompt sẽ được đưa vào mô hình LLM để thực hiện trả lời phản hồi dựa trên các thông tin tìm kiếm được, điều này bao gồm các thông tin lịch sử trò chuyện được tóm tắt trước đó. Sau đó câu trả lời của mô hình sẽ được đưa đến trình hiển thị phản hồi.
* (14.1, 14.2, 14.3, 14.4): Nếu agent ngoài phạm vi trả lời của chatbot, prompt tương ứng sẽ được đưa vào LLM để sinh ra phản hồi phù hợp, sau đó điều hướng người dùng tới cán bộ tuyển sinh nếu người dùng đồng ý. Tất cả những dữ liệu phản hồi này được đưa tới trình hiển thị phản hồi.
* (15, 16): Cuối cùng các đầu ra của các tác nhân này sẽ được truyền tới một bộ phân tích cú pháp để thực hiện phản hồi người dùng cuối cùng (15). Trong suốt quá trình trò chuyện, lịch sử sẽ được ghi lại để cải thiện phản hồi của chatbot và giữ lại được ngữ cảnh và tình huống đang giao tiếp. Đồng thời dữ liệu quan trọng được ghi vào cơ sở dữ liệu người dùng cho các cuộc trò chuyện sau này.
* (17,1 17.2): Trong toàn bộ quá trình xử lý hỏi đáp của hệ thống chatbot, các thông tin người dùng nhập vào sẽ được thu thập tự động bằng cách sử dụng LLM để trích xuất, dựa trên các yêu cầu truy vấn này, hệ thống sẽ biết được các thông tin mong muốn của người dùng (về ngành học, quan tâm tới học phí, các chính sách học bổng…), nhu cầu tìm hiểu về các lĩnh vực quan tâm khác. Những thông tin này sẽ được đưa vào LLM để hỗ trợ cho quá trình phản hồi tới người dùng, khi chatbot hiểu các yêu cầu và lĩnh vực mà người dùng quan tâm, chatbot sẽ trả lời các câu hỏi và gợi ý tìm hiểu các chủ đề xoay quanh các khía cạnh mà người dùng qua tâm. Từ đó giúp duy trì cuộc hội thoại lâu dài hơn và tăng thêm sự chủ động trong việc tìm hiểu nhiều hơn về trường cho người dùng. Các thông tin này sẽ được lưu trữ trong cơ sở dữ liệu MongoDB.

Trên đây là chi tiết toàn bộ quy trình trong hệ thống chatbot tuyển sinh. Trong quá trình trên đã ứng dụng các phương pháp hiện đại trong bối cảnh hiện tại. Đây là một quy trình chặt chẽ và có tính hệ thống cao. Bằng cách kết hợp công nghệ hiện đại và quy trình rõ ràng, phương pháp này giúp chatbot hoạt động hiệu quả, đáp ứng tốt nhu cầu người dùng và dễ dàng mở rộng trong tương lai và tích hợp với các hệ thống khác.

## Thiết kế hệ thống thông tin

Trong phần này, chúng tôi sẽ trình bày thiết kế hệ thống thông tin cho ứng dụng chatbot hỗ trợ hỏi đáp tuyển sinh. Đồ án tập trung vào việc phát triển một hệ thống chatbot có khả năng trả lời chính xác các câu hỏi liên quan đến tuyển sinh, bằng cách ứng dụng công nghệ trí tuệ nhân tạo và phương pháp thiết kế hiện đại. Mục tiêu là tạo ra một chatbot thông minh, không chỉ cung cấp thông tin đúng và kịp thời, mà còn mang lại trải nghiệm tương tác tự nhiên cho người dùng dựa trên dữ liệu đã được thu thập và tổ chức. Hệ thống được thiết kế với chức năng cốt lõi là hỏi đáp, nhưng để đạt được hiệu quả cao nhất, quy trình xây dựng đòi hỏi sự chuẩn bị kỹ lưỡng, thử nghiệm chặt chẽ, và đánh giá toàn diện. Dù chỉ đảm nhận một chức năng chính, hệ thống này là kết quả của việc tích hợp những nghiên cứu chuyên sâu và các công nghệ tiên tiến nhất hiện nay, nhằm đáp ứng nhu cầu thực tiễn trong lĩnh vực giáo dục.

### Chức năng hỏi đáp người dùng

Trọng tâm của chức năng hỏi đáp người dùng là cho phép người dùng đặt câu hỏi liên quan đến tuyển sinh và hệ thống chatbot sẽ dựa vào các thông tin được cung cấp để trả lời câu hỏi cho người dùng. Khi bắt đầu mỗi phiên trò chuyện, hệ thống sẽ yêu cầu người dùng nhập vào các thông tin cá nhân và lưu vào cơ sở dữ liệu MongoDB. Hình 23 dưới đây là sơ đồ mô tả chức năng hỏi đáp của người dùng.

A diagram of a flowchart

Description automatically generated

Hình 23. Biểu đồ chức năng hỏi đáp người dùng.

Đầu tiên khi bắt đầu một cuộc trò chuyện mới, hệ thống sẽ yêu cầu người dùng nhập vào thông tin cá nhân như họ tên, số điện thoại và địa chỉ email. Đầu vào này là một văn bản. Hệ thống sẽ kiểm tra những thông tin này bằng cách sử dụng LLM để kiếm tra người dùng đã nhập tên và các thông tin khác hay chưa. Nếu vẫn chưa thì chatbot sẽ tiếp tục gửi phản hồi lời mời nhập thông tin cá nhân để có thể bắt đầu đoạn chat. Nếu người dùng nhập vào thông tin thì những thông tin này sẽ được hệ thống trích xuất một cách tự động và đưa vào cơ sở dữ liệu MongoDB. Lúc này người dùng có thể nhập bất kỳ câu hỏi nào để hệ thống chatbot xử lý và phản hồi. Vòng lặp hỏi và đáp của người dùng và chatbot sẽ kết thúc nếu người dùng yêu cầu hoặc thoát hệ thống. Các lịch sử trò chuyện sẽ không được lưu vào hệ thống mà sẽ bị xóa đi sau khi phiên trò chuyện kết thúc.

### Biểu đồ use case chức năng hỏi đáp người dùng

Biểu đồ use case chức năng hỏi đáp (hình 24) gồm có các thành phần:

* Tác nhân: Ở đây là người dùng.
* Mô tả: Thực hiện hỏi đáp tư vấn.
* Điều kiện: Người dùng đã nhập thông tin cá nhân vào hệ thống chatbot.
* Luồng xử lý: Khi người dùng nhập câu hỏi cần tư vấn vào chatbox, hệ thống sẽ xem xét yêu cầu của câu hỏi/yêu cầu của người dùng có nằm trong phạm vi hỏi đáp của chatbot hay không. Nếu có thì sẽ trả lời câu hỏi theo dữ liệu được truy vấn từ vector database. Nếu không thì hệ thống chatbot sẽ xin lỗi người dùng và nhắc nhở người dùng rằng câu hỏi mới được nhập và không nằm trong phạm vi xử lý của chatbot.

A diagram of a flowchart

Description automatically generated

Hình 24. Biểu đồ use case chức năng hỏi đáp tuyển sinh.

### Biểu đồ tuần tự chức năng hỏi đáp người dùng

Biểu đồ tuần tự chức năng hỏi đáp người dùng biểu diễn tuần tự quá trình xử lý và phản hồi của hệ thống chatbot và sự tương tác của người dùng. Hệ thống gồm có 3 thành phần chính là người dùng, trình quản lý agent chatbot và các hàm xử lý, tool hành động. Hình 25 là biểu đồ tuần tự mô tả chức năng hỏi đáp người dùng.

Đầu tiên khi bắt đầu một cuộc trò chuyện mới, trình quản lý agent sẽ truyền đến người dùng và yêu cầu nhập vào thông tin cá nhân của người dùng (1) bao gồm họ và tên, số điện thoại và địa chỉ email. Người dùng nhập các thông tin vào hệ thống (2) sẽ được hệ thống kiểm tra độ phù hợp và luu vào hệ thống cơ sở dữ liệu người dùng. Sau đó chatbot sẽ thông báo với người dùng là đã ghi nhớ thông tin và bắt đầu cuộc trò chuyện (3).

Bắt đầu từ bước (4) là quá trình hoạt động chủ yếu của hệ thống chatbot, tại đây hệ thống hoạt động như một vòng lặp để duy trì cuộc hội thoại. Người dùng sẽ nhập vào các câu hỏi liên quan đến tư vấn tuyển sinh. Các câu hỏi này sẽ được trình quản lý agent nhận diện và điều hành (5), bằng cách sử dụng các công cụ và thành phần xử lý, kết quả cuối cùng nhận được là một câu phản hồi cho câu truy vấn đầu vào (6), lúc này hệ thống chatbot app sẽ hiển thị cho người dùng. Nếu người dùng tiếp tục nhập vào câu hỏi, hệ thống sẽ tiếp tục xử lý theo quy trình như bắt đầu từ bước (4).



Hình 25. Biểu đồ tuần tự chức năng hỏi đáp người dùng.

### Cơ sở dữ liệu người dùng

Trong hệ thống chatbot tuyển sinh, chúng tôi xây dựng một cơ sở dữ liệu đơn giản để lưu thông tin các người dùng hệ thống chatbot. Dưới đây là bảng mô tả bảng cơ sở dữ liệu người dùng (bảng 3):

Bảng 3. Bảng mô tả cơ sở dữ liệu người dùng trong hệ thống chatbot.

|  |  |  |
| --- | --- | --- |
| **Tên trường** | **Kiểu dữ liệu** | **Mô tả** |
| ID | String | ID của người dùng |
| NAME | String | Họ và tên người dùng |
| SDT | String | Số điện thoại người dùng |
| EMAIL | String | Địa chỉ email của người dùng |
| USER DESIRE | String | Lưu trữ thông tin thu thập từ người dùng |

## Thu thập và tạo dữ liệu

Dữ liệu tuyển sinh qua các năm thường thay đổi tùy theo cơ chế của từng trường. Việc thu thập dữ liệu chính xác là rất quan trọng, giúp mô hình phân tích đưa ra câu trả lời chính xác hơn thay vì chỉ mang tính chung chung. Trong nghiên cứu này, chúng tôi tập trung thu thập dữ liệu tuyển sinh của Trường Đại học Đông Á qua các năm, bao gồm thông tin chi tiết từng ngành học. Đồng thời thực hiện tạo dữ liệu kịch bản cho những câu hỏi mà sinh viên có thể sẽ quan tâm.

A diagram of a cylinder

Description automatically generated

Hình 26. Tạo dữ liệu tuyển sinh bằng cách thu thập và tạo kịch bản.

Cụ thể, đối với việc thu thập dữ liệu, chúng tôi nhắm đến dữ liệu bao gồm chỉ tiêu tuyển sinh, điểm chuẩn, tổ hợp xét tuyển, thông tin khối ngành xét tuyển và các phương thức xét tuyển như học bạ, điểm thi tốt nghiệp THPT, hoặc đánh giá năng lực. Ngoài ra những thông tin mà phụ huynh và học sinh thường quan tâm là chính sách học bổng và học phí, chúng tôi thu thập thông tin chính sách học bổng các năm, các thông tin và thông báo cũng được xem xét để xây dựng dữ liệu tham khảo đối chiếu. Ngoài ra, chương trình đào tạo ngành học và cơ hội việc làm cũng được quan tâm lớn. Chúng tôi thu thập các module chương trình đào tạo, nêu rõ các môn học cho mỗi ngành qua từng module, đồng thời thu thập những thông tin tổng quan về ngành để sinh viên có thể nắm rõ về ngành cụ thể mà sinh viên có sự quan tâm. Việc thu thập thông tin được lấy từ trang website chính thức của Trường Đại học Đông Á [76] để đảm bảo tính chính xác nhất. Những thông tin này không chỉ giúp chatbot trả lời đúng trọng tâm câu hỏi mong muốn mà còn đáp ứng vấn đề thời gian.

Ngoài việc thu thập dữ liệu trên trang website chính thức của trường, chúng tôi nhận thấy rằng sinh viên có xu hướng tìm hiểu các vấn đề liên quan đến việc nạp hồ sơ, cách đóng học phí, nhu cầu tìm phòng trọ cũng như mức sống tại thành phố Đà Nẵng. Vì thế chúng tôi quyết định sẽ tạo thêm dữ liệu kịch bản thiết thực nhất trong vai trò là một sinh viên mới muốn tìm hiểu về ngành học và trường. Chúng tôi xem xét lựa chọn và thu thập các câu hỏi của sinh viên trong nhóm Facebook tuyển sinh chính thức của Trường Đại học Đông Á năm 2024 [77] cùng với hệ thống website của trường tại [76] để tổng hợp dữ liệu bao quát nhất.

