

Xây dựng Chatbot ứng dụng RAG hỗ trợ sinh viên Trường Đại học Bách Khoa Hà Nội với khả năng tạo biểu mẫu

Báo cáo Bài tập lớn

sinh viên: Dương Quang Hưng - 20225001
Môn học: Nhập môn Học máy và khai phá dữ liệu
Giảng viên: PGS.TS. Phạm Văn Hải



Đại học Bách Khoa Hà Nội
Khoa Công nghệ Thông tin và Truyền thông

Ngày 9 tháng 2 năm 2026

Mục lục

Tóm tắt	3
1 Thu thập và xử lý dữ liệu	4
1.1 Thu thập tài liệu của trường HUST	4
1.2 Thu thập biểu mẫu, đơn	4
1.2.1 Thu thập dữ liệu trên web trường	4
1.2.2 Chuẩn hóa dữ liệu cho hệ thống	5
2 Xây dựng bộ Retrieval	7
2.1 RAG	7
2.2 Mô hình BGE-M3	8
2.3 Docling	8
2.3.1 Khó khăn khi xử lý dữ liệu	8
2.3.2 Docling và giải pháp xử lý bảng biểu	9
2.4 Context Aware Chunking	9
2.4.1 Kỹ thuật chunking thông thường và hạn chế	9
2.4.2 Kỹ thuật Context Aware Chunking với Docling	10
2.5 Reranker	10
2.5.1 Vai trò của Reranker trong RAG	10
2.5.2 BGE-M3 Reranker	11
2.6 Finetune	11
2.6.1 Thách thức về dữ liệu huấn luyện	11
2.6.2 Chiến lược sinh dữ liệu bằng LLM	12
2.6.3 Finetune Embedder và Reranker	12
3 Xây dựng ChatBot với RAG	14
3.1 End-to-end Workflow	14
3.2 Một số Prompt Engineering chính	16
3.2.1 Prompt phát hiện ý định tạo biểu mẫu	16
3.2.2 Prompt trích xuất thông tin từ hội thoại	17
3.2.3 Prompt tạo phản hồi hội thoại với ngữ cảnh RAG	17
3.2.4 Prompt tóm tắt hội thoại	19
3.3 Một số cuộc hội thoại với chatbot	19
3.3.1 Q&A về đầu ra ngoại ngữ, thông tin nằm trong tài liệu của trường	19
3.3.2 Đoạn chat hỗ trợ sinh viên điền mẫu xin cấp học bổng	20
3.3.3 Đoạn chat tạo giấy chứng nhận học bổng	21

Kết luận	23
Tài liệu tham khảo	24

Tóm tắt

Báo cáo này trình bày một hệ thống AI ChatBot hỗ trợ sinh viên Đại học Bách Khoa Hà Nội (HUST) sử dụng kiến trúc RAG (Retrieval-Augmented Generation) nâng cao. Hệ thống không chỉ trả lời các câu hỏi về quy chế đào tạo mà còn tự động tạo các biểu mẫu hành chính dựa trên hội thoại với sinh viên. Báo cáo chia thành 3 chương chính: (1) Thu thập và xử lý dữ liệu từ các nguồn của trường, (2) Xây dựng hệ thống RAG với fine-tuning và xử lý văn bản chuyên sâu, (3) Thiết kế ChatBot thông minh với khả năng đa nhiệm.

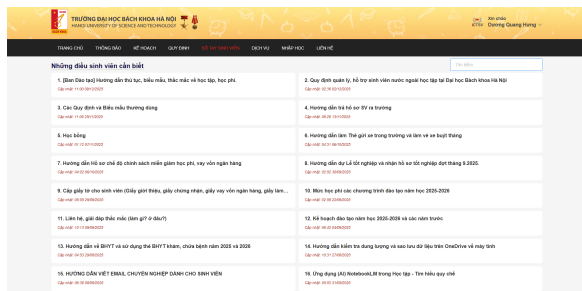
Từ khóa: RAG, Fine-tuning, BGE-M3, Docling, Context Aware Chunking, ChatBot, HUST

1 Thu thập và xử lý dữ liệu

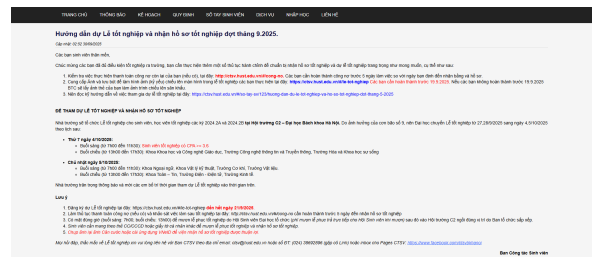
1.1 Thu thập tài liệu của trường HUST

Các tài liệu quy chế, thông báo, hướng dẫn của Trường Đại học Bách Khoa Hà Nội được thu thập từ các nguồn chính thức trên website sở tay sinh viên. Quá trình thu thập:

- **Nguồn dữ liệu:** Các mục trong **trang web sở tay sinh viên** và các trang con trong mỗi mục
- **Định dạng tài liệu:** Công cụ sẽ crawl các tài liệu PDF, DOCX tìm thấy ở trong các link
- **Phương pháp thu thập:** Crawl tự động kết hợp với thu thập thủ công từ các trang web chính thức
- **Lưu trữ:** Các tài liệu sẽ được lưu trữ tập trung ở trên **Google Drive**



(a) Các mục trong sổ tay sinh viên



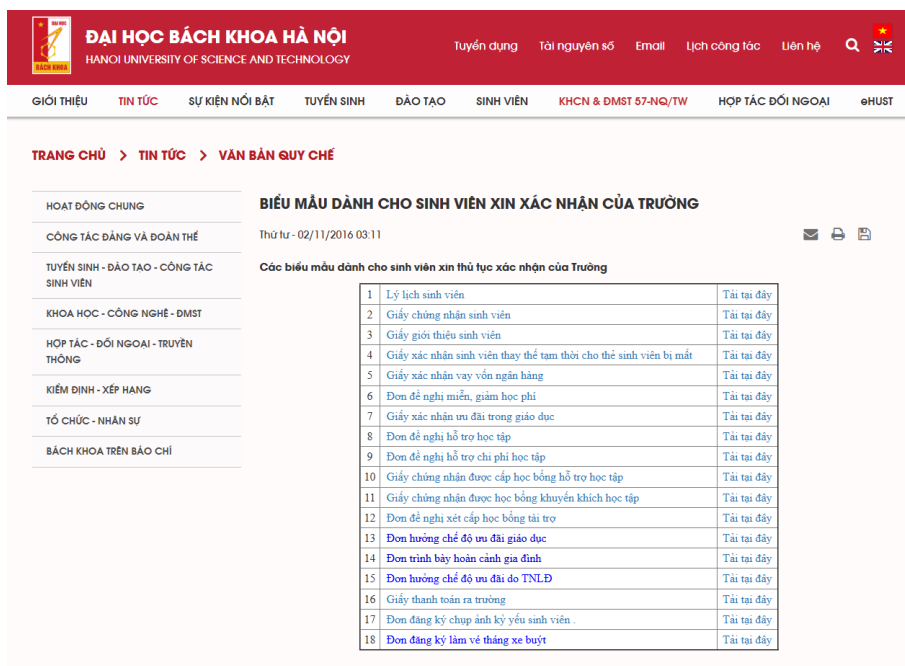
(b) Bên trong một mục và các đường dẫn con

