ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

A Large Language Model Approach to Educational Survey Feedback Analysis

Môn học: Nhập môn xử lý ngôn ngữ tự nhiên

GVHD: Nguyễn Hồng Bửu Long
GVTH: Lê Thanh Tùng
Lương An Vinh
Dương Thị An

Thông tin nhóm

MSSV	Thành viên
21120071	Nguyễn Thị Thanh Hoa (NT)
21120175	Tô Ngọc Hân
21120224	Lều Huy Đức

Thành phố Hồ Chí Minh, 2023

Mục lục

Trang

Danh sách bảng

Danh sách hình vẽ

Chapter I

Giới thiệu

Việc phân tích phản hồi giáo dục là quá trình thu thập, xử lý và phân tích dữ liệu phản hồi của học sinh, giáo viên, phụ huynh và các bên liên quan. Việc này có thể được sử dụng để cải thiện hiệu quả dảng dạy và học tập, cũng như để xác định các lĩnh vực cần cải thiện và phát huy.
Trong bài báo cáo của Michael J. Parker và nhóm tác giả đã sử dụng mô hình ngôn ngữ lớn (Large Language Models) để phân tích phản hồi giáo dục từ các khảo sát, với mô hình GPT-4 và GPT-3.5 để thực hiện nhiều nhiệm vụ NLP khác nhau. Mô hình cũng đề xuất sử dụng chuỗi tư duy (CoT) để đưa ra giải thích kết quả cho mô hình.

1. Cách tiếp cận trước đây

Trước đây việc phân tích các phản hồi đều cần có sự can thiệp của con người trong các giai đoạn quan trọng. Với việc ra đời của học máy (ML) đã giúp giải quyết được những văn bản phi cấu trúc, điều mà các kỹ thuật TF-IDF, Word2Vec, latent semantic analysis,... thuở đầu chưa làm được. Tuy nhiên còn nhiều thách thức khác như việc đào tạo mô hình học máy trước có thể khó khăn trong việc hiểu các thuật ngữ cụ thể của lĩnh vực hay việc phân tích tâm trạng có thể không phù hợp với bối cảnh.

2. Phân loại nhiệm vụ

Việc phân tích những phản hồi của học viên về khóa học yêu cầu xử lý rất nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP) trên dữ liệu phi cấu trúc và phân loại.Vì thế nhóm tác giả đã thực hiện một mô hình để xử lý việc trên.Quy trình công việc bao gồm:

- Phân loại bình luận: Xác định các bình luận là đơn nhãn hay đa nhãn. Các nhãn này có thể được tùy chỉnh tùy thuộc vào mục tiêu nghiên cứu cụ thể.

- Trích xuất văn bản: Trích xuất những thông tin quan trọng trong bình luận giúp dễ dàng phân loại chủ đề cho bình luận.

- Phân tích tâm trạng: Xác định tâm trạng của bình luận là tích cực hay tiêu cực.

3. LLM và nghiên cứu liên quan

Việc phân tích phản hồi giáo dục trích xuất các ý kiến từ đó để nhận được đánh giá giáo viên, trang thiết bị,... được xem là ứng dụng của xử lý ngôn ngữ tự nhiên. Để có được những kết quả như hiện tại là tập hợp của nhiều nghiên cứu trước đó. Cụ thể như:

- Sự ra đời của kiến trúc mạng nơ-ron(neuron network) năm 2017 đã tạo một mô hình mới cho NLP là BERT được phát triển vào 2018 hay RoBERTa,...

- Sự phổ biến của LLMs đa phần là do sự xuất hiện của ChatGPT và các mô hình GPT-3.5 và GPT-4. Các nghiên cứu gần đây đã xem xét việc sử dụng ChatGPT cho các nhiệm vụ ghi chú và phân loại văn bản, với kết quả kết hợp dựa trên sự biến đổi của các lời nhắc, tập dữ liệu, tham số và độ phức tạp của các nhiệm vụ. Có các nghiên cứu đã chỉ ra khả năng của ChatGPT trong việc thực hiện ghi chú văn bản hoặc cung cấp giải thích bằng ngôn ngữ tự nhiên ở mức tiệp cận hoặc tương đương với con người.

4. Ý nghĩa và mục tiêu đề ra

- Trình bày một phương pháp sử dụng LLM để thực hiện nhiều nhiệm vị phân tích văn bản phi cấu trúc trên những câu trả lời khảo sát gồm: phân loại đa nhãn, phân loại nhị phân, trích xuất.

- Đánh giá hiệu suất tiếp cận "zero-shot" trên tất cả các nhiệm vụ.