Nhìn chung trong cả hai trường hợp, quy tắc thu thập dữ theo một khuôn mẫu cố định bao gồm chủ đề quan tâm – nội dung bao quát hoặc câu hỏi - trả lời. Đối với các mẫu có số lượng từ lớn hoặc quá dài, chúng tôi thực hiện chia nhỏ dữ liệu theo từng đoạn theo phương pháp chunking. Sau khi thu thập và tạo dữ liệu, dữ liệu của chúng tôi có 736 mẫu cho tất cả các ngành học và các lĩnh vực dữ liệu liên quan đến tuyển sinh. Bảng 4 dưới đây là minh họa một vài ví dụ về mẫu dữ liệu được thu thập:

Bảng 4. Minh hoạ một số mẫu dữ liệu được thu thập và tạo.

|  |  |
| --- | --- |
| **Câu hỏi** | **Nội dung phản hồi thu thập được** |
| Khi học ngành thiết kế thời trang, sinh viên có cơ hội tham gia các dự án thực tế trong quá trình học không? | Có, nhà trường thường xuyên tổ chức các đồ án thực tế như thiết kế BST (bộ sưu tập) theo chủ đề và trình diễn thời trang. Những dự án này giúp sinh viên trải nghiệm từ việc lên ý tưởng, thiết kế, đến trình diễn sản phẩm. Ngoài ra, sinh viên còn có cơ hội tham gia các sự kiện thời trang của trường hoặc các cuộc thi thiết kế quốc gia và quốc tế, giúp phát triển kỹ năng và nâng cao hồ sơ cá nhân. |
| Vị trí việc làm của ngành Công nghệ thực phẩm? | Là ngành học khá phổ biến nhưng ít ai biết rằng cơ hội việc làm đối với ngành học này rất cao. Ứng dụng của ngành đa dạng và có tính thiết yếu trong cuộc sống. Sinh viên ngay khi ra trường có thể đảm nhận nhiều vị trí sau: - Kỹ sư chế biến, đảm bảo chất lượng, kiểm tra chất lượng sản phẩm, an toàn thực phẩm.  - Quản lý sản xuất, khởi nghiệp trong lĩnh vực thực phẩm, nghiên cứu R&D, nhân viên kiểm định chất lượng, nhân viên kiểm soát chất lượng nguyên liệu.  - Chuyên viên nghiên cứu và phát triển sản phẩm, chuyên viên dinh dưỡng, kỹ thuật viên sản xuất, nhân viên phòng thí nghiệm…  - Và nhiều công việc khác tại nhà máy, công ty thực phẩm như: công ty bia rượu, nước giải khát, công ty chế biến sữa và các thực phẩm từ sữa, công ty chế biến gia vị… |
| Học ngành Quản trị khách sạn là học về cái gì? | Quản trị khách sạn là ngành học dịch vụ, đào tạo chuyên sâu về các hình thức giải trí, nghỉ dưỡng, vui chơi, dịch vụ ăn uống. Nói một cách dễ hiểu, ngành Quản trị khách sạn quản lý, tổ chức, điều phối các hoạt động của khách sạn sao cho hiệu quả và hợp lý nhất. Bên cạnh đó là cách quản lý nhân sự về cả khối lượng công việc cần thực hiện đến yếu tố con người để đảm bảo hiệu quả cho doanh nghiệp, tập đoàn. |
| Học bạ THPT với giấy chứng nhận TN mình công chứng ở đâu ạ? | Em có thể phôtô học bạ và giấy chứng nhận tốt nghiệp đến trường THPT em đã học (nơi cấp) hoặc đến Phường/Xã (Bộ phận Tư pháp), các văn phòng công chứng để chứng thực bạn nhé! |
| Hồ sơ xét tuyển gồm những thủ tục nào? | Hồ sơ xét tuyển đại học Đông Á bao gồm: 1. Đối với thí sinh xét tuyển theo học bạ THPT - Đơn đăng ký xét tuyển (theo mẫu của nhà trường),  - Học bạ THPT/THPT(GDTX) (bản sao có chứng thực); - Bằng tốt nghiệp THPT/THPT(GDTX) (bản sao có chứng thực) hoặc chứng nhận tạm thời tốt nghiệp THPT - Lệ phí xét tuyển: 30.000 đồng. 2. Đối với thí sinh xét tuyển theo kết quả thi THPT QG - Giấy chứng nhận kết quả thi THPT Quốc gia (bản sao có chứng thực); - Giấy chứng nhận tạm thời tốt nghiệp THPT (bản sao có chứng thực); - Lệ phí xét tuyển: 30.000 đồng. |

## Tiền xử lý và phân đoạn dữ liệu

Sau khi hoàn thành thu thập dữ liệu tuyển sinh từ các nguồn tin cậy từ [76], [77], cần kiểm tra các mẫu dữ liệu để đảm bảo tính chính xác và đúng đắn. Hình 27 là các nhiệm vụ được thực hiện trong phần này.

A diagram of a diagram

Description automatically generated

Hình 27. Quy trình tiền xử lý và phân đoạn bộ dữ liệu.

Đầu tiên chúng tôi thực hiện loại bỏ các thông tin có thể gây nhiễu hoặc không cần thiết, làm sạch các lỗi chính tả hoặc các ký tự đặc biệt. Để đảm bảo chatbot có khả năng hiểu tốt hơn về ngữ cảnh, chúng tôi tiến hành xử lý các trường hợp có chữ viết tắt. Nắm bắt xu hướng học sinh thường có hay có thói quen viết tắt các từ khóa thông dụng, chính vì thế chúng tôi thực hiện tái tạo thêm mẫu dữ liệu với các câu trả lời có chữ viết tắt để tăng thêm hiệu quả trả lời của chatbot. Cụ thể chúng tôi tái tạo các câu trả lời với các từ khóa thông dụng như “Công nghệ thông tin” thành các từ khóa “CNTT” hoặc “cntt”, một ví dụ khác như từ khóa “Quản trị kinh doanh” sẽ được tái tạo thêm các mẫu như là “QTKD” hoặc “qtkd” và các từ khóa các ngành các nữa. Sau đó đối với mỗi mẫu dữ liệu, chúng tôi thực hiện kiểm tra lại văn phong câu trả lời và hiệu chỉnh để câu trả lời trôi chảy tự nhiên hơn. Cuối cùng là thực hiện gắn nhãn dữ liệu có cùng chủ đề với nhau, các câu hỏi có tính liên quan này sẽ được đề xuất trong suốt quá trình chatbot, điều này làm tăng kích thích người dùng tìm hiểu sâu hơn về một chủ đề người dùng quan tâm.

Sau khi làm sạch dữ liệu, chúng tôi tiến hành kiểm tra và xử lý các mẫu dữ liệu có độ dài vượt quá giới hạn mà các mô hình embedding và LLM có thể xử lý. Để khắc phục vấn đề này, các mẫu trả lời được phân đoạn lại nhằm đảm bảo phù hợp với kích thước yêu cầu, đồng thời hiệu chỉnh văn phong để dữ liệu trở nên mạch lạc và tự nhiên hơn. Đối với các câu trả lời dài và phức tạp, chúng tôi phân chia thành các đoạn ngắn hơn nhưng vẫn đảm bảo tính liên kết, đầy đủ ý nghĩa và giữ nguyên bối cảnh cần thiết.

Trong bước tiếp theo, dữ liệu đã qua xử lý được chuyển đổi thành định dạng embedding thông qua các mô hình tiên tiến. Đây là giai đoạn quan trọng nhằm mã hóa ý nghĩa ngữ nghĩa của dữ liệu, tạo điều kiện cho việc truy xuất thông tin hiệu quả và đáp ứng nhanh chóng. Việc xây dựng embedding không chỉ tối ưu hóa dung lượng lưu trữ mà còn cải thiện độ chính xác khi chatbot thực hiện tìm kiếm và đưa ra câu trả lời.

## Đánh giá các mô hình Embedding đề xuất

Hình 28 dưới đây mô tả quá trình đánh giá và so sánh hiệu suất của các mô hình embedding.

A diagram of a model embedding

Description automatically generated

Hình 28. Tổng quan quy trình đánh giá hiệu suất các mô hình embedding.

Việc so sánh nhiều mô hình embedding giúp mang lại góc nhìn toàn diện và tối ưu hóa việc lựa chọn mô hình phù hợp. Đầu tiên, việc đưa ra nhiều mô hình và đánh giá giúp hiểu rõ các ưu điểm, hạn chế, cũng như hiệu suất thực tế của từng mô hình trên các tiêu chí cụ thể như tốc độ, độ chính xác, và tài nguyên yêu cầu. Điều này đặc biệt quan trọng khi ứng dụng các mô hình trong bối cảnh cụ thể như hệ thống chatbot, nơi hiệu suất thực tế có thể thay đổi tùy thuộc vào dữ liệu hoặc nhu cầu. Hơn thế nữa, so sánh nhiều mô hình mang lại lợi ích về mặt kinh tế và chiến lược giúp chọn được giải pháp tối ưu về chi phí và hiệu năng, đồng thời giảm thiểu rủi ro khi triển khai​ [78], [79].

Dưới đây là 4 mô hình được cân nhắc lựa chọn chọn cho nhiệm vụ embedding dữ liệu, bao gồm:

* Mô hình *Sentence Transformers - all-MiniLM-L6-v2*: Được phát triển bởi nhóm nghiên cứu Sentence Transformers, dẫn đầu bởi Nils Reimers, trong cộng đồng Hugging Face. Được công bố vào năm 2020-2021 như một phần của bộ mô hình sử dụng MiniLM của Microsoft. Mô hình này được tối ưu hóa để mã hóa văn bản thành vector 384 chiều, hỗ trợ các tác vụ như tìm kiếm ngữ nghĩa (semantic search), cụm văn bản (clustering) và đo độ tương tự (sentence similarity) [80], [81]. Nó nổi bật vì sự cân bằng giữa hiệu suất và tốc độ xử lý. Dữ liệu huấn luyện trên hơn 1 tỷ cặp câu sử dụng kỹ thuật học tương phản (contrastive learning), kết hợp dữ liệu từ nhiều nguồn như Reddit, WikiAnswers, và Stack Exchange [81]. Mô hình này miễn phí và mã nguồn mở, đây là một mô hình nhẹ và rất phổ biến trong các ứng dụng tìm kiếm ngữ nghĩa và embedding văn bản
* Mô hình *Alibaba-NLP/gte-multilingual-base*: Ra đời bởi Alibaba NLP Team. Mô hình này được công bố vào khoảng năm 2022. Mục tiêu của nó là hướng tới việc xử lý ngôn ngữ đa ngữ (multilingual), mô hình này hỗ trợ embedding cho hơn 100 ngôn ngữ. Đây là một giải pháp mạnh mẽ cho các hệ thống toàn cầu, yêu cầu hỗ trợ nhiều ngôn ngữ trong các tác vụ như tìm kiếm đa ngữ hoặc dịch thuật. Ưu điểm tích hợp dễ dàng vào các pipeline đa ngữ, với hiệu suất cao trên nhiều ngữ cảnh [82]. Mô hình này miễn phí và mã nguồn mở, được cung cấp bởi nhóm Alibaba NLP. Mô hình này có thể tải về từ Hugging Face [83].
* Mô hình *BAAI/bge-m3*: Ra đời bởi Beijing Academy of Artificial Intelligence (BAAI), Trung Quốc. Được giới thiệu trong năm 2021. Mục tiêu là được thiết kế đặc biệt cho các tác vụ tìm kiếm văn bản và hiểu ngữ nghĩa sâu hơn trong văn bản lớn. Mô hình này hướng tới cải thiện độ chính xác của các tác vụ embedding thông qua việc sử dụng kỹ thuật học sâu (deep learning) với kiến trúc hiện đại. Dữ liệu huấn luyện được tận dụng kho dữ liệu phong phú từ các nguồn Trung Quốc và quốc tế, giúp hỗ trợ tốt cả ứng dụng đơn ngữ và đa ngữ. Mô hình này miễn phí do Beijing Academy of Artificial Intelligence phát hành trên nền tảng Hugging Face. Nó được sử dụng rộng rãi trong các tác vụ tìm kiếm và xử lý ngữ nghĩa [84], [85]​.
* Mô hình *nomic-ai/nomic-embed-text-v1.5*: Được hát hành vào năm 2024, mô hình này được đào tạo trình embedding bằng cách sử dụng một pipline đào tạo nhiều giai đoạn. Bắt đầu từ mô hình BERT ngữ cảnh dàitrên lượng dữ liệu lớn được tạo từ các cặp văn bản có liên quan yếu, các cặp tiêu đề - nội dung. Trong giai đoạn tinh chỉnh, các tập dữ liệu được gắn nhãn chất lượng cao hơn như các truy vấn tìm kiếm và trả lời từ các tìm kiếm web được tận dụng. Mô hình này hiện đang mở cho mục đích nghiên cứu và có thể tham khảo từ Hugging Face [86], [87].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bảng 5. So sánh và đánh giá hiệu suất của các mô hình embedding đề xuất. | **4Trung bình thời gian embed câu hỏi** | 0,005 | 0,010 | 0,021 | 0,012 |
| **Trung bình độ dài câu hỏi** | 14,681 | 166,484 | 9,734 | 14,681 |
| **Trung bình thời gian embed câu** | 0,013 | 0,056 | 0,024 | 0,014 |
| **Trung bình độ dài câu** | 389,149 | 1.744,195 | 373,787 | 398,149 |
| **Không gian GPU** | 0,109 | 1,466 | 2,725 | 0,656 |
| **Số lượng tham số** | 22.713.216 | 305.368.320 | 567.754.762 | 136.731.648 |
| **Độ dài bối cảnh** | 256 | 8.192 | 8.192 | 8.192 |
| **Số chiều** | 384 | 768 | 1.024 | 1.024 |
| **Top 10** | 0,468 | 0,872 | 0,851 | 0,574 |
| **Top 5** | 0,351 | 0,840 | 0,798 | 0,436 |
| **Top 3** | 0,234 | 0,755 | 0,787 | 0,351 |
| **Top 1** | 0,096 | 0,500 | 0,521 | 0,213 |
| **Tên mô hình** | Sentence-transformer/all-MiniLM-L6-v2 | Alibaba-NLP/gte-multilingual-base | BAAI/bge-m3 | nomic-ai/nomic-embed-text-v1.5 |