1.2 Thu thập biểu mẫu, đơn

1.2.1 Thu thập dữ liệu trên web trường

Các biểu mẫu hành chính được thu thập từ các trang web cung cấp biểu mẫu của trường:

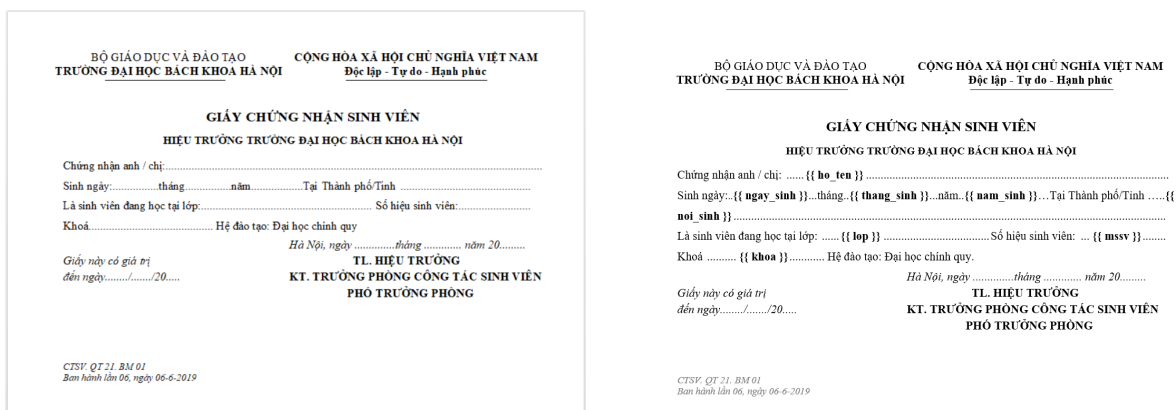
- **Nguồn dữ liệu:** **Trang tổng hợp các biểu mẫu**
- **Loại biểu mẫu thu thập:** Đơn xin miễn học phần, đăng ký tốt nghiệp, xin cấp bằng điểm, đơn xin bảo lưu kết quả, đơn xin chuyển ngành
- **Định dạng gốc:** DOCX với các trường điền sẵn và định dạng bảng
- **Phương pháp:** Tải trực tiếp từ các biểu mẫu và lưu trữ trên **Google Drive**



Hình 2: Trang web cung cấp biểu mẫu DOCX

1.2.2 Chuẩn hóa dữ liệu cho hệ thống

Tạo template DOCX: Giữ lại định dạng gốc làm template, đánh dấu các vị trí điền dữ liệu bằng placeholder để lập template chung cho hệ thống xử lý (phù hợp với công cụ `docxptl` trong python)



(a) Form DOCX ban đầu

(b) Form DOCX template

Hình 3: Chuẩn hóa biểu mẫu DOCX sang template chung

Tạo spec file JSON: Mỗi biểu mẫu có file JSON mô tả chi tiết các trường, quy tắc validate, định dạng dữ liệu

```

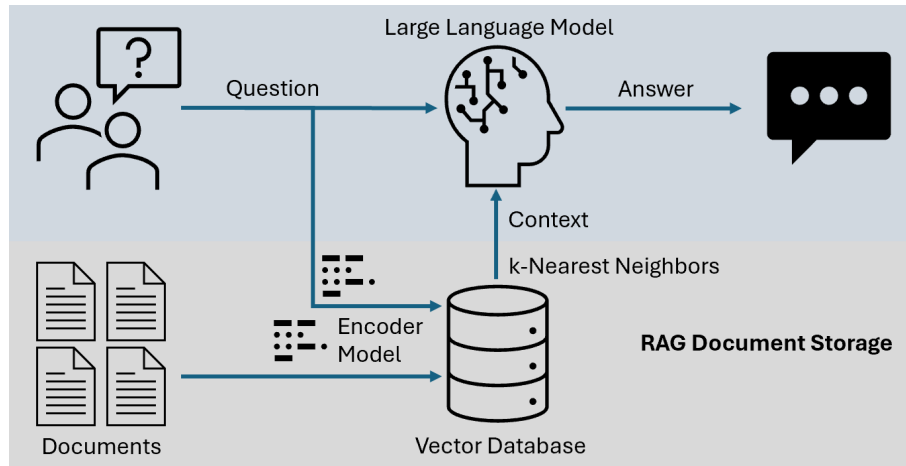
chung_nhan_vv-spec.json - Notepad
File Edit Format View Help
{
  "description": "Giấy chứng nhận sinh viên do Trường Đại học Bách khoa Hà Nội cấp, xác nhận người được cấp đang là sinh viên chính quy của Trường.",
  "data": {
    "ho_ten": {
      "label": "Họ và tên sinh viên",
      "required": true,
      "meaning": "Họ và tên đầy đủ của sinh viên theo giấy tờ pháp lý"
    },
    "ngay_sinh": {
      "label": "Ngày sinh",
      "required": true,
      "meaning": "Ngày sinh của sinh viên (ngày trong tháng)",
      "pattern": "^\\d{1,2}$"
    },
    "thang_sinh": {
      "label": "Tháng sinh",
      "required": true,
      "meaning": "Tháng sinh của sinh viên",
      "pattern": "^\\d{1,2}$"
    },
    "nam_sinh": {
      "label": "Năm sinh",
      "required": true,
      "meaning": "Năm sinh của sinh viên",
      "pattern": "^\\d{4}$"
    },
    "noi_sinh": {
      "label": "Tỉnh/Thành phố nơi sinh",
      "required": true,
      "meaning": "Tỉnh hoặc thành phố nơi sinh của sinh viên theo giấy khai sinh"
    },
    "lop": {
      "label": "Lớp",
      "required": true,
      "meaning": "Lớp hành chính mà sinh viên đang theo học"
    },
    "mssv": {
      "label": "Mã số sinh viên",
      "required": true,
      "meaning": "Mã số sinh viên do Trường Đại học Bách khoa Hà Nội cấp",
      "pattern": "^\\d{6,10}$"
    },
    "khoa": {
      "label": "Khóa",
      "required": true,
      "meaning": "Khóa học của sinh viên (ví dụ: 65, 66, 67)"
    }
  }
}