- Cho thấy tiềm năng của LLMs để giải thích cách người dùng đưa ra câu trả lời.

Chapter II

Phương pháp

1. Dữ liệu khảo sát:

Phương pháp nghiên cứu của cuộc khảo sát nhằm thu thập phản hồi về các khóa học khoa học y sinh, đặc biệt là các khóa học liên quan đến di truyền học, miễn dịch học và dược học. Dưới đây là phân tích chi tiết:

1.1 Lựa chọn Mẫu:

- Một phần con của 2500 câu trả lời từ khảo sát được chọn ngẫu nhiên từ một tập lớn hơn của các câu trả lời khảo sát.

- Những phản hồi này được thu thập như là ý kiến phản hồi cuối khóa học trên nhiều khóa học khoa học y sinh khác nhau.

1.2 Nội dung Khảo sát:

Câu hỏi Mở: Khảo sát bao gồm bốn câu hỏi mở được thiết kế để đưa ra phản hồi chất lượng khóa học từ người tham gia:

- Q1: "Please describe the best parts of this course."

- Q2: "What parts of the experience enhanced your learning of the concepts most?"

- Q3: "What can we do to improve this course?"

- Q4: "Please provide any further suggestions, comments, or ideas you have for this series."

Một số ý kiến khảo sát được chọn đặc biệt làm bộ phát triển. Bộ này được sử dụng để điều chỉnh Hướng Dẫn của Mô Hình Ngôn Ngữ (LLM). Mục đích của việc điều chỉnh có thể là để cải thiện khả năng của mô hình tạo ra các phản ứng hoặc hiểu biết liên quan dựa trên dữ liệu khảo sát cụ thể này.

1.3 Quá trình thu thập và xử lý các phản hồi từ cuộc khảo sát:

- Sử dụng nền tảng Qualtrics để thu thập các câu trả lời và ý kiến từ người tham gia khảo sát.

- Sử dụng thư viện Pandas phiên bản 2.0.1 để thực hiện xử lý dữ liệu cơ bản. Cụ thể, loại bỏ các khoảng trắng không cần thiết ở đầu và cuối câu và tự động loại bỏ các phản hồi không có nội dung (các biến thể NA hoặc None).

- Duyệt qua các phản hồi từ khảo sát một cách thủ công để kiểm tra chất lượng và xác nhận rằng các bước xử lý dữ liệu đã diễn ra đúng đắn.

- Sử dụng công nghệ nhận diện thực thể đặt tên (NER) chạy ở mức độ địa phương để đảm bảo rằng không có thông tin riêng tư hay nhạy cảm nào trong phản hồi được truyền tải đến các mô hình ngôn ngữ lớn công khai (LLMs).

2. Hệ thống nhãn dán

2.1 Định nghĩa nhãn dán trong xử lý ngôn ngữ tự nhiên:

- Tag hay được biết đến là gán nhãn được biết là một quá trình của xử lý ngôn ngữ tự nhiên(NLP) trong đó mỗi từ trong văn bản được gắn nhãn với phần nội dung tương ứng điều này có thể gồm danh từ,tính từ, dộng từ,...

- Việc gắn nhãn có nhiều tác dụng cho việc phân tích văn bản như trích xuất thông tin nhận dạng thực thể hay để xác định cấu trúc ngữ pháp trong câu.

2.2 Xây dựng nhãn dán

Dựa trên dữ liệu bài báo cáo, dựa vào dữ liệu đánh giá khóa học ở trong báo cáo trên các tác giả đã dùng những dữ liệu nhận xét về khóa học ở trong web để lấy dữ liệu cho mục miêu tả dữ liệu Áp dụng theo phương pháp được sử dụng trong bài báo cáo trên các tác giả đã đưa được hơn 70 nhãn dán cơ bản như bảng dưới:

Bảng II.1: Mô tả tag.

Tag	Description
Course logistics and fit	Course delivery (policy, support), cost, difficulty, time commitment, grading, credit, schedule, user fit, access, background (e.g., prereqs and appropriateness of course level).
Curriculum	Course content, curriculum, specific topics, course structure. This focuses on the content and the pedagogical structure of the content, including flow and organization. This also includes applied material such as clinical cases and case studies. Includes references to pre-recorded discussions between experts or between a doctor and a patient. Includes specific suggestions for additional courses or content.
Teaching modality	Video, visual, interactive, animation, step-by-step, deep dive, background builder (the format rather than the content/topic)
Teaching	Instructors, quality of teaching and explanations
Assessment	Quizzes, exams
Resources	Note taking tools, study guides, notepads, readings. Includes other potential static resources like downloadable video transcripts.
Peer and teacher interaction	Includes chances for the student to interact with another person in the course (teacher or student). This includes discussion forums, teacher-student or student-student interactions.Includes requests for live sessions with teachers or live office hours.
Other	Catch-all for the rarer aspects that we'll encounter and also the 'na', 'thank you', etc. comments that don't really belong in the above bins. Also for sufficiently general comments like 'all the course was terrific that can't be narrowed down to one of the other categories.