Trong bảng 3 so sánh hiệu suất tăng cường truy xuất tạo sinh (RAG) cho một số mô hình LLM nổi bật miễn phí. Trong đó, chỉ số Top (Top-K Accuracy) biểu thị tỷ lệ phần trăm các câu trả lời đúng xuất thành công xuất hiện trong danh sách kết quả đứng đầu (tương ứng với 1, 3, 5 hoặc 10 kết quả đầu tiên). Các giá trị này đánh giá khả năng truy xuất chính xác của mô hình. Mô hình có giá trị cao hơn có nghĩa là tốt hơn trong việc xếp hạng các kết quả phù hợp với câu hỏi truy vấn [78], [79].

Số chiều (Dimensions) là kích thước của vector embedding được tạo bởi mô hình. Ví dụ: 384, 768, hoặc 1.024 chiều. Vector có số chiều cao hơn thường nắm bắt được nhiều đặc trưng ngữ nghĩa hơn, nhưng cũng đòi hỏi nhiều bộ nhớ và tính toán hơn khi thực hiện so sánh vector [78], [79].

Độ dài bối cảnh (Context Length) là số token tối đa mà mô hình có thể xử lý trong một lần (context window). Mô hình có độ dài bối cảnh cao hơn có thể xử lý các câu dài hoặc đoạn văn lớn hơn, phù hợp cho các ứng dụng như tìm kiếm hoặc phân tích văn bản dài [78], [79].

Số lượng tham số (Parameters) là tổng số tham số trong mô hình, thường biểu thị kích thước của mô hình. Mô hình lớn hơn (số tham số cao) thường mạnh mẽ hơn trong việc nắm bắt các mối quan hệ ngữ nghĩa phức tạp, nhưng tốn nhiều tài nguyên tính toán hơn [78], [79].

Không gian GPU là lượng bố nhớ GPU chiếm dụng khi triển khai truy vấn.

Trung bình độ dài câu biểu thị độ dài trung bình của các câu hoặc đoạn văn được embed. Thường được tính bằng số token hoặc ký tự. Độ dài lớn hơn có thể yêu cầu mô hình thực hiện nhiều bước tính toán hơn, ảnh hưởng đến tốc độ và hiệu suất [78], [79].

Trung bình thời gian embed câu (Seconds) là thời gian trung bình để mã hóa một đoạn văn bản (hoặc câu) thành vector embedding. Giá trị này quan trọng trong các ứng dụng xử lý dữ liệu lớn, vì nó ảnh hưởng đến hiệu suất tổng thể khi xử lý hàng loạt dữ liệu [78], [79].

Trung bình độ dài câu hỏi biểu thị độ dài trung bình của các câu hỏi (số lượng token hoặc ký tự) được sử dụng để kiểm tra mô hình. Độ dài này ảnh hưởng đến tốc độ xử lý vì câu hỏi dài hơn yêu cầu nhiều tính toán hơn [78], [79].

Trung bình thời gian embedding câu hỏi (Seconds) là thời gian trung bình (tính bằng giây) để mã hóa một câu hỏi thành vector embedding. Mô hình có giá trị thấp hơn sẽ nhanh hơn trong việc xử lý các truy vấn, phù hợp cho các ứng dụng thời gian thực hoặc yêu cầu độ trễ thấp [78], [79].

Mỗi mô hình embedding đều có những thông số xử lý tùy thuộc vào từng dự án cụ thể và ưu nhược điểm riêng. Sau khi phân tích và so sánh các chỉ số hiệu suất trong bảng 3, chúng tôi quyết định chọn mô hình *nomic-ai/nomic-embed-text-v1.5* làm mô hình embedding chính vì những lý do sau:

* Hiệu suất Top-K: Mô hình *nomic-ai* cho thấy khả năng truy xuất chính xác vượt trội với giá trị Top-K Accuracy cao trong các mẫu thử nghiệm. Điều này chứng tỏ Nomic có hiệu quả trong việc xếp hạng các kết quả liên quan, phù hợp cho việc tăng cường truy xuất tạo sinh (RAG).
* Số chiều vector tối ưu: Mô hình cung cấp số chiều embedding cân bằng, vừa đủ lớn để nắm bắt đặc trưng ngữ nghĩa quan trọng nhưng không quá lớn để gây ra chi phí tính toán vượt mức. Điều này giúp giảm thiểu gánh nặng bộ nhớ và thời gian xử lý mà vẫn đảm bảo chất lượng embedding.
* Thời gian embedding: *nomic-ai* có thời gian trung bình để embedding câu và câu hỏi thấp hơn so với nhiều mô hình khác. Điều này rất quan trọng đối với các ứng dụng yêu cầu tốc độ xử lý nhanh, như chatbot thời gian thực hoặc hệ thống hỏi đáp.
* Khả năng xử lý bối cảnh dài: Với độ dài bối cảnh cao, cho phép tối đa khoảng 8,000 tokens, *nomic-ai* phù hợp để xử lý các đoạn văn bản dài và câu hỏi phức tạp, nâng cao khả năng phục vụ trong các trường hợp truy vấn ngữ nghĩa sâu.
* Tối ưu hóa tài nguyên GPU: Mô hình Nomic có mức chiếm dụng bộ nhớ GPU tối ưu hơn so với các mô hình lớn khác. Điều này giúp dễ dàng triển khai trên các hệ thống có tài nguyên phần cứng hạn chế mà không ảnh hưởng đến hiệu năng.
* Độ trễ thấp và hiệu quả xử lý hàng loạt: Khả năng mã hóa nhanh của *nomic-ai* giúp xử lý dữ liệu lớn trong thời gian ngắn, tối ưu hóa cho các kịch bản yêu cầu tốc độ như phân tích dữ liệu hàng loạt hoặc cung cấp phản hồi người dùng tức thì.
* Miễn phí thử nghiệm: Mô hình *nomic-ai/nomic-embed-text-v1.5* cung cấp lượng tokens thử nghiệm, điều này phù hợp để lựa chọn xây dựng và thử nghiệm cho đề tài chatbot có quy mô nhỏ hoặc vừa phải.

## Phương pháp đánh giá hiệu quả retrieval dữ liệu

Sau khi thực hiện embedding dữ liệu và lưu vào vector databse, các mô hình retrieval được sử dụng để đánh giá hiệu quả embedding, chất lượng dữ liệu và độ chính xác retrieval dữ liệu trước khi đưa vào mô hình LLM. Một phương pháp retrieval tốt cung cấp câu trả lời tốt. Trong dự án này chúng tôi sử dụng 3 phương pháp retrieval bao gồm retrieval theo vector, retrieval theo keyword và retireval theo hybrid. Trong hình 29 là mô tả quá trình thực hiện đánh giá hiệu suất retrieval.

A diagram of a model

Description automatically generated

Hình 29. Quy trình đánh giá hiệu suất retrieval dữ liệu.

Với Retrieval theo Vector, sử dụng các vector embedding để đại diện cho các đối tượng thông tin (như tài liệu, câu hỏi, hoặc câu trả lời). Đây là một kỹ thuật phổ biến trong các hệ thống tìm kiếm thông tin hiện đại, đặc biệt khi kết hợp với các mô hình học sâu. Bên cạnh đó, keyword retrieval là phương pháp truy xuất thông tin truyền thống hơn, nơi các truy vấn của người dùng được so khớp với các từ khóa trong cơ sở dữ liệu. Phương pháp này thường sử dụng các kỹ thuật tìm kiếm chuỗi, như tìm kiếm từ khóa chính xác hoặc biến thể từ khóa trong văn bản. Ngoài ra, phương pháp hybrid retrieval kết hợp cả retrieval theo vector và retrieval theo keyword để kết hợp những ưu điểm của cả hai phương pháp, từ đó cải thiện chất lượng và hiệu quả của việc tìm kiếm.

Theo đó chúng tôi sử dụng các chỉ số đánh giá hiệu quả retrieval bao gồm: chỉ số P@3, Recall@3, và MRR@3, đây là các chỉ số đánh giá phổ biến trong các hệ thống retrieval thông tin, đặc biệt là trong các hệ thống trả lời câu hỏi như chatbot. Mỗi chỉ số này giúp đo lường một khía cạnh khác nhau của hiệu suất tìm kiếm, đặc biệt là trong việc so sánh kết quả trả về và so khớp với các tài liệu hay câu trả lời đúng.

* Chỉ số P@3 (Precision at 3): Precision at 3 (P@3) đo lường độ chính xác của kết quả tìm kiếm dựa trên top 3 kết quả đầu tiên trả về từ hệ thống.

P@3 giúp đo lường khả năng của hệ thống trong việc trả về các kết quả chính xác ngay từ đầu (3 kết quả đầu tiên). Một giá trị P@3 cao cho thấy hệ thống có xu hướng đưa ra các kết quả có liên quan cao ngay lập tức.

* Chỉ số Recall@3: Recall at 3 (Recall@3) đo lường khả năng của hệ thống trong việc lấy lại tất cả các tài liệu hoặc câu trả lời đúng từ bộ dữ liệu, trong số 3 kết quả đầu tiên.

Recall@3 đo lường khả năng của hệ thống trong việc không bỏ sót các câu trả lời đúng trong số 3 kết quả đầu tiên. Một giá trị Recall cao nghĩa là hệ thống có khả năng tìm thấy phần lớn các kết quả đúng trong số ít các kết quả đầu tiên.

* Chỉ số MRR@3 (Mean Reciprocal Rank at 3): Mean Reciprocal Rank at 3 (MRR@3) đo lường mức độ tốc độ mà hệ thống trả về câu trả lời đúng trong top 3 kết quả đầu tiên. Đây là một chỉ số đánh giá xếp hạng cho biết vị trí của câu trả lời đúng trong danh sách trả về.

MRR@3 tập trung vào vị trí của câu trả lời đúng, với giá trị càng cao khi câu trả lời đúng xuất hiện càng sớm trong danh sách trả về. MRR đánh giá khả năng của hệ thống trong việc đưa câu trả lời đúng lên vị trí cao trong danh sách kết quả.