```

Hình 4: Ví dụ file **spec** của một form

2 Xây dựng bộ Retrieval

2.1 RAG



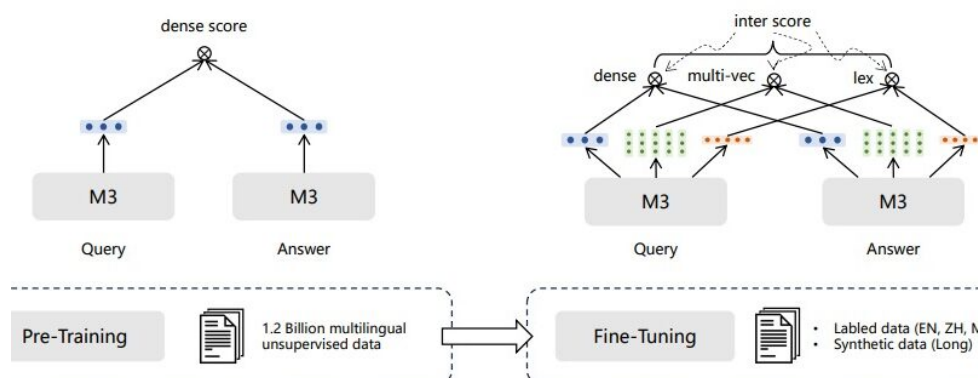
Hình 5: Kiến trúc RAG

RAG (Retrieval-Augmented Generation) là kiến trúc kết hợp giữa hai thành phần chính:

- **Retriever (Bộ truy xuất):** Truy xuất các tài liệu liên quan từ cơ sở tri thức dựa trên truy vấn đầu vào
- **Generator (Bộ sinh văn bản):** Sử dụng thông tin từ các tài liệu truy xuất được để tạo câu trả lời chính xác và ngữ cảnh

Trong hệ thống của chúng tôi, RAG được sử dụng để đảm bảo câu trả lời luôn dựa trên thông tin chính xác từ tài liệu của trường, tránh hiện tượng "hallucination" (tạo thông tin sai) của các mô hình ngôn ngữ lớn.

2.2 Mô hình BGE-M3



Hình 6: BGE-M3

BGE-M3 (BAAI General Embedding) là mô hình embedding đa ngôn ngữ của BAAI, được sử dụng bên trong hệ thống Chatbot của nhóm. Nó phù hợp vì:

- **Đa ngôn ngữ:** Được huấn luyện trên 100+ ngôn ngữ, trong đó tiếng Việt là ngôn ngữ chính với lượng dữ liệu lớn
- **Hỗ trợ đa chức năng:** Có thể sử dụng cho cả retrieval, classification, clustering, và semantic similarity
- **Tối ưu cho văn bản dài:** Xử lý tốt các văn bản dài đặc trưng của tài liệu hành chính và quy chế

2.3 Docling

2.3.1 Khó khăn khi xử lý dữ liệu

- **PDF dạng ảnh:** Một số tài liệu PDF mà trường cung cấp không thể copy chữ mà mang hoàn toàn định dạng ảnh.
- **Ô gộp (merged cells):** Khó trích xuất thông tin chính xác do các ô được merge theo cả chiều ngang và dọc
- **Bảng biểu có bố cục phức tạp:** Các tài liệu PDF của trường chứa nhiều bảng biểu phức tạp, ô merge ngang và dọc

1. Bảng quy đổi tương đương các chứng chỉ tiếng Anh

Bậc cơ sở		CEFR	PEIC	TOEIC 2 kỹ năng	TOEIC 4 kỹ năng				TOEFL		VSTEP	IELTS Academic	APTIS ESOL	Cambridge English Scale & Tests (KET, PET, FCE)
NLNNVN*					Nghe	Độc	Nói	Viết	iBT	iTP				
Bậc 1	Bậc 1.1	A1	Level A1	<250	60÷80	60÷80	50÷60	30÷40	≤ 10	≤ 337	Không có mức tương đương	≤ 2.0	≤ 45	100÷110 (KET)
	Bậc 1.2			250÷345	85÷105	85÷110	70÷80	50÷60	11÷20	338÷360		2.5	46÷65	111÷119 (KET)
Bậc 2	Bậc 2.1	A2	Level 1	350÷395	110÷150	115÷155	90	70÷80	21÷23	361÷400		3.0	66÷75	120÷129 (KET)
	Bậc 2.2			400÷445	155÷195	160÷195	100	90÷100	24÷26	401÷430		3.5	76÷86	130÷135 (KET)
	Bậc 2.3			450÷495	200÷235	200÷235	110	110	27÷29	431÷449		3.5	87÷101	136÷139 (KET)
Bậc 3	Bậc 3.1	B1	Level 2	500÷545	240÷270	240÷270	120	120	30÷33	450÷459	4.0	4.0	102÷115	140÷149 (KET)
	Bậc 3.2			550÷595	275÷295	275÷290	120	120	34÷36	460÷469	4.5	4.0	116÷127	140÷149 (PET)
	Bậc 3.3			600÷645	300÷320	295÷215	130	130	37÷39	470÷479	5.0	4.5	128÷135	150÷155 (PET)
	Bậc 3.4			650÷700	325÷350	220÷340	130	130	40÷42	480÷489	5.0	4.5	136÷143	156÷159 (PET)
	Bậc 3.5			705÷780	355÷395	345÷380	140	140	43÷45	490÷499	5.5	5.0	144÷152	156÷159 (FCE)
Bậc 4		B2	Level 3	≥ 785	≥ 400	≥ 385	≥ 150	≥ 150	≥ 46	≥ 500	≥ 6.0	≥ 5.5	153÷183	≥ 160 (PET/FCE)

Từ viết tắt “NLNNVN”: Khung năng lực ngoại ngữ 6 bậc dùng cho Việt Nam, sử dụng làm cơ sở tham chiếu quy đổi các chứng chỉ khác, trong

Hình 7: Bảng biểu phức tạp

2.3.2 Docling và giải pháp xử lý bảng biểu

Docling là thư viện chuyên dụng để trích xuất nội dung từ tài liệu với khả năng bảo toàn cấu trúc. Docling giúp giải quyết các khó khăn trên bằng cách:

- **Nhận diện bảng chính xác:** Sử dụng thuật toán computer vision kết hợp với heuristic để phát hiện bảng trong PDF
- **Trích xuất cấu trúc bảng:** Giữ nguyên cấu trúc hàng, cột, ô gộp, nhận diện được header và body của bảng
- **Chuyển đổi linh hoạt:** Có thể xuất bảng ra nhiều định dạng như Markdown, HTML, JSON với cấu trúc rõ ràng
- **Xử lý đa định dạng:** Hỗ trợ PDF, DOCX chứa hình ảnh

PHỤ LỤC 2 Bảng quy đổi tương đương các chứng chỉ tiếng Anh

1. Bảng quy đổi tương đương các chứng chỉ tiếng Anh

Bậc cơ sở	Bậc cơ sở	CEFR	PEIC	TOEIC 2 kỹ năng	TOEIC 4 kỹ năng	TOEIC 4 kỹ năng	TOEIC 4 kỹ năng	TOEIC 4 kỹ năng	TOEFL	TOEFL	VSTEP	IELTS Academic	APTIS ESOL	Cambridge English Scale & Tests (KET, PET, FCE)