2.3 Mô tả cho nhãn dán

- Với mỗi nhãn dán các tác giả đã miêu tả 2 đến 3 câu. Với cách này hệ thống có thể nhận diện được ngữ cảnh thích hợp hơn. Người dùng có thể tự chỉnh sửa miêu tả để hệ thống phù hợp với mục đích riêng.

- Áp dụng mô hình ngôn ngữ lớn (LLM) để phân loại đa nhãn như một cách để phân loại chủ đề hay ngữ cảnh nào đó.

- Vì 4 tác giả trong bài báo cáo này gán nhãn độc lập với nhau nên việc lặp đi lặp lại thử nghiệm trong khoảng 100 bộ câu trả lời giúp xác định tính nhất quán giữa các thẻ. Với phương pháp Jaccard giúp xác định được sự tương đồng giữa các kết quả thử của các tác giả với nhau để đưa ra con số tương đồng giữa các nhãn dán các tác giả tạo ra.

3. LLM processing

- Sử dụng các mô hình ngôn ngữ học máy của OpenAI, bao gồm GPT-3.5 và GPT-4, cho nhiều nhiệm vụ như phân loại đa nhãn, đa lớp, nhị phân và phân tích tâm trạng.

- Các cuộc gọi API được thực hiện không đồng bộ để xử lý nhiều cuộc gọi mô hình đồng thời, và sử dụng Function calling để tạo ra đầu ra có cấu trúc cho mọi nhiệm vụ một cách đáng tin cậy.

- Các thử nghiệm được thực hiện với các tham số được điều chỉnh, bao gồm nhiệt độ và các tham số khác của mô hình.

- Kỹ thuật zero-shot Chain-of-Thought (CoT) prompting được áp dụng để yêu cầu mô hình suy luận từng bước mà không có ví dụ.

- Các kỹ thuật khác như bổ sung ngữ cảnh mô tả và sử dụng thông tin mô tả cho các nhãn được áp dụng để cải thiện hiệu suất mô hình:

  – Sử dụng Zero-shot CoT và kiểm tra suy luận CoT để tinh chỉnh prompt cho các nhiệm vụ phân loại và phân tích tình cảm.

  – Bổ sung ngữ cảnh mô tả vào prompt để cải thiện khả năng giải thích và hiệu suất của mô hình.

  – Thêm thông tin mô tả cho các nhãn, bổ sung ngữ cảnh thông qua câu hỏi khảo sát để cải thiện hiệu suất mô hình trong việc hiểu nhiệm vụ.

  – Sử dụng function calling để tạo ra đầu ra có cấu trúc cho mọi nhiệm vụ một cách đáng tin cậy.

4. So sánh với mô hình khác

Ngoài việc so sánh với các nhãn thực tế do con người cung cấp, đối với phân loại đa nhãn cũng so sánh các mô hình với SetFit, một phương pháp tinh chỉnh SentenceTransformers dựa trên Sentence-BERT và chỉ yêu cầu rất ít dữ liệu có nhãn; đối với phân tích cảm xúc cũng đã so sánh với một mô hình dựa trên RoBERTa có sẵn trên công khai, được huấn luyện trên 124 triệu Tweets. Những so sánh này cung cấp một số ngữ cảnh cho hiệu suất của các LLMs so với các mô hình chuyên biệt gần đây.

5. Các chỉ số đánh giá

Nhóm tác giả sử dụng Scikit-learn 1.2.0, numpy 1.23.5, và Pandas 2.0.1 để thực hiện kiểm định thống kê và phân tích dữ liệu trong mô hình học máy. Weights và Biases được sử dụng để theo dõi đánh giá mô hình.

Đối với nhiệm vụ phân loại đa nhãn, nhóm tác giả tạo ra nhãn thực tế từ nhiều người gán nhãn, sử dụng cả hai phương pháp: 1) hàng đồng thuận và 2) nhãn đồng thuận. Mô hình được tinh chỉnh trên một phần của tập dữ liệu thực tế và đánh giá thông qua nhiều chỉ số.