Dưới đây là bảng so sánh các chỉ số đánh giá retrieval của các phương pháp trên:

Bảng 6. Bảng so sánh hiệu suất các phương pháp retrieval.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **P@3** | **Recall@3** | **MRR@3** |
| VECTOR | 0,173 | 0,52 | 0,363 |
| KEYWORD | 0,197 | 0,59 | 0,477 |
| HYBRID | 0,213 | 0,64 | 0,44 |

## Các mô hình LLM được sử dụng

Sự phát triển mạnh mẽ của các mô hình LLM trong thời gian gần đây trong cuộc đua AI đã tạo nên sự đa dạng mô hình LLM, khi ngày càng có nhiều mô hình được ra mắt với khả năng không ngừng cải tiến về độ chính xác, tốc độ và khả năng xử lý ngữ cảnh [88], [89]. Các LLM hiện đại không chỉ tập trung vào việc sinh văn bản mà còn mở rộng ứng dụng vào các lĩnh vực như tìm kiếm ngữ nghĩa (semantic search), lập luận logic, và hỗ trợ đa ngôn ngữ, đáp ứng nhu cầu ngày càng tăng trong nghiên cứu và công nghiệp [88], [89]. Các mô hình LLM mạnh mẽ hầu hết được phát triển cho mục đích công nghiệp, tuy nhiên vẫn có những mô hình LLM được phát hành miễn phí dành cho các nhà phát triển thử nghiệm và nghiên cứu [88], [89], [90]. Dưới đây là một vài mô hình LLM mở được cân nhắc sử dụng trong báo cáo này:

#### Mô hình LLM Llama 3

Llama 3 là một mô hình ngôn ngữ lớn (LLM) mới của Meta, tiếp nối thành công của Llama 2. Nó được thiết kế với cấu trúc Transformer tối ưu và có sẵn dưới dạng hai kích thước: 8 tỷ và 70 tỷ tham số. Llama 3 được huấn luyện trên tập dữ liệu lớn gấp bảy lần so với Llama 2, bao gồm hơn 15 nghìn tỷ token, giúp tăng khả năng xử lý ngôn ngữ và giảm tỉ lệ từ chối câu trả lời sai. Mô hình này hỗ trợ ngữ cảnh dài tới 128.000 token, vượt trội so với các mô hình trước đó. Trong các thử nghiệm, Llama 3 có hiệu suất tốt hơn nhiều đối thủ như Gemini 1.5 và Claude trên các chỉ số hiệu năng quan trọng. Mô hình này cũng hỗ trợ các ứng dụng sử dụng LangChain và LlamaIndex để tích hợp vào các hệ thống chatbot hoặc AI khác

#### Mô hình LLM Gemini 1.5

Gemini 1.5 là phiên bản nâng cấp của dòng Gemini do Google phát triển. Mô hình này tập trung vào khả năng sinh ngữ cảnh chi tiết và hỗ trợ đa ngôn ngữ. Với các cải tiến về huấn luyện và xử lý token, Gemini 1.5 đã tăng độ chính xác và tốc độ xử lý trong các tác vụ như dịch thuật và hội thoại. Dù không đạt được mức hiệu suất của Llama 3 trên một số chỉ số, Gemini 1.5 vẫn là lựa chọn mạnh mẽ cho các ứng dụng AI tập trung vào xử lý hội thoại tự nhiên và ngôn ngữ toàn cầu​

#### Mô hình LLM Grok

Grok là mô hình AI do X phát triển, được tích hợp vào nền tảng mạng xã hội X để hỗ trợ viết bài đăng và tương tác với người dùng. Đây là một LLM hướng đến khả năng tạo nội dung ngắn gọn, súc tích, với trọng tâm là tối ưu trải nghiệm người dùng trong các tác vụ hội thoại đơn giản và viết nội dung truyền thông. Mặc dù Grok không được thiết kế để xử lý các tác vụ phức tạp như các mô hình lớn khác, nó vẫn đóng vai trò quan trọng trong các ứng dụng hỗ trợ người dùng trực tiếp.

Việc lựa chọn nhiều LLM để thực hiện xây dựng chatbot là vì để tận dụng tối đa những gì miễn phí mà các mô hình này có như tài nguyên tính toán, lượng truy vấn giới hạn của mỗi mô hình, ngoài ra khi thực hiện một nhiệm vụ được yêu cầu từ người dùng, việc phân chia các nhiệm vụ cho các mô hình này thực hiện cùng lúc sẽ làm giảm thời gian trong quá trình xử lý, do đó tăng hiệu suất trả lời và giảm độ trễ phản hồi của chatbot.

## Prompt engineering trong hệ thống chatbot

Prompt Engineering (lời nhắc) đóng vai trò cốt lõi trong việc giao tiếp với mô hình LLM. Giúp định hướng mô hình xử lý thông tin chính xác, hiểu ngữ cảnh của câu hỏi từ người dùng và tạo ra câu trả lời phù hợp với kỳ vọng. Việc xây dựng prompt tốt giúp tăng độ chính xác, đảm bảo mô hình LLM hiểu được mục tiêu câu hỏi. Giảm sai lệch, tránh việc mô hình đưa ra các phản hồi không liên quan hoặc không chính xác. Tối ưu hóa trải nghiệm, đưa ra câu trả lời đầy đủ, dễ hiểu, và ngắn gọn. Dưới đây là quy trình triển khai prompt cho hệ thống. Cấu trúc của một prompt trong hệ thống chatbot hỏi đáp tuyển sinh này bao gồm các thành phần sau (hình 30):

A screenshot of a phone

Description automatically generated

Hình 30. Cấu trúc prompt trong hệ thống chatbot hỏi đáp tuyển sinh.

Trong đó:

* Vai trò: Xác định vai trò hoặc vị trí mà hệ thống AI cần đóng. Mục tiêu là định hướng hệ thống vào một cách tư duy hoặc phong cách phản hồi cụ thể.
* Bối cảnh: Cung cấp thông tin cần thiết để AI hiểu tình huống hoặc nội dung liên quan.
* Câu hỏi: Câu hỏi hoặc yêu cầu chính mà hệ thống giải quyết.
* Các ví dụ: Là các bản hướng dẫn, cung cấp các ví dụ minh họa để hệ thống có thể hiểu rõ kỳ vọng về đầu ra.
* Hướng dẫn: Đưa ra hướng dẫn chi tiết về cách hệ thống nên trả lời
* Lưu ý: Các lưu ý quan trọng giúp chatbot có thể tránh các điều cần tránh trả lời và các điều mà chatbot nên làm, ví dụ như prompt yêu cầu chatbot hạn chế đưa ra các từ toxic.

Prompt được cấu trúc thành một chuỗi văn bản đầy đủ bao gồm các thành phần chính bao gồm ngữ cảnh chứa nội dung liên quan từ cơ sở dữ liệu (ví dụ: thông tin truy vấn, dữ liệu lịch sử của người dùng) và câu hỏi của người dùng.

Việc xây dựng prompt diễn ra tại bước (9.1), sau khi thông tin được trích xuất từ cơ sở dữ liệu vector (8). Quy trình cụ thể bao gồm:

Thu thập thông tin từ truy vấn:

* Dữ liệu từ người dùng được gửi qua bước trích xuất thông tin (7, 8) để lấy ra các đoạn thông tin phù hợp từ cơ sở dữ liệu vector.
* Các đoạn thông tin này đóng vai trò cung cấp ngữ cảnh cho prompt.

Prompt sau khi được định dạng sẽ được gửi đến mô hình LLM để tạo phản hồi. Tạo điều kiện để LLM xử lý các truy vấn có liên quan trực tiếp đến dữ liệu được lưu trữ. Ngoài ra, tùy chỉnh lời nhắc theo loại tác nhân để phản hồi phù hợp để tối ưu hóa chất lượng câu trả lời, giảm thiểu các lỗi như câu trả lời quá ngắn, lan man, hoặc không chính xác.

## Tác nhân (Agents)

Tác nhân (agent) là các thành phần trung gian, hoạt động độc lập hoặc phối hợp với LLM để giải quyết các tác vụ cụ thể. Trong hệ thống này, các tác nhân giúp chia nhỏ công việc của chatbot thành các nhiệm vụ, đảm bảo chatbot có thể xử lý đa dạng các yêu cầu từ người dùng, tăng khả năng mở rộng của hệ thống bằng cách dễ dàng thêm mới các tác nhân với chức năng đặc thù. Hệ thống chatbot tuyển sinh triển khai ba loại tác nhân chính như sau: Tác nhân hỏi đáp, tác nhân điều hướng người dùng cho các truy vấn ngoài phạm vi.

### Tác nhân hỏi đáp (Q&A Agent)

Hình 31 dưới đây mô tả luồng hoạt động của một Q&A agent từ đầu đến cuối.

A diagram of a company

Description automatically generated

Hình 31. Mô tả luồng hoạt động của Q&A agent.

Question & Answering Agent (Tác nhân Hỏi & Đáp – Q&A agent) là thành phần quan trọng nhất trong hệ thống, chịu trách nhiệm xử lý và trả lời hầu hết các câu hỏi của người dùng. Phạm vi câu hỏi được giới hạn trong dữ liệu tuyển sinh đã được thu thập và lưu trữ trong cơ sở dữ liệu vector, đảm bảo các câu trả lời luôn bám sát thông tin đã được xác minh.

* Mô tả chức năng của Q&A agent: Trả lời các câu hỏi thường gặp và những câu hỏi liên quan đến thông tin tuyển sinh trong cơ sở dữ liệu vector. Đảm bảo phản hồi chính xác và nhanh chóng bằng cách tận dụng cơ sở dữ liệu vector và sức mạnh xử lý ngôn ngữ tự nhiên của mô hình LLM. Cung cấp câu trả lời trực tiếp, ngắn gọn nhưng đầy đủ, nhằm tối ưu trải nghiệm người dùng.
* Quy trình hoạt động: Khi hệ thống phát hiện truy vấn của người dùng nằm trong phạm vi thông tin đã được lưu trữ, Q&A agent sẽ được agent manager điều phối. Cơ chế nhận diện dựa trên việc so khớp ngữ nghĩa của truy vấn với nội dung trong cơ sở dữ liệu vector. Hệ thống tìm kiếm và lấy ngữ cảnh liên quan từ cơ sở dữ liệu vector. Thông tin này sau đó được chuyển tiếp đến mô hình LLM thông qua cơ chế lời nhắc (prompting) để hỗ trợ tạo câu trả lời chính xác và phù hợp
* Tạo phản hồi: Sau khi xử lý thông tin, mô hình LLM sẽ đưa ra phản hồi trực tiếp dựa trên dữ liệu ngữ cảnh đã được cung cấp. Câu trả lời cuối cùng được hệ thống gửi lại cho người dùng thông qua giao diện tương tác.

### Tác nhân ngoài phạm vi (Out-of-Domain Agent)

Tác nhân trả lời câu hỏi ngoài phạm vi được thiết kế nhằm đảm bảo trải nghiệm người dùng không bị gián đoạn ngay cả khi chatbot không thể trực tiếp trả lời câu hỏi. Nó hoạt động như một cầu nối, hỗ trợ người dùng tiếp cận thông tin đúng cách hoặc kết nối với nguồn hỗ trợ phù hợp. Hình 26 dưới đây mô tả luồng hoạt động của một out\_of\_domain agent (OOD agent).

A diagram of a process

Description automatically generated

Hình 32. Mô tả luồng hoạt động của OOD agent.

* Mô tả: Xử lý các yêu cầu không thuộc phạm vi kiến thức hiện tại của chatbot. Khi truy vấn không phù hợp, hệ thống ngay chuyển quyền xử lý cho tác nhân này. Trong trường hợp cần thiết, tác nhân này sẽ chuyển yêu cầu đến các cán bộ hoặc nhân viên có chuyên môn để đảm bảo người dùng nhận được hỗ trợ đầy đủ.
* Quy trình hoạt động: Dựa trên cơ chế nhận diện ngữ nghĩa, tác nhân xác định rằng truy vấn của người dùng không khớp với dữ liệu hiện có trong cơ sở dữ liệu vector. Hệ thống sẽ tự động yêu cầu người dùng cung cấp thêm thông tin để làm rõ truy vấn, hoặc điều hướng tới cán bộ tư vấn trực tiếp. Những phản hồi này được LLM đưa ra thông qua cơ chế prompt.
* Tạo phản hồi: Hệ thống sẽ nhận phản hồi từ mô hình LLM và cung cấp cho agent manager để hiển thị đầu ra cho người dùng.

### Quản lý tác nhân (Agent Management)

Vai trò: Điều phối các tác nhân trong hệ thống để đảm bảo chọn đúng tác nhân dựa trên loại truy vấn. Quản lý giao tiếp giữa tác nhân, cơ sở dữ liệu và mô hình LLM.

Quy trình:

* Phân loại truy vấn sau khi người dùng nhập câu hỏi có nằm trong phạm vi trả lời của chatbot hay không.
* Kích hoạt tác nhân phù hợp thông qua Agent Call (11, 12.x).
* Gửi thông tin trả về từ tác nhân đến bước xử lý tiếp theo.