Bậc 1.1	A1	Level 1	<250	60÷80	60÷80	50÷60	30÷40	≤ 10	≤ 337	Không có mức tương	≤ 2.0	≤ 45	100÷110 (KET)	
Bậc 1.2	A1	Level 1	250÷345	85÷105	85÷110	70÷80	50÷60	11÷20	338÷360	Không có mức tương	2.5	46÷65	111÷119 (KET)	
Bậc 2.1	A2	Level 1	350÷395	110÷150	115÷155	90	70÷80	21÷23	361÷400	Không có mức tương	3.0	66÷75	120÷129 (KET)	
Bậc 2.2	A2	Level 1	400÷445	155÷195	160÷195	100	90÷100	24÷26	401÷430	Không có mức tương	3.5	76÷86	130÷135 (KET)	
Bậc 2.3	A2	Level 1	450÷495	200÷235	200÷235	110	110	27÷29	431÷449	Không có mức tương	3.5	87÷101	136÷139 (KET)	
Bậc 3.1	B1	Level 2	500÷545	240÷270	240÷270	120	120	30÷33	450÷459	4.0	4.0	102÷115	140÷149 (KET)	
Bậc 3.2	B1	Level 2	550÷595	275÷295	275÷290	120	120	34÷36	460÷469	4.5	4.0	116÷127	140÷149 (PET)	
Bậc 3.3	B1	Level 2	600÷645	300÷320	295÷215	130	130	37÷39	470÷479	5.0	4.5	128÷135	150÷155 (PET)	
Bậc 3.4	B1	Level 2	650÷700	325÷350	220÷340	130	130	40÷42	480÷489	5.0	4.5	136÷143	156÷159 (PET)	
Bậc 3.5	B1	Level 2	705÷780	355÷395	345÷380	140	140	43÷45	490÷499	5.5	5.0	144÷152	156÷159 (FCE)	
Bậc 4	B2	Level 3	≥ 785	≥ 400	≥ 385	≥ 150	≥ 150	≥ 46	≥ 500	≥ 6.0	≥ 5.5	153÷183	≥ 160 (PET/FCE)	

Hình 8: Dữ liệu bảng được xử lý bởi Docling

2.4 Context Aware Chunking

2.4.1 Kỹ thuật chunking thông thường và hạn chế

Chunking truyền thống thường sẽ tách văn bản thành các Chunk đơn thuần theo độ dài Chunk, ví dụ: cứ mỗi 800 token thì thành 1 Chunk. Tuy nhiên nó có các hạn chế khi xử lý tài liệu phức tạp:

- **Cắt ngang câu/đoạn văn:** Chia văn bản tại vị trí cố định mà không quan tâm đến ranh giới ngữ nghĩa

- **Mất ngữ cảnh:** Thông tin quan trọng có thể bị chia cắt giữa các chunk, làm mất ý nghĩa tổng thể
- **Không phù hợp với cấu trúc phức tạp:** Đặc biệt với văn bản hành chính có nhiều mục, điều, khoản, bảng biểu
- **Xử lý bảng kém:** Bảng biểu thường bị cắt ngang, làm mất cấu trúc và ý nghĩa của bảng

2.4.2 Kỹ thuật Context Aware Chunking với Docling

Chúng tôi sử dụng Docling để thực hiện chunking dựa trên cấu trúc văn bản:

- **Phân cấp theo cấu trúc:** Chunking dựa vào cấu trúc phân cấp của văn bản (tiêu đề chương, mục, tiểu mục, đoạn văn)
- **Bảo toàn ngữ cảnh:** Mỗi chunk chứa đầy đủ một ý hoàn chỉnh, không cắt ngang nội dung
- **Xử lý bảng biểu riêng biệt:** Mỗi bảng là một chunk riêng, không bị cắt ngang, giữ nguyên toàn bộ cấu trúc

CÔNG BÁO/Số 1003 + 1004/Ngày 25-9-2015	5
Điều 7. Cơ sở vật chất, thiết bị dạy học	
1. Các cơ sở giáo dục có trung tâm, khoa hoặc bộ môn GDQP&AN thực hiện nhiệm vụ GDQP&AN phải có phòng học chuyên dùng, thao trường tổng hợp, có đủ cơ sở vật chất, thiết bị dạy học GDQP&AN theo quy định hiện hành về danh mục thiết bị dạy học tối thiểu môn học GDQP&AN.	
2. Quản lý, sử dụng và bảo quản vũ khí, trang thiết bị GDQP&AN thực hiện theo quy định hiện hành.	
Điều 8. Quản lý môn học và tổ chức dạy, học	
1. Các cơ sở giáo dục nghề nghiệp tổ chức dạy, học GDQP&AN theo kế hoạch đào tạo của cơ sở.	
2. Các cơ sở giáo dục đại học có khoa, bộ môn GDQP&AN tổ chức dạy, học tập trung theo kế hoạch đào tạo của cơ sở.	
3. Các trung tâm GDQP&AN tổ chức dạy, học tập trung theo kế hoạch đào tạo của trung tâm.	
4. Các cơ sở giáo dục tổ chức dạy, học thực hành kỹ thuật, chiến thuật, thuốc nổ và bắn đạn thật phải hợp đồng với cơ quan quân sự địa phương, đơn vị quân đội, công an để bảo đảm tuyệt đối an toàn về người, vũ khí, trang thiết bị.	
5. Lớp học lý thuyết nếu bố trí lớp ghép phải phù hợp với phương pháp dạy học và điều kiện cụ thể của các cơ sở giáo dục, nhưng không quá 150 người; lớp học thực hành không quá 40 người.	

(a) Dữ liệu chữ được phân cấp trong tài liệu

CÔNG BÁO/Số 1003 + 1004/Ngày 25-9-2015	5
Chunk 1 Điều 7. Cơ sở vật chất, thiết bị dạy học	
1. Các cơ sở giáo dục có trung tâm, khoa hoặc bộ môn GDQP&AN thực hiện nhiệm vụ GDQP&AN phải có phòng học chuyên dùng, thao trường tổng hợp, có đủ cơ sở vật chất, thiết bị dạy học GDQP&AN theo quy định hiện hành về danh mục thiết bị dạy học tối thiểu môn học GDQP&AN.	
2. Quản lý, sử dụng và bảo quản vũ khí, trang thiết bị GDQP&AN thực hiện theo quy định hiện hành.	
Chunk 2 Điều 8. Quản lý môn học và tổ chức dạy, học	
1. Các cơ sở giáo dục nghề nghiệp tổ chức dạy, học GDQP&AN theo kế hoạch đào tạo của cơ sở.	
2. Các cơ sở giáo dục đại học có khoa, bộ môn GDQP&AN tổ chức dạy, học tập trung theo kế hoạch đào tạo của cơ sở.	
3. Các trung tâm GDQP&AN tổ chức dạy, học tập trung theo kế hoạch đào tạo của trung tâm.	
4. Các cơ sở giáo dục tổ chức dạy, học thực hành kỹ thuật, chiến thuật, thuốc nổ và bắn đạn thật phải hợp đồng với cơ quan quân sự địa phương, đơn vị quân đội, công an để bảo đảm tuyệt đối an toàn về người, vũ khí, trang thiết bị.	
5. Lớp học lý thuyết nếu bố trí lớp ghép phải phù hợp với phương pháp dạy học và điều kiện cụ thể của các cơ sở giáo dục, nhưng không quá 150 người; lớp học thực hành không quá 40 người.	