Đối với nhiệm vụ phân loại nhị phân, các chỉ số như accuracy, precision, recall, và F1 score được tính toán so sánh với người gán nhãn chuyên gia.

Đối với trích xuất đoạn văn, đánh giá bởi GPT-4 dựa trên nhiều khía cạnh, bao gồm sự chính xác của việc trích xuất và sự có mặt của đoạn văn ảo.

Phương pháp suy luận trong phân tích chủ đề không có phương pháp đánh giá chấp nhận được. Kết quả cảm xúc của GPT-3.5 và GPT-4 được so sánh với một bộ phân loại cảm xúc RoBERTa và người gán nhãn, sử dụng các chỉ số như accuracy, precision, recall, và F1 score.

6. Phương pháp cho các quy trình LLM:

Các loại quy trình chính:

- Phân tích ở cấp độ cao:
  - Inductive Thematic Analysis (ITA): là một phương pháp "từ dưới lên" hỗ trợ trường hợp sử dụng khi không có nhãn trước đã được định nghĩa.
  - Multi-label Classification (MLC): Sử dụng nhãn được định nghĩa trước, là một phương pháp "từ trên xuống"hay còn gọi là phân tích chủ đề suy luận. Khi các hạng mục quan trọng đã biết trước, MLC là bước quan trọng đầu tiên, phân loại các phản hồi thành các hạng mục liên quan. như việc mô hình chủ đề.
- Phân tích tập trung:
  - Extraction: là bước quan trọng để tập trung vào một mục tiêu cụ thể trong phân tích.
  - Multi-class Classification: có thể được sử dụng sau khi trích xuất để cung cấp đầu ra cho phân tích và đánh giá tiếp.
  - Sentiment Analysis: Áp dụng cuối cùng để xác định sự tích cực hoặc tiêu cực trong các phản hồi.

7. Ví dụ về quy trình làm việc

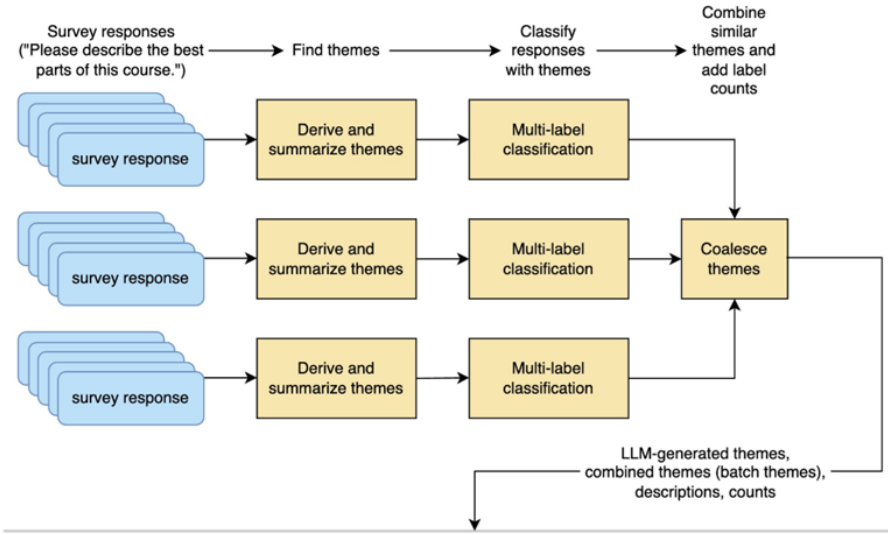
7.1 Ví dụ - Phân tích chủ đề bằng phương pháp quy nạp (tiếp cận "từ dưới lên")

Qua 3 bước LLM: Đầu tiên, các chủ đề được xác định và tóm tắt cho từng lô nhận xét. Mỗi lô được thiết lập với kích thước phù hợp với cửa sổ ngữ cảnh của mô hình sử dụng (ví dụ: 8K ký tự). Tiếp theo, phân loại các nhận xét bằng cách sử dụng các chủ đề được xác định trong

bước thứ nhất (phân loại đa nhãn). Bước 3 là hợp nhất các chủ đề thành bộ chủ đề cuối cùng và số lượng nhãn được tổng hợp từ các chủ đề được kết hợp.

Trong phương pháp định tính (thu thập các phi số cho quá trình nghiên cứu), các bước 1,3 là các bước phân tích chuyên đề quy nạp (tương tự mô hình hóa chủ đề). Các chủ đề được quy nạp từ các nhận xét. Kết quả các tác giả nhận được là 625 chủ đề bao gồm thông tin như số lượng ý kiến tương ứng với mỗi chủ đề, cùng với tiêu đề và mô tả của chúng. Số nhận xét mà mô hình ngôn ngữ lớn(LLM) xác định được tương ứng với từng chủ đề, tiêu đề và mô tả.