Thông qua việc xây dựng lời nhắc và tác nhân cho hệ thống chatbot, sự phối hợp giữa kỹ thuật tạo lời nhắc và các tác nhân giúp hệ thống chatbot trở nên mạnh mẽ, linh hoạt và dễ mở rộng. Trong đó Prompt Engineering tối ưu hóa tương tác giữa người dùng và LLM. Các tác nhân đảm bảo hệ thống có thể xử lý nhiều loại nhiệm vụ khác nhau, từ việc hỏi đáp đơn giản đến điều hướng và xử lý các yêu cầu phức tạp. Việc triển khai kỹ thuật này không chỉ nâng cao độ chính xác mà còn tăng cường trải nghiệm người dùng và khả năng ứng dụng của hệ thống chatbot trong các lĩnh vực khác nhau.

## Cá nhân hóa người dùng

Một trong những thành phần quan trọng nhất trong hệ thống chatbot là làm thế nào để giúp chatbot hiểu người dùng, từ đó chatbot sẽ đưa ra những câu trả lấy người dùng làm trung tâm. Điều này tạo nên trải nghiệm tương tác thân thiện, đồng thời mang lại nhiều lợi ích hơn trong việc tư vấn tuyển sinh, như:

* Tăng cường trải nghiệm người dùng: Khi chatbot hiểu rõ nhu cầu, sở thích, và ngữ cảnh của từng cá nhân, câu trả lời sẽ trở nên phù hợp hơn. Điều này giúp người dùng cảm thấy được quan tâm và hài lòng khi sử dụng hệ thống.
* Tăng khả năng tương tác lâu dài: Chatbot được cá nhân hóa có khả năng tạo ra các cuộc hội thoại thú vị và gần gũi, từ đó khuyến khích người dùng quay lại và tương tác nhiều hơn. Điều này đặc biệt quan trọng đối với các sinh viên tiềm năng có nhu cầu tìm hiểu về trường.
* Đưa ra các câu hỏi gợi ý cá nhân hóa: Đối với các nhiệm vụ tư vấn tuyển sinh, việc chatbot được cá nhân hóa cho người dùng có thể gợi ý các câu hỏi, lĩnh vực phù hợp dựa trên sở thích và nhu cầu tìm hiểu, từ đó tăng khả năng tin tưởng và tạo thiện cảm với người dùng.

A diagram of a flowchart

Description automatically generated

Hình 33. Quy trình thu thập thông tin và cá nhân hóa người dùng.

Hình 33 mô tả quy trình thu thập thông tin truy vấn đầu vào và thực hiện cá nhân hóa người dùng dựa trên những thông tin thu thạp được. Đầu tiên người dùng sẽ nhập vào các thông tin muốn tìm hiểu.Các thông tin này sẽ được trích xuất thông minh để nhận diện các chủ đề, lĩnh vực hoặc khía cạnh mà người dùng quan tâm và sau đó lưu vào cơ sở dữ liệu MongoDB, đồng thời đưa tới LLM, kết hợp với các thông tin truy xuất cho mục đích trả lời, từ đó chatbot sẽ đưa ra câu trả lời được cá nhân hóa cho người dùng hiện tại.

# TRIỂN KHAI HỆ THỐNG CHATBOT TUYỂN SINH

## Triển khai embedding dữ liệu và truy vấn người dùng

Sau khi dữ liệu được tiền xử lý và chuẩn hóa, hệ thống sử dụng mô hình embedding để chuyển đổi dữ liệu này thành các vector số, giúp biểu diễn thông tin theo dạng có thể tính toán. Đồng thời, câu truy vấn của người dùng cũng được chuyển đổi thành vector tương tự bằng cách sử dụng cùng mô hình embedding, giúp dễ dàng so sánh và tìm kiếm thông tin liên quan trong cơ sở dữ liệu. Hình 34 là quá trình triển khai embedding dữ liệu và câu truy vấn người dùng sử dụng mô hình embedding *nomic-ai/nomic-embed-text-v1.5*:

A diagram of a question

Description automatically generated with medium confidence

Hình 34. Quá trình triển khai embedding bộ dữ liệu sử dụng mô hình embedding.

Đầu tiên dữ liệu được chuẩn hóa sẽ được xử lý bằng chương trình Python, sau khi định dạng phù hợp thì đưa vào mô hình embedding, tại đây mô hình sẽ tự động thực hiện embedding từ văn bản sang svector embedding. Các vector embedding này sẽ được sử dụng để đưa vào lưu trữ trong Milvus.

## Triển khai lưu trữ Vector embedding vào Vector database

Vector embeddings được sinh ra từ dữ liệu câu hỏi và câu trả lời bằng mô hình embedding, sau đó được tổ chức và lưu trữ trong một vector database sử dụng Milvus, với ưu điểm dễ sử dụng và khả năng quản lý dữ liệu hiệu quả, là lựa chọn phù hợp cho dự án nhỏ và trung bình như chatbot tuyển sinh. Bước lưu trữ này thường được thực hiện sau khi đã embedding dữ liệu từ mô hình, nhằm chuẩn bị cho các hoạt động tìm kiếm thông tin hoặc truy vấn tương tự (vector search). Đây là nền tảng để chatbot có thể tìm kiếm và phản hồi chính xác hơn. Các bước thực hiện lưu vector vào database Milvus thông thường bao gồm các bước như sau:

* Cài đặt các thư viện cần thiết.
* Khởi chạy Milvus trên máy tính cục bộ bằng Docker:
* Sử dụng pymilvus để kết nối tới Milvus:
* Chuẩn bị dữ liệu embeddings: Sau khi đã thực hiện embedding từ mô hình, chuẩn bị dữ liệu theo định dạng phù hợp:
* Lưu embeddings vào Milvus: Sử dụng Milvus client để gửi dữ liệu vector vào collection.

Hình 35 là quá trình thực hiện xử lý và lưu vector embedding vào vector database bằng Milvus.

A diagram of a software development process

Description automatically generated

Hình 35. Quá trình lưu trữ vector embedding vào vector database.

Tại bước này, chúng tôi định dạng metadata có cấu trúc như sau:

*# Định dạng metadata để lưu vào vector database*

metadata = [

    {

        "id": 1,

        "question": "Khi nào bắt đầu tuyển sinh 2025?",

        "answer": "Tháng 6 năm 2025",

*# Embeddings được tạo từ mô hình embedding*

        "embedding": [0.1, 0.2, ...]

    },

    {

        "id": 2,

        "question": "Chỉ tiêu tuyển sinh năm 2025?",

        "answer": "10,000 sinh viên",

        "embedding": [0.4, 0.5, ...]

    },

]

Id định danh cho mẫu dữ liệu, số lượng mẫu trong data được chuẩn hóa tương ứng với số Id trong vector database, meta bao gồm câu hỏi, câu trả lời và vector embedding đã được chuẩn hóa trước đó.

## Retrieval cho truy vấn người dùng

Việc retrieval cho một truy vấn người dùng được nhập vào trong hệ thống chatbot được thực hiện sau khi embedding dữ liệu thành vector và lưu trữ trong vector database Milvus. Hệ thống sẽ tìm kiếm dựa trên vector để liên kết các câu hỏi của người dùng với các thông tin đã được lưu trữ. Phương pháp kết hợp embedding với tìm kiếm vector giúp nâng cao khả năng tìm kiếm theo ngữ cảnh, cho phép chatbot truy xuất các kết quả sát nghĩa nhất với câu truy vấn, đồng thời tối ưu hóa thời gian phản hồi. Trong phần này, chúng tôi trình bày cách thực hiện truy vấn từ cơ sở dữ liệu Milvus bằng cách sử dụng embedding của câu truy vấn, sau đó bằng cơ chế tìm kiếm vector của Milvus để trả về các câu trả lời phù hợp nhất. Hình 36 dưới đây là quy trình thực hiện truy xuất dữ liệu từ đầu vào của người dùng, bao gồm embedding truy vấn từ vector database Milvus.

A diagram of a process

Description automatically generated

Hình 36. Các bước thực hiện retrieval cho truy vấn người dùng.

Đầu tiên, người dùng nhập vào một câu truy vấn, câu hỏi sẽ được embedding bằng mô hình nomic-ai/nomic-embed-text-v1.5 embedding (1). Sau đó chương trình sẽ dùng vector embediding này (2), bằng thuật toán retrieval để tìm kiếm dữ liệu liên quan trong vector database (3). Sau đó trả về các kết quả truy vấn được (4) theo chỉ định, trong dự án của này chúng tôi chỉ định trả về 5 kết quả liên quan nhất trong vector database (5). Các kết quả này bao gồm một điểm số score biểu thị mức độ liên quan nhất được sắp xếp theo thứ tự liên quan nhất. Các kết quả này sẽ được dùng để tạo câu trả lời cho sau này.

## Triển khai agent cho hệ thống chatbot

Agent đóng vai trò quan trọng trong hệ thống chatbot hỏi đáp tuyển sinh. Nó giúp hệ thống có khả năng thực hiện các hành động nhất định, như giao tiếp với người dùng, xử lý thông tin, truy vấn dữ liệu, và phản hồi. Một agent có thể sử dụng prompt để giao tiếp với mô hình AI. Agent sẽ chuẩn bị prompt bao gồm các câu hỏi, thông điệp hoặc yêu cầu và gửi nó đến mô hình LLM để nhận được phản hồi. Trong trường hợp này, prompt là thành phần mà agent sử dụng để tương tác với mô hình. Trong phần này chúng tôi sẽ trình bày các thành phần của agent cho hệ thống chatbot hỏi đáp và cách triển khai cho từng thành phần của agent.

### Xây dựng cấu trúc Prompt

Prompt được xây dựng dựa trên thiết kế trong hình 30 của phương pháp đề xuất. Đây là thành phần quan trọng giúp định hướng cách thức mà LLM sẽ trả lời câu hỏi của người dùng, đặc biệt chất lượng phản hồi của chatbot phụ thuộc phần lớn vào prompt được xây dựng. Khi các kết quả tìm kiếm từ cơ sở dữ liệu được truy xuất dựa trên truy vấn của người dùng đã được thu thập và xếp hạng, chúng sẽ được sử dụng để tạo ra prompt cho mô hình. Prompt có thể được xây dựng bằng cách kết hợp câu hỏi người dùng và các thông tin từ kết quả tìm kiếm, nhằm cung cấp ngữ cảnh rõ ràng và chính xác hơn cho mô hình LLM. Thông qua việc sử dụng các kết quả tìm kiếm này, hệ thống có thể cung cấp câu trả lời phù hợp hơn, dựa trên kiến thức có sẵn từ cơ sở dữ liệu. Quy trình xây dựng prompt cho hệ thống chatbot tuyển sinh như sau:

* Nhắc nhở cho LLM biết nó là ai và vai trò của nó hiện tại, điều này giúp chatbot có thể xưng hô với người dùng dựa theo tên được cung cấp. Ví dụ như cho chatbot biết rằng nó là UDAChat, nhiệm vụ của nó là sử dụng các thông tin quan trọng để cung cấp cho người dùng…
* Cung cấp một tóm tắt bối cảnh dành cho chatbot bao gồm cung cấp thông tin bối cảnh hiện tại, cung cấp lời nhắc rằng chatbot sẽ làm những nhiệm vụ mà chatbot có thể làm. Ví dụ như nhiệm vụ của chatbot là sẽ trả lời các câu hỏi tư vấn tuyển sinh, các chủ đề bao gồm…
* Cung cấp cho chatbot câu hỏi mà người dùng cần tư vấn do người dùng nhập.
* Nhận dạng kết quả từ retrieval và rerank: Sau khi thực hiện truy xuất retrieval và có kết quả, cần lọc ra các văn bản có liên quan nhất từ cơ sở dữ liệu hoặc vector database. Các kết quả này là các câu trả lời từ cơ sở dữ liệu mà chatbot có thể sử dụng để cung cấp thông tin cho người dùng. Các thông tin này sẽ được đưa vào prompt kết hợp với câu hỏi do người dùng nhập vào.
* Cung cấp cho chatbot những hướng dẫn cần thiết và các lưu ý để chatbot có thể trả lời một cách hiệu quả và tốt nhất.

### Triển khai prompt hỏi đáp

Prompt hỏi đáp có nhiệm vụ cung cấp cho LLM thông tin để trả lời câu hỏi tư vấn tuyển sinh. Dựa trên cấu trúc prompt được đề xuất trong dự án này. Prompt hỏi đáp cũng sẽ bao gồm các thành phần cơ bản của prompt. Dưới đây là một prompt cho tác nhân hỏi đáp hệ thống chatbot tuyển sinh:

Bạn là UDAChat.

{human\_preference}. Sử dụng những thông tin quan trọng này để tăng tương tác cũng như mang đến thông tin cần thiết và hữu ích cho người dùng.