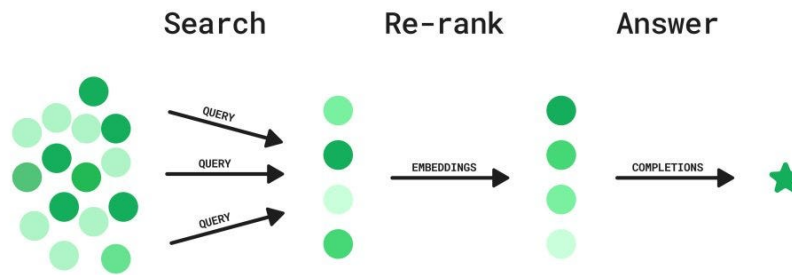
(b) Chunking phân cấp tự động

Hình 9: Context Aware Chunking với Docling

2.5 Reranker

2.5.1 Vai trò của Reranker trong RAG

Reranker đóng vai trò quan trọng trong việc cải thiện chất lượng retrieval:



Hình 10: Reranker

- **Sắp xếp lại kết quả:** Từ kết quả ban đầu của bộ embedder, reranker đánh giá lại độ liên quan chi tiết hơn
- **Hiểu ngữ cảnh sâu:** **Cross-encoder** xem xét tương tác toàn cục giữa query và toàn bộ document
- **Loại bỏ nhiễu:** Giảm các kết quả không liên quan nhưng có embedding tương tự do semantic overlap
- **Tăng độ chính xác:** Đưa các tài liệu thực sự liên quan lên top đầu

2.5.2 BGE-M3 Reranker

Chúng tôi sử dụng **BGE-M3 Reranker** với các ưu điểm:

- **Tương thích với BGE-M3 embedder:** Cùng hệ thống mô hình, tối ưu hiệu suất và độ tương thích
- **Hỗ trợ tiếng Việt tốt:** Được huấn luyện trên dữ liệu đa ngôn ngữ bao gồm tiếng Việt với chất lượng cao

2.6 Finetune

2.6.1 Thách thức về dữ liệu huấn luyện

Đối với trường Đại học Bách Khoa Hà Nội, việc thu thập dữ liệu huấn luyện cho hệ thống RAG gặp nhiều thách thức:

- **Dữ liệu ít:** Các câu hỏi liên quan đến quy chế đào tạo ít và không up-to-date
- **Đa dạng ngữ nghĩa:** Nếu thập các QA thì các cách biểu diễn dữ liệu khác nhau và dữ liệu khó kiểm và phải làm thủ công

2.6.2 Chiến lược sinh dữ liệu bằng LLM

Để giải quyết vấn đề thiếu dữ liệu, chúng tôi sử dụng **LLaMA 3.1 8B** qua **Groq API** để sinh dữ liệu huấn luyện tự động:

- **Template sinh câu hỏi:** Thiết kế prompt chuyên biệt để LLM sinh câu hỏi dựa trên nội dung chunk
- **Kỹ thuật mở rộng ngữ cảnh:** Mỗi chunk khi đưa vào LLM không đứng riêng lẻ mà kết hợp với chunk trước và sau để LLM hiểu rõ ngữ cảnh
- **Đa dạng hóa câu hỏi:** Yêu cầu LLM sinh nhiều loại câu hỏi khác nhau (chi tiết, tổng quan, điều kiện, quy trình)
- **Kiểm soát chất lượng:** Thiết lập các rule để đảm bảo câu hỏi chỉ dựa trên thông tin trong chunk, không suy diễn

Prompt Template Sinh Câu Hỏi

Bạn đang tạo dữ liệu huấn luyện cho hệ thống RAG.

Dựa trên NỘI DUNG bên dưới:

- Viết 7-8 câu hỏi thực tế mà sinh viên có thể hỏi
- Mỗi câu hỏi PHẢI trả lời được CHỈ dựa vào nội dung được cung cấp
- KHÔNG được suy đoán hay bịa thêm thông tin

NỘI DUNG:

"" ""

{ nội dung chunk đó và các chunk xung quanh để bổ nghĩa }

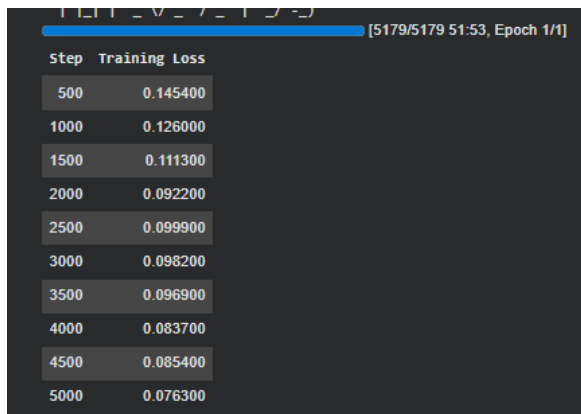
"" ""

	A	B	C	D	E	F
1	source	chunk_id	section	type	context	question
2					Học sinh, sinh viên là người khuyết tật, có giấy xác nhận khuyết tật theo quy định của pháp luật. Học sinh, sinh viên không đủ sức khỏe về thể lực hoặc mắc những bệnh lý thuộc diện miễn I.	
3					Học sinh, sinh viên đã hoàn thành nghĩa vụ quân sự, công an nhân dân. Đã tương đương tạm hoàn học môn học GDQP&AN.	Ai được coi là học sinh, sinh viên?
4	01_1%/202015%		7 ROOT	section	Học sinh, sinh viên là người khuyết tật, có giấy xác nhận khuyết tật theo quy định của pháp luật. Học sinh, sinh viên không đủ sức khỏe về thể lực hoặc mắc những bệnh lý thuộc diện miễn I. Học sinh, sinh viên đã hoàn thành nghĩa vụ quân sự, công an nhân dân. Đã tương đương tạm hoàn học môn học GDQP&AN.	
5					Học sinh, sinh viên có đủ sức khỏe về thể lực hoặc mắc những bệnh lý thuộc diện miễn I. Học sinh, sinh viên là người khuyết tật, có giấy xác nhận khuyết tật theo quy định của pháp luật. Học sinh, sinh viên không đủ sức khỏe về thể lực hoặc mắc những bệnh lý thuộc diện miễn I.	Học sinh, sinh viên nào được tạm hoàn học môn học GDQP&AN vì lý do sức khỏe?