Hình dưới đây miêu tả qui trình dẫn xuất các chủ đề từ nhận xét của sinh viên (kết quả hiển thị bằng GPT-4). Với phản hồi khảo sát từ câu hỏi Q1("Please describe the best parts of this course?").

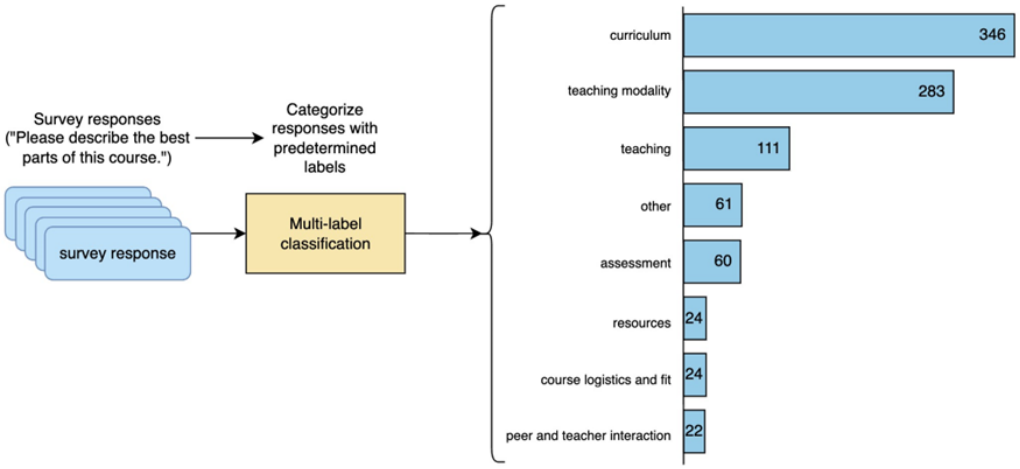


Theme title	Combined themes	Theme description	Total count
Real-world Applications	Clinical Applications, Patient Cases, Real-world Applications, Value of Real-World Examples and Clinical Applications, Engagement with Real-Life Patient Cases	Students found the real-world examples, clinical applications, and patient cases particularly engaging and useful for understanding the material.	287
Visual Learning Tools	Helpful Videos, Graphics and Animations, Visual Learning Tools, Appreciation for Videos and Animations	Students appreciated the use of videos, animations, and visual aids in the course, finding them helpful for understanding complex concepts.	285
Concept Understanding and Application	Concept Understanding, Application of Theory, Content Depth and Breadth, Current Advances, Clear and Concise Explanations	Students appreciated the course's focus on understanding concepts, application of theory, and the depth and breadth of content covered, including the latest advances in the field.	199
Interactive Elements	Interactive Elements, Quizzes and Interactivity, Usefulness of Quizzes and Assessments	Several students appreciated the interactive elements of the course, including quizzes and interactive diagrams, finding them effective for reinforcing the material and testing their understanding.	87
Course Organization and Flexibility	Course Organization, Course Flexibility, Flexibility	The course's organization, structure, and flexibility were praised by many students, who found the flow of the course and the breakdown of work manageable and self-paced.	68
Professional Relevance and Discussions	Professional Discussions, Relevance	Students enjoyed the discussions and interviews with professionals and researchers in the field, and found the course relevant to their current positions or fields of study.	48
Treatment Development	Treatment Development	Students appreciated the course's focus on the development of new treatments and therapies.	29

Hình II.1: Lấy chủ đề từ nhận xét của sinh viên (kết quả hiển thị bằng GPT-4).

7.2 Ví dụ - Phân tích cấp cao bằng cách phân loại nhận xét của sinh viên (tiếp cận "từ trên xuống")

Sử dụng những nhãn được xác định và phát triển trước đó đã được thử nghiệm trên 625 câu nhận xét từ câu hỏi Q1(“Please describe the best parts of this course?”) để phân loại đa nhãn các câu trả lời cho khảo sát. Các nhận xét được phân loại để có thể sử dụng để phân tích (phân loại các câu trả lời cho các câu hỏi khác nhau) hay có thể làm điểm đầu cho các công việc tiếp theo, chẳng hạn như phân tích chi tiết hơn về từng nhóm phản hồi hoặc đề xuất các cải tiến cụ thể cho khóa học dựa trên những gì sinh viên đã chia sẻ.



Hình II.2: Phân loại đa nhãn của nhận xét của sinh viên (kết quả được thể hiện bằng GPT-4).