Nhiệm vụ của bạn là trả lời các câu hỏi về thông tin tuyển sinh và thông tin chung của trường Đại học Đông Á như ngành học, học phí, chính sách học bổng, môi trường học tập, hồ sơ đăng kí xét tuyển, giấy tờ hồ sơ cần thiết, điểm chuẩn xét học bạ, điểm chuẩn xét tuyển thi...

Số điện thoại hotline: *\*0236.351.9929\**

<Câu hỏi:>

<Thông tin:>

Ví dụ:

Các thông tin bạn sẽ phản hồi về tuyển sinh và về trường Đại học Đông Á

Lưu ý:

- Bạn nên xưng hô là `mình` cho tất cả phản hồi của bạn.

- IMPORTANT: Your answer should be in **\*\*Vietnamese\*\***.

- Hãy chọn lọc cẩn thận những *\*nội dung được trích xuất\** để trả lời yêu cầu của người dùng một các phù hợp.

Trong đoạn prompt trên, chúng tôi tạo một lời nhắc cho biết vai trò và nhiệm vụ, đồng thời cung cấp thông tin số điện thoại của ban tuyển sinh (lấy số máy của Trường Đại học Đông Á làm ví dụ). Tiếp đó, prompt sẽ được cung cấp câu hỏi của người dùng và các thông tin truy xuất được. Cuồi cùng là cung cấp những ví dụ và các nhắc nhở lưu ý để LLM có thể trả lời tốt hơn. Prompt cho nhiệm vụ hỏi đáp này có thể tùy chỉnh, lợi ích của việc tạo prompt tùy chỉnh có thể dễ dàng linh hoạt nâng cấp, cải thiện và sửa đổi sau này.

Quy trình sử dụng prompt hỏi đáp câu hỏi tuyển sinh được xử lý bởi hầu hết các thành phần trong hệ thống. Hình 37 dưới đây là cách mà prompt được sử dụng cho nhiệm vụ hỏi đáp.

A diagram of a process

Description automatically generated

Hình 37. Cách một prompt được hoàn chỉnh và sử dụng trong hệ thống chatbot.

Trong đó prompt template là một đoạn prompt được tạo từ trước, sau đó prompt này được cung cấp thêm thông tin nhờ kết quả truy xuất từ cơ sở dữ liệu để tạo thành prompt hoàn chỉnh. Lúc này prompt hoàn chỉnh sẽ được đưa vào LLM để sinh ra câu hỏi. Câu trả lời sẽ được hệ thống chatbot xử lý và hiển thị cho người dùng.

### Triển khai prompt điều hướng người dùng

Quy trình triển khai prompt nhắc nhở người dùng yêu cầu nằm ngoài phạm vi và thực hiện điều hướng người dùng tương tự với quy trình triển khai prompt trả lời câu hỏi tuyển sinh, đó là dựa theo cấu trúc cơ bản prompt được đê cập trong phương pháp đề xuất. Điểm khác là prompt điều hướng người dùng cần phải cho chatbot biết xin lỗi người dùng vì câu hỏi không nằm trong phạm vi của chatbot. Ngoài ra trong prompt điều hướng người dùng sẽ không bao gồm các thông tin được truy xuất từ cơ sở dữ liệu và cách sử dụng prompt điều hướng người dùng cũng sẽ khác với prompt trả lời câu hỏi tuyển sinh. Dưới đây là prompt điều hướng người dùng khi quy vấn ngoài phạm vi của chatbot tuyển sinh:

Bạn là UDAChat.

{human\_preference} một trợ lý ảo chuyên hỗ trợ thông tin tuyển sinh và thông tin chung về trường Đại học Đông Á.

Nhiệm vụ chính của bạn bao gồm: Trả lời các câu hỏi về thông tin tuyển sinh: ngành học, học phí, môi trường học tập, hồ sơ đăng ký xét tuyển, giấy tờ cần thiết, điểm chuẩn xét học bạ, điểm chuẩn xét tuyển thi, v.v.

Tuy nhiên người dùng đã nhập vào một thông tin không liên quan đến tuyển sinh, bạn cần nhắc người dùng rằng bạn chỉ hỗ trợ thông tin tuyển sinh. Lịch sự giải thích rằng bạn không thể hỗ trợ chính xác câu hỏi này.

Gợi ý liên hệ với đội ngũ chuyên trách hoặc cung cấp nguồn thông tin phù hợp.

Đảm bảo giữ tương tác thân thiện và hỗ trợ người dùng tiếp cận thông tin cần thiết.

- Số điện thoại hotline: *\*0236.351.9929\**

Lưu ý:

  - Bạn nên xưng hô là `mình` trong tất cả phản hồi của bạn.

  - IMPORTANT: Your answer should be in **\*\*Vietnamese\*\***.

  - Các thông tin nằm trong ngữ cảnh trò chuyện bạn hãy phản hồi một các tự nhiên và phù hợp nhất.

  - *Không trả lời và xin lỗi một cách lịch sự là bạn không biết* đối với các câu hỏi nằm ngoài hoàn toàn phạm vi tuyển sinh *năm 2024*, thông tin về Trường Đại học Đông Á hoặc các câu hỏi ở các lĩnh vực khác như thời tiết, môn học, thể thao, giải trí, v.v. Bạn chỉ có thể trả lời các thông tin về trường và tuyển sinh của Trường Đại học Đông Á.

  - *Không trả lời* các thông tin *có thể liên quan đến giáo dục và tuyển sinh của trường* nhưng không phải của *năm 2024* (Ví dụ: năm ngoái(2023), năm tới(2025), đề thi trung học phổ thông quốc gia 2025). Trong trường hợp này hãy xin lỗi và điều hướng người dùng sang cán bộ tuyển sinh.(Thời gian hiện tại: năm 2024)

Hình 38 dưới đây là minh họa cách mà prompt điều hướng người dùng được sử dụng trong hệ thống chatbot.

A diagram of a person's process

Description automatically generated

Hình 38. Prompt điều hướng người dùng trong hệ thống chatbot.

Đầu tiên (1), việc đặt giám định phát hiện truy vấn ngoài phạm vi là điều cần thiết, tránh cho việc chatbot có thể trả lời các thông tin không chính xác hoặc không chính thống. Điều này đảm bảo rằng chatbot chỉ trả lời duy nhất về các thông tin liên quan đến tuyển sinh của trường dựa trên thông tin đã được thu thập mà thôi. Nếu phát hiện ra truy vấn của người dùng (2) không nằm trong phạm vi, hệ thống sẽ trực tiếp sử dụng prompt này vào LLM (3) để tạo ra phản hồi (4) và nhắc nhở người dùng rằng chatbot không hỗ trợ phạm vi câu hỏi của người dùng.

### Trình quản lý agent và function calling

Để chatbot có thể hoạt động hiệu quả từ bước người dùng nhập truy vấn tới tới bước đưa ra phản hồi hoàn chỉnh, cần thiết lập một trình quản lý agent (agent management) để điều hướng các chức năng và giải quyết nhiệm vụ trong chatbot. Cụ thể tại đây sẽ nhận đầu vào là truy vấn của người dùng cùng với danh sách các hàm có thể gọi, đi kèm với đó là thông tin các tham số cần được truyền vào hàm, một mô tả ngắn về chức năng và nhiệm vụ của mỗi hàm. Những thông tin này sẽ được đưa vào mô hình LLM, sau đó LLM sẽ phân tích ngữ nghĩa của câu hỏi và lựa chọn hàm phù hợp, sau đó trả về một phản hồi ở định dạng JSON, trong đó bao gồm:

* Tên của hàm cần gọi (function\_calling): Đây là tên hàm mà LLM xác định là phù hợp nhất để xử lý câu hỏi.
* Các tham số cần thiết (parameters): Một đối tượng JSON chứa dữ liệu đầu vào cần thiết cho hàm đó.

Bằng cách này, hệ thống có thể tự động điều hướng đến đúng chức năng (hàm) dựa trên yêu cầu của người dùng. Đồng thời giảm độ phức tạp của logic xử lý trong Manager Agent, vì trách nhiệm chọn hàm được giao cho LLM. Hình 39 là minh họa cách hoạt động của một agent management.

A diagram of a company

Description automatically generated

Hình 39. Minh họa chức năng của agent management.

## Triển khai trình quản lý bộ nhớ chatbot

Trình quản lý bộ nhớ chatbot MemGPT là một giải pháp tiên tiến trong việc xây dựng chatbot, đặc biệt phù hợp cho các ứng dụng đòi hỏi quản lý bối cảnh dài hạn và phản hồi chính xác. Nhờ khả năng tóm tắt và lưu trữ thông tin, MemGPT giúp chatbot duy trì ngữ cảnh hội thoại trong giới hạn token và ghi nhớ thông tin quan trọng từ lịch sử tương tác trước. Điều này tạo nên trải nghiệm cá nhân hóa, hiệu quả hơn trong việc giải quyết các truy vấn phức tạp như tư vấn tuyển sinh. Lợi ích lớn nhất của MemGPT là kết hợp giữa tính chính xác, tính liên tục của ngữ cảnh và khả năng thích ứng với người dùng, giúp cải thiện cả hiệu suất và sự hài lòng trong giao tiếp.

Trong hệ thống chatbot này, chúng tôi sử dụng mô hình kiến trúc của MemGPT để làm bộ nhớ lưu trữ ngữ cảnh cho hệ thống chatbot trong một phiên trò chuyện của người dùng. Kiến trúc của MemGPT được thể hiện trong hình 40, bao gồm các thành phần và miêu tả chức năng như sau:

A diagram of a workflow

Description automatically generated

Hình 40. Mô hình kiến trúc hệ thống MemGPT [91].

Kiến trúc hệ thống của MemGPT được sử dụng lại bao gồm đầy đủ các thành phần trong hệ thống. Gồm các thành phần chính như sau [91]:

* LLM Finite Context Window: Đây là vùng làm việc chính của LLM, giới hạn bởi số lượng token tối đa (ví dụ: 8k token). Bao gồm hai phần chính:
  + Prompt Tokens: Các thông tin đầu vào để LLM thực hiện suy luận.
  + Completion Tokens: Đầu ra của LLM sau khi xử lý.
* System Instructions và MemGPT System Prompt (Các hướng dẫn hệ thống và MemGPT System Prompt), trong đó:
  + System Instructions: Các chỉ dẫn cốt lõi được hệ thống định nghĩa trước, không thay đổi, nhằm đảm bảo LLM hoạt động theo một cách thống nhất.
  + MemGPT System Prompt: Một vùng dữ liệu chỉ đọc, cung cấp ngữ cảnh nền tảng và các hướng dẫn cơ bản để hệ thống vận hành.
* Working Context và FIFO Queue:
  + Working Context: Là vùng lưu trữ tạm thời trong LLM để xử lý thông tin ngắn hạn. Có thể đọc-ghi, cho phép thêm thông tin mới thông qua các hàm xử lý (Functions).
  + FIFO Queue (First-In-First-Out): Một hàng đợi đơn giản, nơi thông tin được thêm vào theo thứ tự và dữ liệu cũ nhất sẽ bị đẩy ra khi đạt giới hạn. Dùng để quản lý luồng thông tin trong ngữ cảnh của LLM, giảm nguy cơ quá tải.
* Output Buffer: Là khu vực lưu trữ tạm thời kết quả (Completion Tokens) trước khi trả về cho người dùng hoặc các thành phần khác. Đảm bảo đầu ra được định dạng đúng và sẵn sàng sử dụng.
* Archival Storage (Lưu trữ dài hạn): Mục đích là để lưu trữ dữ liệu quan trọng để truy xuất trong tương lai. Ghi và truy vấn thông tin thông qua các Function Executor, cho phép lưu giữ thông tin không phụ thuộc vào ngữ cảnh hiện tại.
* Recall Storage (Lưu trữ ngắn hạn/trung hạn): Dùng để giữ thông tin cần truy cập nhanh hoặc thường xuyên trong quá trình xử lý. Kết nối chặt chẽ với Queue Manager để đảm bảo dữ liệu ngắn hạn được xử lý hiệu quả.
* Function Executor: Là thành phần trung gian chịu trách nhiệm ghi dữ liệu vào Archival Storage, truy xuất dữ liệu cần thiết từ các bộ lưu trữ, chuyển dữ liệu qua lại giữa các thành phần như FIFO Queue và Recall Storage. Những chức năng này giúp mở rộng khả năng ghi nhớ và xử lý của hệ thống.
* Queue Manager có chức năng: Điều phối hoạt động của FIFO Queue và Recall Storage, giúp đảm bảo thông tin trong FIFO Queue không bị quá tải, đồng thời sắp xếp và quản lý truy xuất dữ liệu trong Recall Storage và hỗ trợ chuyển giao dữ liệu giữa các vùng lưu trữ và LLM.