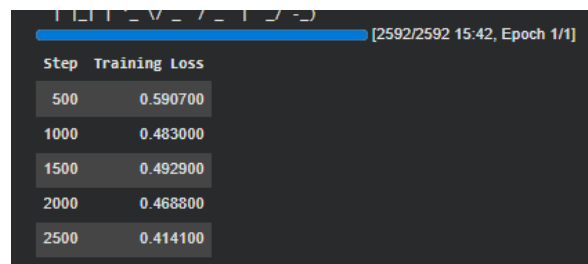
Hình 11: Bộ dữ liệu gen từ LLM

2.6.3 Finetune Embedder và Reranker

Sau khi có dataset Q&A, chúng tôi tiến hành finetune cả hai mô hình với 1 epoch để tránh overfit.



(a) Finetune Embedder



(b) Finetune Reranker

6. **Trích xuất thông tin (Information Extraction):** Nếu cần tạo biểu mẫu, trích xuất thông tin từ hội thoại (tên, MSSV, học phần muốn miễn, lý do, ...). Nếu không đủ thì phản hồi với sinh viên để sinh viên cung cấp thêm.
7. **Xác thực dữ liệu (Data Validation):** Kiểm tra tính đầy đủ và hợp lệ của thông tin (định dạng MSSV, tên học phần có tồn tại không, ...)
8. **Tạo phản hồi bằng LLM:** Sử dụng **LLaMA 3.3 70B** với ngữ cảnh từ các chunk retrieval để tạo câu trả lời chính xác và tự nhiên
9. **Tạo biểu mẫu tự động:** Nếu có yêu cầu và dữ liệu đã đầy đủ, tự động điền thông tin vào template DOCX và tạo file biểu mẫu với công cụ **docxptl** của python
10. **Trả kết quả cho sinh viên:** Hiển thị câu trả lời văn bản và/hoặc cung cấp file biểu mẫu đã điền sẵn cho sinh viên tải về

3.2 Một số Prompt Engineering chính

3.2.1 Prompt phát hiện ý định tạo biểu mẫu

Prompt cho phát hiện ý định được thiết kế nhằm phân biệt giữa câu hỏi thông thường và yêu cầu tạo biểu mẫu.

Prompt phát hiện ý định tạo biểu mẫu

Bạn là ****hệ thống Chat Bot hỗ trợ sinh viên Đại học Bách Khoa Hà Nội (HUST)**** trả lời bằng JSON được chỉ định.

Chức năng chính của bạn là hỗ trợ, trả lời các câu hỏi và cung cấp thông tin cho sinh viên.

Ngoài ra, hệ thống có thể tạo các đơn liên quan tới Đại học Bách Khoa Hà Nội, biểu mẫu nếu được yêu cầu nằm trong DANH SÁCH BIỂU MẪU.

NHIỆM VỤ:

- Phân tích tin nhắn người dùng và tóm tắt trò chuyện để xem họ có cần làm biểu mẫu hay không, và nếu có thì biểu mẫu nào.
- Nếu không có biểu mẫu nào phù hợp với ý định của sinh viên thì để `form_name` là `null`.
- `is_recommend` là `true` nếu người dùng chưa rõ cần lập đơn nào, và bạn đưa ra đề xuất.
- `is_recommend` là `false` nếu người dùng đang yêu cầu lập đơn đã mô tả.

DANH SÁCH BIỂU MẪU:

(Nếu không có biểu mẫu nào phù hợp trực tiếp thì để `form_name` là `null`)
{form_list}

TÓM TẮT TRÒ CHUYỆN:

{summary}

TIN NHẮN TRƯỚC CỦA HỆ THỐNG TỚI SINH VIÊN:

"{prev_sys_message}"

TIN NHẮN CỦA SINH VIÊN:

"{user_message}"

CHỈ TRẢ VỀ JSON THEO ĐỊNH DẠNG SAU, KHÔNG GIẢI THÍCH THÊM:

```
{{
  "form_name": Tên biểu mẫu hoặc null,
  "is_recommend": true/false hoặc null
}}
```

3.2.2 Prompt trích xuất thông tin từ hội thoại

Prompt này được thiết kế để trích xuất các thông tin cần thiết từ hội thoại nhằm tự động điền vào biểu mẫu.

Prompt trích xuất thông tin từ hội thoại

```
Bạn là hệ thống Chat Bot hỗ trợ sinh viên Đại học Bách Khoa Hà Nội (HUST), chuyên gia trích xuất dữ liệu.
Nhiệm vụ: Tìm kiếm thông tin từ TIN NHẮN MỚI NHẤT và TÓM TẮT TRÒ CHUYỆN để điền vào form.

ĐƠN: {form_name}
CÁC TRƯỜNG THÔNG TIN CẦN TÌM:
{fields_desc}
DỮ LIỆU CÓ SẴN:
{user_data}
TÓM TẮT TRÒ CHUYỆN:
{summary}
TIN NHẮN TRƯỚC CỦA HỆ THỐNG TỚI SINH VIÊN:
"{prev_sys_message}"
TIN NHẮN CỦA SINH VIÊN:
"{user_message}"
YÊU CẦU:
- Ưu tiên thông tin trong tin nhắn mới nhất và lịch sử trò chuyện.
- Nếu không có trong chat, mới lấy từ Hồ sơ sinh viên.
- Nếu không tìm thấy hoặc không thể suy luận ra, bỏ qua (không bịa đặt)

CHỈ TRẢ VỀ JSON THEO ĐỊNH DẠNG SAU, KHÔNG GIẢI THÍCH THÊM:
{{
  "values": [{"field_key": "extracted_value"}]
}}
```

3.2.3 Prompt tạo phản hồi hội thoại với ngữ cảnh RAG

Prompt này điều khiển chatbot tạo câu trả lời dựa hoàn toàn trên thông tin truy xuất từ tài liệu chính thức của nhà trường.

Prompt tạo phản hồi với ngữ cảnh RAG

THÔNG TIN TRA CỨU TỪ TÀI LIỆU NHÀ TRƯỜNG (RAG):

{rag_context}

==== KẾT THÚC THÔNG TIN TRA CỨU TỪ TÀI LIỆU NHÀ TRƯỜNG ====

Bạn là **hệ thống Chat Bot hỗ trợ sinh viên Đại học Bách Khoa Hà Nội (HUST)**.

Bạn đang trò chuyện với sinh viên ĐANG THEO HỌC tại Đại học Bách Khoa Hà Nội.

Chức năng chính của bạn là hỗ trợ, trả lời các câu hỏi và cung cấp thông tin cho sinh viên.