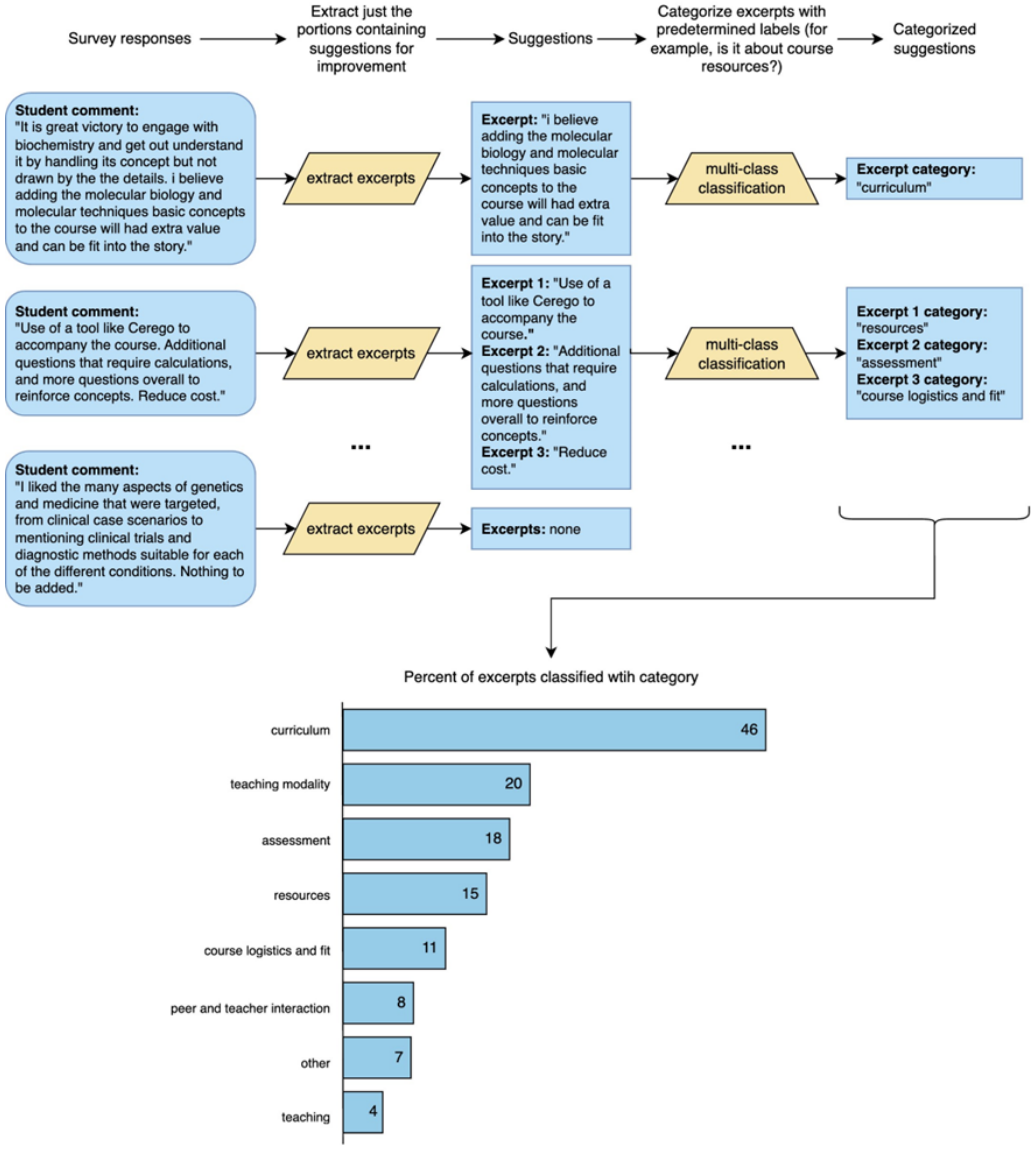
7.3 Ví dụ - Tìm kiếm gợi ý để cải thiện khóa học

Quy trình thực hiện tìm kiếm và định lượng các cách cải thiện khóa học gồm các bước:

- Trích xuất các trích đoạn có liên quan. Các kết quả từ bước trích xuất được giả định là đủ để mỗi nhãn dán có thể được phân loại là đối tượng duy nhất trong các nhãn dán có sẵn.

- Phân loại nhiều lớp dựa vào nhãn để tạo định lượng cũng như định tuyến cho những nhận xét cho các bên liên quan.