Cách nó hoạt động và tương tác với hệ thống chatbot như sau [91]:

* Khi nhận yêu cầu từ người dùng, Prompt Tokens được nạp vào Working Context. Tại đây nếu ngữ cảnh quá dài, MemGPT sẽ lưu trữ các thành phần ngữ cảnh cũ hơn vào Archived Storage, đồng thời đẩy ngữ cảnh ít quan trọng hơn vào FIFO Queue để giải phóng dung lượng xử lý.
* Trong quá trình xử lý (Processing Stage), MemGPT sẽ tự động kích hoạt Function Executor. Khi thông tin cần thiết không còn nằm trong ngữ cảnh hiện tại (Working Context), Function Executor sẽ truy xuất thông tin từ Recall Storage hoặc Archival Storage, đồng thời ghi thông tin cần thiết trở lại ngữ cảnh làm việc. Dữ liệu không còn quan trọng trong FIFO Queue được xóa hoặc lưu dài hạn vào Archival Storage.
* Khi trả lời người dùng (Output Stage), Kết quả (Completion Tokens) sau khi được xử lý xong sẽ được lưu vào Output Buffer, dữ liệu bổ sung từ Archival Storage hoặc Recall Storage được ghi nhớ cho các lần truy vấn sau này và trả lời cuối cùng được gửi đến người dùng qua giao diện chatbot.

Những chức năng của MemGPT giúp tối ưu hóa bối cảnh hội thoại, giữ thông tin quan trọng trong phạm vi token cho phép từ hệ thống. Nhờ vào khả năng tóm tắt và cắt giảm thông minh, MemGPT không chỉ giúp duy trì thông tin quan trọng mà còn giảm thiểu tình trạng quá tải bộ nhớ, đảm bảo rằng hệ thống luôn hoạt động hiệu quả. Điều này giúp chatbot duy trì sự liên tục trong cuộc trò chuyện, ngay cả khi số lượng tin nhắn hoặc thông tin người dùng ngày càng tăng, từ đó cung cấp các phản hồi chính xác và phù hợp hơn trong từng tình huống.

## Cơ sở dữ liệu người dùng

Trước khi bắt đầu một đoạn chat, yêu cầu người dùng nhập vào họ tên, số điện thoại để có thể bắt đầu đoạn chat. Sau khi nhập dữ liệu này thì được lưu vào MongoDB. Bằng cách này, hệ thống có thể quản lý thông tin người dùng một cách hiệu quả, đảm bảo rằng mỗi phiên trò chuyện đều được liên kết với một hồ sơ cụ thể. Việc lưu trữ thông tin trong MongoDB giúp dễ dàng truy xuất và phân tích dữ liệu người dùng, hỗ trợ các chiến lược chăm sóc khách hàng và tối ưu hóa trải nghiệm. Hình 41 minh hoạ cách thông tin người dùng được thu thập cho mục đích cá nhân hóa trải nghiệm hệ thống.

A diagram of a flowchart

Description automatically generated

Hình 41. Dữ liệu người dùng được thu thập và sử dụng cho mục đích cá nhân hóa.

Trong suốt quá trình diễn ra hành động hỏi đáp giữa người dùng và hệ thống, các thông tin của người dùng liên tục được thu thập và gửi tới hệ thống xử lý chatbot, nhằm cung cấp thêm thông tin và nhận diện được các khía cạnh mà người dùng quan tâm. Điều này giúp chatbot trả lời thân thiện, và xoay quanh vấn đề của người dùng và làm tăng thêm tính hiểu người dùng.

## Triển khai hệ thống chatbot hoàn chỉnh

Trong dự án này, chúng tôi sử dụng Framework Chainlit để tận dụng những công cụ và tính năng mạnh mẽ để xây dựng giao diện người dùng (UI). Với mục tiêu tối ưu hóa trải nghiệm người dùng, Chainlit cung cấp các thành phần giao diện, các tùy chọn tuỳ chỉnh giao diện, và khả năng tích hợp dễ dàng với các nền tảng khác. Việc có thể triển khai một cách nhanh chóng bằng công cụ giúp tiết kiệm thời gian xây dựng, hơn thế nữa có thể dánh nhiều thời gian hơn cho việc thiết kế, triển khai và tối ưu hóa hệ thống chatbot một cách hiệu quả hơn. Các hình ảnh dưới đây là kết quả minh hoạ triển khai hệ thống chatbot hỏi đáp tuyển sinh trường đại học Đông Á sử dụng LLM.

A screenshot of a computer

Description automatically generated

Hình 42. Minh họa bắt đầu trò chuyện sau skhi triển khai chatbot hoàn chỉnh.

A screenshot of a computer

Description automatically generated

A text on a white background

Description automatically generated

Hình 43. Minh họa kết quả triển khai hệ thống hoàn chỉnh.

A screenshot of a chat

Description automatically generated

Hình 44. . Minh họa kết quả triển khai hệ thống hoàn chỉnh.s

A screenshot of a chat

Description automatically generated

Hình 45. Minh họa kết quả triển khai hệ thống hoàn chỉnh.

A screenshot of a computer

Description automatically generatedA screenshot of a chat

Description automatically generated

Hình 46. Minh họa kết quả triển khai hệ thống hoàn chỉnh.

A screenshot of a chat

Description automatically generated

Hình 47. Minh họa kết quả triển khai hệ thống hoàn chỉnh.

Như vậy, đồ án xây dựng chatbot tuyển sinh trong phạm vi Trường Đại học Đông Á sử dụng LLM đã được hoàn thành. Trong báo cáo này chúng tôi đã trình bày về cơ sở lý thuyết liên quan, từ những kiến thức nền móng cho tới những chủ đề phức tạp xoay quanh vấn đề xây dựng hệ thống chatbot. Cụ thể trong chương 1, chúng tôi đã đi từ những phương pháp học máy, các mô hình học sâu là cảm hứng cho những mô hình LLM. Sau đó chúng tôi cũng đề cập đến các phần mềm và công cụ hỗ trợ, đồng thời nêu lên lợi ích và cách triển khai cho những bộ công cụ hỗ trợ ấy. Trong chương 2, chúng tôi đã nêu rõ phương pháp đề xuất của chúng tôi. Bằng cách đưa ra một chiến lược tối ưu ngay từ lúc bắt đầu, chúng tôi đã triển khai thành công các thành phần quan trọng cần được xây dựng để đáp ứng cho đề tài của chúng tôi. Cuối cùng, trong chương 3 chúng tôi đã đi từ những bước đầu cho đến bước hoàn thành mà chúng tôi đã thực hiện trong quá trình xây dựng đồ án. Cụ thể chúng tôi đã nêu rõ cách dữ liệu cần được xử lý trước khi đưa vào hệ thống chatbot, cách vận dụng các công cụ để trích xuất thông tin và điều khiển chúng hoạt động, đi cùng với đó chúng tôi đã nêu trọng tâm cách triển khai các thành phần của hệ thống chatbot, đồng thời miêu tả cách nó hoạt động cùng với chức năng đi kèm. Để xem thêm kết quả của hệ thống chatbot, chúng tôi trình bày các thử nghiệm của chúng tôi trong phần phụ lục để có thể tham khảo thêm.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

#### Kết luận chung

Đồ án *"Xây dựng chatbot tuyển sinh trường Đại học Đông Á sử dụng LLM"* đã giải quyết thành công bài toán xây dựng một hệ thống hỗ trợ tuyển sinh thông minh, tận dụng các tiến bộ trong công nghệ xử lý ngôn ngữ tự nhiên và học máy. Với tầm quan trọng ngày càng tăng của chuyển đổi số trong giáo dục, việc triển khai chatbot không chỉ giúp tự động hóa quy trình tư vấn tuyển sinh mà còn nâng cao trải nghiệm người dùng thông qua phản hồi nhanh chóng và chính xác.

Đồ án đã không chỉ xây dựng một hệ thống chatbot hoàn chỉnh mà còn triển khai một nền tảng lý thuyết vững chắc để hỗ trợ các khía cạnh kỹ thuật phức tạp. Các công nghệ như lưu trữ cơ sở dữ liệu vector, mô hình embedding, quản lý agents, và kỹ thuật truy xuất thông tin đều được tích hợp và vận hành mượt mà trong hệ thống. Hệ thống không chỉ đáp ứng được yêu cầu ban đầu mà còn chứng minh tiềm năng ứng dụng thực tiễn cao, mở ra hướng đi mới trong việc tự động hóa quản lý giáo dục tại trường Đại học Đông Á.

Quan trọng hơn, đồ án đã đạt được mục tiêu chính: thiết kế một chatbot không chỉ cung cấp thông tin tuyển sinh mà còn hỗ trợ giải đáp câu hỏi tuyển sinh, các vấn đề ngoài phạm vi và quản lý dữ liệu người dùng. Các thành quả này khẳng định tính hiệu quả và khả năng ứng dụng rộng rãi của hệ thống, góp phần cải thiện chất lượng dịch vụ của nhà trường.

#### Các kết quả đạt được

Với sự triển khai toàn diện, đồ án đã đạt được nhiều kết quả nổi bật. Về mặt lý thuyết, đồ án đã nghiên cứu và áp dụng các công nghệ hiện đại như mô hình Transformer, LLM, và kỹ thuật Prompt Engineering, đồng thời xây dựng nền tảng lý thuyết chặt chẽ về xử lý ngôn ngữ tự nhiên và học máy. Những nền tảng này không chỉ giúp phát triển hệ thống chatbot mà còn cung cấp một khung tham chiếu cho các nghiên cứu tương lai.

Về thực tiễn, chatbot đã được thiết kế và triển khai với nhiều tính năng quan trọng. Hệ thống sử dụng cơ sở dữ liệu vector để lưu trữ và truy xuất thông tin nhanh chóng, tích hợp các thuật toán tiên tiến để tăng độ chính xác của kết quả. Các tác nhân như Q&A Agent, và Out-of-Domain Agent đã hoạt động hiệu quả, mang đến trải nghiệm mượt mà cho người dùng. Đặc biệt, hệ thống quản lý lời nhắc (Prompt Engineering) và trình quản lý tác nhân giúp chatbot có khả năng phân luồng câu hỏi và xử lý linh hoạt các tình huống khác nhau.

Hệ thống chatbot đã được đánh giá hiệu quả thông qua các chỉ số chính như độ chính xác của phản hồi, thời gian truy vấn, và tính ổn định. Các tính năng như đặt lịch hẹn tư vấn hay quản lý dữ liệu người dùng đều hoạt động chính xác, phù hợp với nhu cầu thực tế của bộ phận tuyển sinh.

#### Phương hướng phát triển tiếp theo

Mặc dù đạt được những thành quả đáng kể, hệ thống chatbot vẫn còn nhiều tiềm năng phát triển nhằm nâng cao hiệu quả và mở rộng phạm vi ứng dụng. Một định hướng quan trọng là mở rộng khả năng hỗ trợ của chatbot sang các ngành học hoặc bộ phận khác trong nhà trường. Điều này có thể bao gồm việc tư vấn học thuật, hỗ trợ thủ tục hành chính, hoặc cung cấp thông tin chi tiết cho từng ngành học và phòng ban, giúp chatbot trở thành công cụ toàn diện phục vụ cho cả sinh viên và cán bộ nhà trường.

Hệ thống cũng cần được tối ưu hóa về hiệu suất để đảm bảo khả năng xử lý đồng thời cho nhiều người dùng. Việc áp dụng các thuật toán retrieval và rerank tiên tiến hơn có thể giúp cải thiện tốc độ phản hồi và độ chính xác của kết quả. Ngoài ra, tích hợp chatbot vào các nền tảng phổ biến như mạng xã hội, ứng dụng di động hoặc hệ thống nội bộ sẽ tăng cường khả năng tiếp cận người dùng, tạo ra trải nghiệm nhất quán trên nhiều môi trường.

Một yếu tố quan trọng khác là cải thiện khả năng tự động học hỏi từ dữ liệu mới. Điều này sẽ giúp chatbot thích nghi với những thay đổi trong thông tin tuyển sinh hoặc quy trình của trường mà không cần can thiệp thủ công. Việc tích hợp công cụ phân tích dữ liệu người dùng để dự đoán xu hướng hoặc đưa ra gợi ý thông minh cũng sẽ gia tăng giá trị của hệ thống.