Ngoài ra, hệ thống có thể tạo các đơn liên quan tới Đại học Bách Khoa Hà Nội, biểu mẫu nếu được yêu cầu nằm trong DANH SÁCH BIỂU MẪU (nếu biểu mẫu sinh viên mong muốn không nằm trong đó thì suy luận và đưa phương án khác).

Từ nội dung tin nhắn và tóm tắt, suy luận xem sinh viên có cần tạo ra biểu mẫu nào trong DANH SÁCH BIỂU MẪU thì hãy đề xuất.

DANH SÁCH BIỂU MẪU:

{form_list}

DỮ LIỆU CỦA SINH VIÊN:

{user_data}

TÓM TẮT TRÒ CHUYỆN:

{summary}

TIN NHẮN TRƯỚC CỦA HỆ THỐNG TỚI SINH VIÊN:

"{prev_sys_message}"

TIN NHẮN CỦA SINH VIÊN:

"{user_message}"

NHIỆM VỤ:

- Trả lời LỊCH SỰ, TRỌNG TÂM như cán bộ hỗ trợ sinh viên.
- Trả lời bằng ngôn ngữ viết trang trọng và KHÔNG bao gồm các ký hiệu, tên biến, ngôn ngữ lập trình.
- Trình diễn câu trả lời dễ đọc, cách dòng, lề đầy đủ.
- Ưu tiên dùng RAG, thông tin tra cứu từ tài liệu nhà trường.
- Bỏ qua tài liệu RAG không liên quan.
- Trích dẫn nguồn nếu có thể.
- Sau khi trả lời, xem xét lại xem có biểu mẫu nào có thể sinh viên cần không thì hãy đề xuất.

3.2.4 Prompt tóm tắt hội thoại

Prompt này được sử dụng để duy trì bộ nhớ ngắn hạn của hệ thống thông qua việc tóm tắt hội thoại.

Prompt tóm tắt hội thoại

Bạn là ****hệ thống Chat Bot hỗ trợ sinh viên Đại học Bách Khoa Hà Nội (HUST)****.

Bạn đang thực hiện công việc tóm tắt văn bản hội thoại giữa bạn (Chatbot hỗ trợ) và một sinh viên.

LƯU Ý:

- Tóm tắt đầy đủ thông tin trong nội dung cùng với cuộc hội thoại
- Chú ý ghi lại cụ thể các THÔNG TIN CÁ NHÂN hay học tập mà sinh viên cung cấp, có thể hữu ích cho việc làm đơn.
- TRẢ VỀ TRỰC TIẾP NỘI DUNG PHẦN TÓM TẮT, KHÔNG MỞ ĐẦU, TIÊU ĐỀ, ĐỀ MỤC HAY GIẢI THÍCH GÌ THÊM TRƯỚC VÀ SAU ĐÂY.

HÃY TÓM TẮT:

Nội dung: {old_summary}

Sinh viên (chú ý vào câu trả lời của sinh viên vì nó có thể mang thông tin sinh viên): {user_message}

Bạn: {bot_response}

3.3 Một số cuộc hội thoại với chatbot

3.3.1 Q&A về đầu ra ngoại ngữ, thông tin nằm trong tài liệu của trường

Sinh viên: Quy định đầu ra ngoại ngữ của trường mình là gì, tôi học lớp đại trà

ChatBot:

Xin chào sinh viên Phạm Thành Khương,

Cảm ơn bạn đã liên hệ với hệ thống hỗ trợ sinh viên của Đại học Bách Khoa Hà Nội. Tôi xin trả lời câu hỏi của bạn về quy định đầu ra ngoại ngữ của trường.

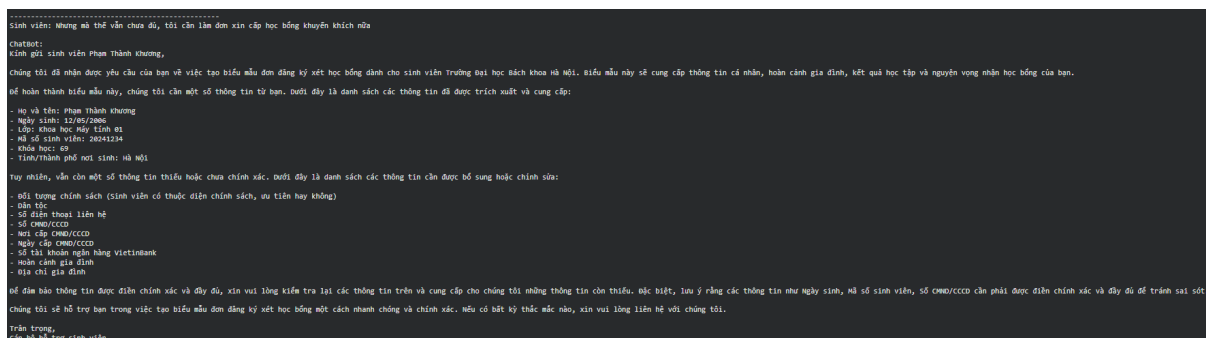
Theo tài liệu quy định về công tác tổ chức đánh giá và phân loại trình độ ngoại ngữ đầu vào, chương trình môn học ngoại ngữ, điều kiện được miễn học các học phần ngoại ngữ, chuẩn ngoại ngữ yêu cầu theo số lượng tín chỉ tích lũy và chuẩn ngoại ngữ đầu ra (QD_ngoi_ngu_tu_K68_CQ_final.pdf), quy định đầu ra ngoại ngữ của trường như sau:

- Để đạt chuẩn đầu ra ngoại ngữ, sinh viên cần tích lũy đủ số tín chỉ yêu cầu và đạt trình độ tiếng Anh tương ứng.
- Cụ thể, theo Bảng 3.2 Yêu cầu chuẩn Tiếng Anh theo số tín chỉ tích lũy và chuẩn đầu ra, sinh viên cần đạt tất cả học phần Tiếng Anh cơ sở gồm FL1131, FL1132, FL1133 hoặc đạt tối thiểu bậc 2.1 khi tích lũy đến 63 tín chỉ.

Về việc bạn học lớp đại trà, bạn sẽ cần tuân theo quy định chung về chuẩn ngoại ngữ đầu ra của trường. Nếu bạn có bất kỳ câu hỏi nào khác hoặc cần hỗ trợ thêm, xin đừng ngần ngại liên hệ với tôi.