- Kết quả cho ra một số nhận xét thực tế đại diện cho tập hợp các nhận xét được đại diện từ tập hợp lớn hơn.



Hình II.3: Tìm kiếm gợi ý để cải thiện khóa học từ nhận xét của sinh viên .

7.4 Ví dụ - Những nội dung hoặc chủ đề nào khác mà sinh viên quan tâm muốn được bao gồm?

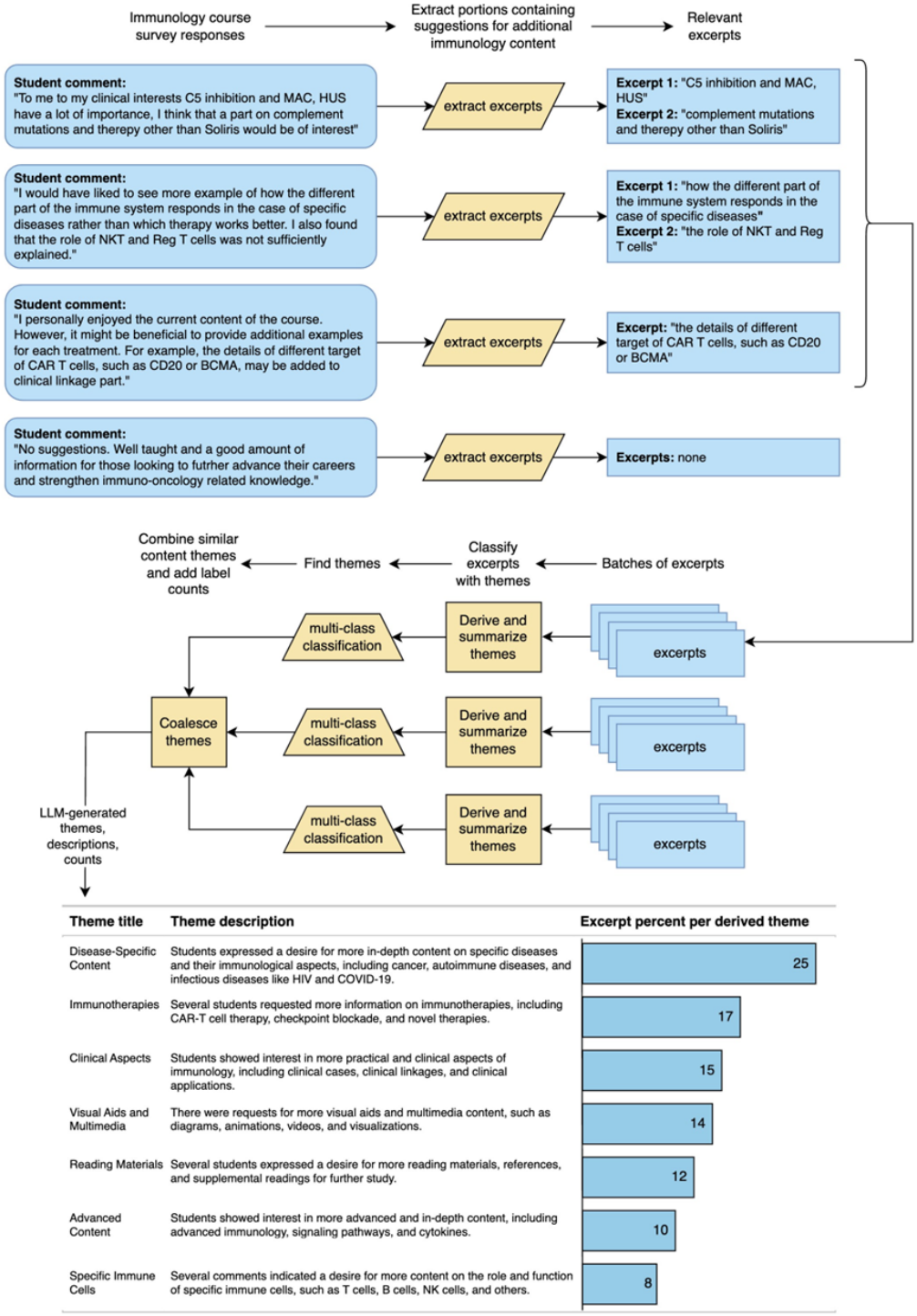
Từ các phản hồi của khảo sát các câu hỏi liên quan đến khóa học miễn dịch (Q3 và Q4) qua các qui trình làm việc:

- Đầu tiên, trích xuất các phần chỉ chứa gợi ý nội dung hoặc chủ đề mới từ câu phản hồi của sinh viên. Các chủ đề gợi ý nội dung sau đó được rút ra và tóm tắt từ các đoạn trích; điều này được thực hiện theo lô nếu chúng không thể phù hợp trong một lời nhắc duy nhất cho LLM (tức là, nếu có quá nhiều trích đoạn để phù hợp với kích thước ngữ cảnh tối đa của mô hình).