Ngoài ra, để đảm bảo hệ thống hoạt động an toàn và bền vững, cần tiếp tục nghiên cứu và áp dụng các biện pháp bảo mật dữ liệu cá nhân, đặc biệt khi triển khai trên các môi trường công cộng. Những định hướng này không chỉ giúp hoàn thiện hệ thống hiện tại mà còn tạo tiền đề cho việc mở rộng nghiên cứu và ứng dụng trong các lĩnh vực khác, góp phần vào sự phát triển toàn diện của nhà trường.

# CÁC TÀI LIỆU THAM KHẢO

[1] S. C. Mana, G. Kalaiarasi, Y. R, L. S. Helen, and R. Senthamil Selvi, “Application of Machine Learning in Healthcare: An Analysis,” in *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2022, pp. 1611–1615. doi: 10.1109/ICESC54411.2022.9885296.

[2] S. Das, A. dey, A. Pal, and N. Roy, “Applications of Artificial Intelligence in Machine Learning: Review and Prospect,” *Int J Comput Appl*, vol. 115, pp. 31–41, Nov. 2015, doi: 10.5120/20182-2402.

[3] M. Abedi, I. Alshybani, M. R. B. Shahadat, and M. Murillo, “Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education,” Nov. 2023. doi: 10.48550/arXiv.2309.13059.

[4] B. M. Gurusamy *et al.*, “An analysis of large language models: their impact and potential applications,” *Knowl Inf Syst*, vol. 66, pp. 1–24, Nov. 2024, doi: 10.1007/s10115-024-02120-8.

[5] S. Brightwood, “Implementing AI Chatbots for Real-Time Supply Chain Monitoring and Risk Management,” Nov. 2024.

[6] Y. Zhang, S. Kumar, D. Paikar, H. Ali, and K. Vutukuri, “KatzBot: Revolutionizing Academic Chatbot for Enhanced Communication,” Nov. 2024. doi: 10.13140/RG.2.2.12847.96162.

[7] M. J. El Naqa Issam and Murphy, “What Is Machine Learning?,” in *Machine Learning in Radiation Oncology: Theory and Applications*, R. and M. M. J. El Naqa Issam and Li, Ed., Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3\_1.

[8] V. Nasteski, “An overview of the supervised machine learning methods,” *HORIZONS.B*, vol. 4, pp. 51–62, Nov. 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.

[9] M. Mohri, “Foundations of machine learning,” 2018, *MIT press*.

[10] Wikipedia, “Supervised learning,” p. How supervised learning algorithms work, Nov. 2024.

[11] Z. Ghahramani, “Unsupervised Learning,” in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, U. and R. G. Bousquet Olivier and von Luxburg, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 72–112. doi: 10.1007/978-3-540-28650-9\_5.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[13] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021, doi: 10.1007/s12525-021-00475-2.

[14] K. Moore *et al.*, “A comprehensive solution to retrieval-based chatbot construction,” *CoRR*, vol. abs/2106.06139, 2021, [Online]. Available: https://arxiv.org/abs/2106.06139

[15] P. Domingos, “A few useful things to know about machine learning,” *Commun ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[16] nhi.lty, “Học máy (Machine Learning) – Khái niệm, phương pháp và ứng dụng,” 2024.

[17] T. Group, “Machine learning là gì? Ứng dụng phổ biến của Machine learning,” 2024.

[18] hiroshidigiuni, “Top 10 ứng dụng của Machine Learning,” 2024.

[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] M. Abedi, I. Alshybani, M. R. B. Shahadat, and M. Murillo, “Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education,” Nov. 2023. doi: 10.48550/arXiv.2309.13059.

[21] Amazon, “Điểm khác biệt giữa máy học và học sâu là gì?,” 2024, *https://aws.amazon.com/vi/compare/the-difference-between-machine-learning-and-deep-learning/*.

[22] S. Shanmuganathan, “Artificial Neural Network Modelling: An Introduction,” in *Artificial Neural Network Modelling*, S. Shanmuganathan Subana and Samarasinghe, Ed., Cham: Springer International Publishing, 2016, pp. 1–14. doi: 10.1007/978-3-319-28495-8\_1.

[23] B. M. Gurusamy *et al.*, “An analysis of large language models: their impact and potential applications,” *Knowl Inf Syst*, vol. 66, pp. 1–24, Nov. 2024, doi: 10.1007/s10115-024-02120-8.

[24] H. Pothina and K. V Nagaraja, “Artificial Neural Network and Math Behind It,” in *Smart Trends in Computing and Communications*, Y.-D. Zhang, T. Senjyu, C. So-In, and A. Joshi, Eds., Singapore: Springer Nature Singapore, 2023, pp. 205–221.

[25] S. Haykin, *Neural networks and learning machines, 3/E*. Pearson Education India, 2009.

[26] R. Raj, “Components of an Artificial Neural Network,” 2024.

[27] A. D. Rasamoelina, F. Adjailia, and P. Sinčák, “A Review of Activation Function for Artificial Neural Network,” in *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2020, pp. 281–286. doi: 10.1109/SAMI48414.2020.9108717.

[28] Wikipedia, “Activation function,” p. https://en.wikipedia.org/wiki/Activation\_function, Nov. 2024.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[30] T. Babs, “The Mathematics of Neural Networks,” 2018.

[31] L. F. Guilhoto, “An overview of artificial neural networks for mathematicians,” *Univ. Chicago*, 2018.

[32] J. Schmidhuber, “A ‘self-referential’weight matrix,” in *ICANN’93: Proceedings of the International Conference on Artificial Neural Networks Amsterdam, The Netherlands 13–16 September 1993 3*, 1993, pp. 446–450.

[33] S. Amari, “Backpropagation and stochastic gradient descent method,” *Neurocomputing*, vol. 5, no. 4–5, pp. 185–196, 1993.

[34] J. Mao and A. K. Jain, “Artificial neural networks for feature extraction and multivariate data projection,” *IEEE Trans Neural Netw*, vol. 6, no. 2, pp. 296–317, 1995.

[35] indrasingh52, “Layers in Artificial Neural Networks (ANN),” 2024.

[36] anshumanm2fja, “What is Forward Propagation in Neural Networks?,” 2024.

[37] T. Beysolow II, “What Is Natural Language Processing?,” in *Applied Natural Language Processing with Python : Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing* , Berkeley, CA: Apress, 2018, pp. 1–12. doi: 10.1007/978-1-4842-3733-5\_1.

[38] J. H. Cole Stryker, “What is NLP?,” 2024.

[39] A. Vaswani, “Attention is all you need,” *Adv Neural Inf Process Syst*, 2017.

[40] 4technews.net, “Ứng dụng xử lý ngôn ngữ tự nhiên trong kinh doanh,” 2024.

[41] J. Alammar, “The Illustrated Transformer,” 2024.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” *CoRR*, vol. abs/1904.09675, 2019, [Online]. Available: http://arxiv.org/abs/1904.09675

[43] M. Sharma, “What is Add & Norm, as quick as possible?,” 2024.

[44] S. Targ, D. Almeida, and K. Lyman, “Resnet in Resnet: Generalizing Residual Architectures,” *CoRR*, vol. abs/1603.08029, 2016, [Online]. Available: http://arxiv.org/abs/1603.08029

[45] N. T. H. Trần Hồng Việt, “TRANSFORMERS MODEL AND APPLY IN NATURAL LANGUAGE PROCESSING,” *Tài liệu NCKH GOV*, 2017.

[46] Cloudflare, “Mô hình ngôn ngữ lớn (LLM) là gì?,” 2024.

[47] R. Dilts, R. Bandler, J. Grinder, and J. DeLozier, “What is NLP,” *Режим доступу: http://www. nlpu. com/NLPU\_WhatIsNLP. html*, 1999.

[48] M. Gothankar, “What is LLM & How to Build Your Own Large Language Models?,” 2024.

[49] J. C. Luna, “What is an LLM? A Guide on Large Language Models and How They Work,” 2024.

[50] Nithyashree, “What is Chunking in Natural Language processing?,” 2021.

[51] D. Raikar, “How Chunking Strategies Work: Paragraph, Sentence and Smart Techniques,” 2024.

[52] T. ML, “Embedding,” 2024.

[53] P. Phan, “ChatGPT Series 5: Tìm hiểu về Retrieval Augmented Generation (RAG),” 2023.

[54] S. Ahmed, “What is Retrieval-Augmented Generation(RAG) in LLM and How it works?,” 2024.

[55] D. Shah, “Reciprocal Rank Fusion (RRF) explained in 4 mins — How to score results form multiple retrieval methods in RAG,” 2024.

[56] Z. Rackauckas, “RAG-Fusion: a New Take on Retrieval-Augmented Generation,” 2024. doi: https://doi.org/10.5121/ijnlc.2024.13103.

[57] T. Q. Huy, “LLM agent: Tổng quan về LLM agent,” 2024.

[58] T. Palaniyappan, “How LLM Agent works?,” 2024.

[59] L. Weng, “LLM Powered Autonomous Agents,” 2023.

[60] R. Graph, “Prompt Engineering,” 2024.

[61] L. Phan, “Tất tần tật những kĩ thuật Prompt Engineering hữu ích nhất cho chatGPT,” 2024.

[62] S. Ahmed, “LLM Prompt Engineering for Beginners: What It Is and How to Get Started,” 2024.

[63] J. Liu *et al.*, “Generated Knowledge Prompting for Commonsense Reasoning,” *CoRR*, vol. abs/2110.08387, 2021, [Online]. Available: https://arxiv.org/abs/2110.08387

[64] M. T. School, “Ưu điểm vượt trội của lập trình trí tuệ nhân tạo python?,” 2024.

[65] Wikipedia, “Python (ngôn ngữ lập trình),” 2024.

[66] C. Team, “Chainlit,” 2024.

[67] T. Rasul, “Building a Chatbot Application with Chainlit and LangChain,” 2024.

[68] Milvus, “Milvus,” 2024.

[69] Milvus, “introducing-milvus-lite-start-building-a-genai-application-in-seconds,” 2024.

[70] MongoDB, “MongoDB,” 2024.

[71] M. Narang, “MemGPT with Real-life Example: Bridging the Gap Between AI and OS,” 2024.

[72] M. Raspuzzi, “what is memGPT?,” 2023.

[73] C. Packer *et al.*, “MemGPT: Towards LLMs as Operating Systems,” 2024. [Online]. Available: https://arxiv.org/abs/2310.08560

[74] M. L. Expert, “MemGPT - Unlimited Context (Memory) for LLMs,” 2024.

[75] L. Teixeira, “MemGPT: Unlimited Memory without Token Constraints for Generative AI Platforms, like GPT-4, LaMDA, PaLM, LLAMA, CLAUDE, and others,” 2023.

[76] donga, “ĐẠI HỌC ĐÔNG Á,” 2024.

[77] T. sinh Đại học Đông Á, “Tuyển sinh Đại học Đông Á Official 2024,” 2024.

[78] R. Khawaja, “Demystifying Embeddings 101: The Foundation of Large Language Models,” 2023.

[79] novita.ai, “What is LLM Embeddings: All You Need To Know,” 2024.

[80] Sbert, “Pretrained Models: Sentence Transformers,” 2024.

[81] Deepinfra, “sentence-transformers/all-MiniLM-L6-v2,” 2024.

[82] Sbert, “Pretrained Models: Multilingual Models,” 2024.

[83] H. Face, “Alibaba-NLP/gte-multilingual-base,” 2024.

[84] huggingface, “BAAI/bge-m3,” 2024. [Online]. Available: https://huggingface.co/BAAI/bge-m3

[85] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” 2024. [Online]. Available: https://arxiv.org/abs/2402.03216

[86] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, “Nomic Embed: Training a Reproducible Long Context Text Embedder,” 2024.

[87] H. Face, “nomic-ai/nomic-embed-text-v1.5,” 2024.

[88] A. Trivedi, “Top 10 Open-Source LLMs for 2025 and Their Uses,” 2024.

[89] J. Sumrak, “7 LLM use cases and applications in 2024,” 2024.

[90] J. Fell, “Stanford just released its annual AI Index report. Here’s what it reveals,” 2024.

[91] MemGPT, “MemGPT,” 2024.

--HẾT--

1. Một số tài liệu sử dụng thuật ngữ *residual connection*. Trong tài liệu của [43] tác giả sử dụng *skip connection* thay vì *residual connection*. Tuy nhiên hai thuật ngữ này đều chỉ chung cùng mục đích trong trường hợp này. [↑](#footnote-ref-1)
2. Vấn đề biến mất đạo hàm (vanishing gradient) xảy ra khi các gradient trở nên rất nhỏ, làm cho việc cập nhật trọng số không thay đổi. [↑](#footnote-ref-2)