Hình 14: Q&A về đầu ra ngoại ngữ

3.3.2 Đoạn chat hỗ trợ sinh viên điền mẫu xin cấp học bổng



Hình 15: Đoạn chat hỗ trợ sinh viên tạo đơn xin cấp học bổng

<div style="border: 1px solid black; width: 80px; height: 80px; margin: 0 auto; display: flex; align-items: center; justify-content: center;">ảnh 3x4</div>	<p>CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM <u>Độc lập – Tự do – Hạnh phúc</u></p> <p>ĐƠN ĐĂNG KÝ HỌC BỔNG</p> <p>Kính gửi: Phòng Công tác Sinh viên</p> <p>1. Sơ lược về bản thân: Họ và tên...Phạm Thành Khương ; Ngày sinh: ..12/05/2006 Nơi sinh:Hà Nội; Giới tính:Nam MSSV:20241234 ; Lớp:Khoa học Máy tính 01 Khóa:69; ViệnCNTT Đối tượng chính sách: ...Con nhà thương binh ; Dân tộc: ...Kinh Điện thoại liên hệ: ...0869346316 ; E-mail: ..so1bk@hust.edu.vn Số CMND/ Thẻ CCCD:027204000222 ; Nơi cấp CATP; Ngày cấp..21/03/2021..... Số tài khoản ngân hàng Vietinbank: ..101234567890..... Chi nhánh.....</p> <p>2. Hoàn cảnh gia đình (trình bày sơ lược hoàn cảnh gia đình): ...Gia đình gặp bão lũ Tình hình học tập: - Điểm trung bình tích lũy (CPA): ..3.5..... - Các thành tích học tập/ NCKH khác (kèm bản sao các giấy chứng nhận có giá trị):</p> <p>3. Các thành tích hoạt động xã hội (trong thời gian học tại Trường ĐHBKHN): .. 4. Địa chỉ gia đình sinh viên (địa chỉ gửi thư): ...Hai Bà Trưng, Hà Nội Nay em viết đơn này xin được tham gia xét nhận học bổng:Học bổng khuyến khích..... Em xin chân thành cảm ơn.</p> <p style="text-align: right;"><i>Hà Nội, ngày tháng năm 20.....</i> Sinh viên <i>(Ký và ghi rõ họ tên)</i></p>
---	---

Hình 16: Đơn xin cấp học bổng tạo tự động

3.3.3 Đoạn chat tạo giấy chứng nhận học bổng

```
Sinh viên: Tôi cũng mới nhận được học bổng của trường xong, giờ tôi cần chứng nhận để đem đi khoe.
ChatBot:
Kính gửi sinh viên Phạm Thành Khương,

Chúng tôi đã nhận được yêu cầu của bạn về việc tạo biểu mẫu Giấy chứng nhận sinh viên được cấp học bổng hỗ trợ học tập. Biểu mẫu này sẽ được sử dụng để xác nhận thông tin sinh viên và loại học bổng mà bạn được cấp trong năm học cụ thể.
Để hoàn thành biểu mẫu này, chúng tôi cần một số thông tin từ bạn. Dưới đây là thông tin mà chúng tôi đã trích xuất được từ yêu cầu của bạn:

- Họ và tên: Phạm Thành Khương
- Ngày sinh: 12
- Tháng sinh: 05
- Năm sinh: 2006
- Lớp: Khoa học Máy tính 01
- Mã số sinh viên: 20241234
- Khóa học: 69
- Tỉnh/Thành phố nơi sinh: Hà Nội

Tuy nhiên, chúng tôi vẫn còn thiếu một số thông tin quan trọng để hoàn thành biểu mẫu này. Cụ thể, chúng tôi cần biết:

- Loại học bổng hỗ trợ học tập mà bạn được cấp (ví dụ: toàn phần, một phần, học bổng khuyến học...)
- Năm học mà bạn được cấp học bổng (ví dụ: 2024-2025)

Xin vui lòng cung cấp thông tin này để chúng tôi có thể hoàn thành biểu mẫu của bạn. Nếu bạn có bất kỳ câu hỏi nào, xin đừng ngần ngại liên hệ với chúng tôi.
Chúng tôi sẽ gửi biểu mẫu hoàn thành cho bạn sau khi nhận được thông tin còn thiếu.

Trân trọng,
Cán bộ hỗ trợ sinh viên.

-----
Sinh viên: Đây là học bổng toàn phần, mới ăm nay 2026 ý
ChatBot:
Đã tạo thành công biểu mẫu tại output/giay_chung_nhan_hb.docx theo yêu cầu của sinh viên.
```

Hình 17: Đoạn chat tạo giấy chứng nhận học bổng

BỘ GIÁO DỤC VÀ ĐÀO TẠO TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI	CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM <u>Độc lập - Tự do - Hạnh phúc</u>
GIẤY CHỨNG NHẬN SINH VIÊN ĐƯỢC CẤP HỌC BỔNG HỖ TRỢ HỌC TẬP	
HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI	
Chứng nhận anh / chị: Phạm Thành Khương	
Sinh ngày:.. 12...tháng..05 ...năm.. 2006 ...Tại Thành phố/Tỉnh Hà Nội	
Là sinh viên đang học tại lớp: Khoa học Máy tính 01 Số hiệu sinh viên: 20241234	
Khóa 69Hệ đào tạo: Đại học chính quy.	
Được cấp học bổng hỗ trợ học tập loại: ... toàn phầnnăm học: ... 2025-2026	
Hà Nội, ngàytháng năm 20.....	
TL. HIỆU TRƯỞNG KT. TRƯỞNG PHÒNG CÔNG TÁC SINH VIÊN PHÓ TRƯỞNG PHÒNG	
<i>CTSV. QT 18. BM 02</i> <i>Ban hành lần 03, ngày 06-6-2019</i>	

Hình 18: Giấy chứng nhận học bổng tạo tự động

Source

Ngoại trừ phần data crawl thì source code đều được để trên **Google Colab**.

1. Sử dụng Docling để xử lý và chunk dữ liệu
2. Tạo bộ dữ liệu finetune với LLM
3. Finetune Embedder
4. Finetune Reranker
5. Index lại dữ liệu trong VectorDB với embedder đã finetune
6. Hust ChatBot

Kết luận

Hệ thống AI ChatBot hỗ trợ sinh viên HUST với RAG nâng cao đã giải quyết thành công bài toán hỗ trợ thông tin và hành chính cho sinh viên. Với việc kết hợp nhiều kỹ thuật hiện đại như Context Aware Chunking, fine-tuning embedding và reranker, cùng với chiến lược sinh dữ liệu huấn luyện thông minh bằng LLM, hệ thống đạt được độ chính xác cao trong việc trả lời câu hỏi và tự động hóa quy trình tạo biểu mẫu. Các kỹ thuật prompt engineering được thiết kế tỉ mỉ giúp hệ thống hiểu chính xác ý định người dùng và trích xuất thông tin hiệu quả. Trong tương lai, hệ thống có thể được mở rộng để hỗ trợ thêm nhiều loại dịch vụ và tích hợp sâu hơn với các hệ thống thông tin của trường.

Tài liệu tham khảo

1. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33.
2. Xiao, S., et al. (2023). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings. *arXiv preprint arXiv:2312.15547*.
3. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
4. Đại học Bách Khoa Hà Nội. (2024). Các văn bản quy chế đào tạo và biểu mẫu hành chính.
5. Docling Documentation. (2024). <https://docling-project.github.io/>
6. Groq API Documentation. (2024). <https://console.groq.com/docs>
7. Sentence-Transformers Documentation. (2024). <https://www.sbert.net/>