- Phân loại đa lớp trên các đoạn trích với các chủ đề đã xác định.

- Tìm kiếm và xác định các chủ đề từ các đoạn trích đã phân loại.

- Nếu phân tích theo chủ đề được thực hiện theo lô, tập hợp các chủ đề từ các lô này sau đó được kết hợp lại và thêm số lượng nhãn để đi đến một bộ chủ đề nội dung cuối cùng.

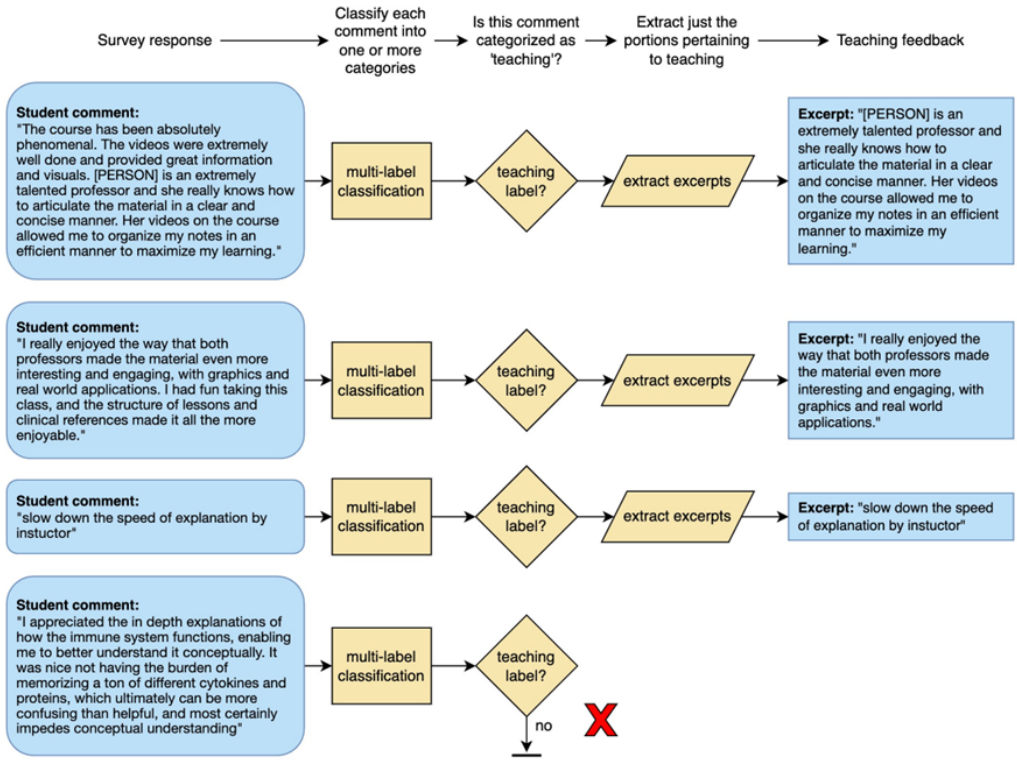


Hình II.4: Tìm kiếm các đề xuất nội dung miễn dịch học mới từ các bình luận của sinh viên .

7.5 Ví dụ - Sinh viên đưa ra những phản hồi gì về giảng dạy và giải thích?

Từ các phản hồi của khảo sát các câu hỏi liên quan đến khóa học miễn dịch (Q3 và Q4) qua các qui trình làm việc:

- Sử dụng phân loại đa nhãn như một bước khởi đầu, sử dụng các nhãn có sẵn đã được phát triển.
- Tiếp theo là trích xuất các trích đoạn có liên quan từ các nhận xét được phân loại vào danh mục 'giảng dạy' để tập trung vào ý kiến liên quan đến giảng dạy (nhóm xác định trước).
- Sử dụng trích xuất để hạn chế phạm vi của phân tích. Bước trích xuất làm tinh tế thông tin, tuy nhiên có thể xuất hiện lỗi (như lọc ra các bình luận liên quan mặc dù chúng có tham chiếu đến chất lượng giải thích), nổi bật cần phải rõ ràng trong mục tiêu trích xuất.
- Định rõ mục tiêu thông qua zero-shot prompting để giảm thiểu lỗi.

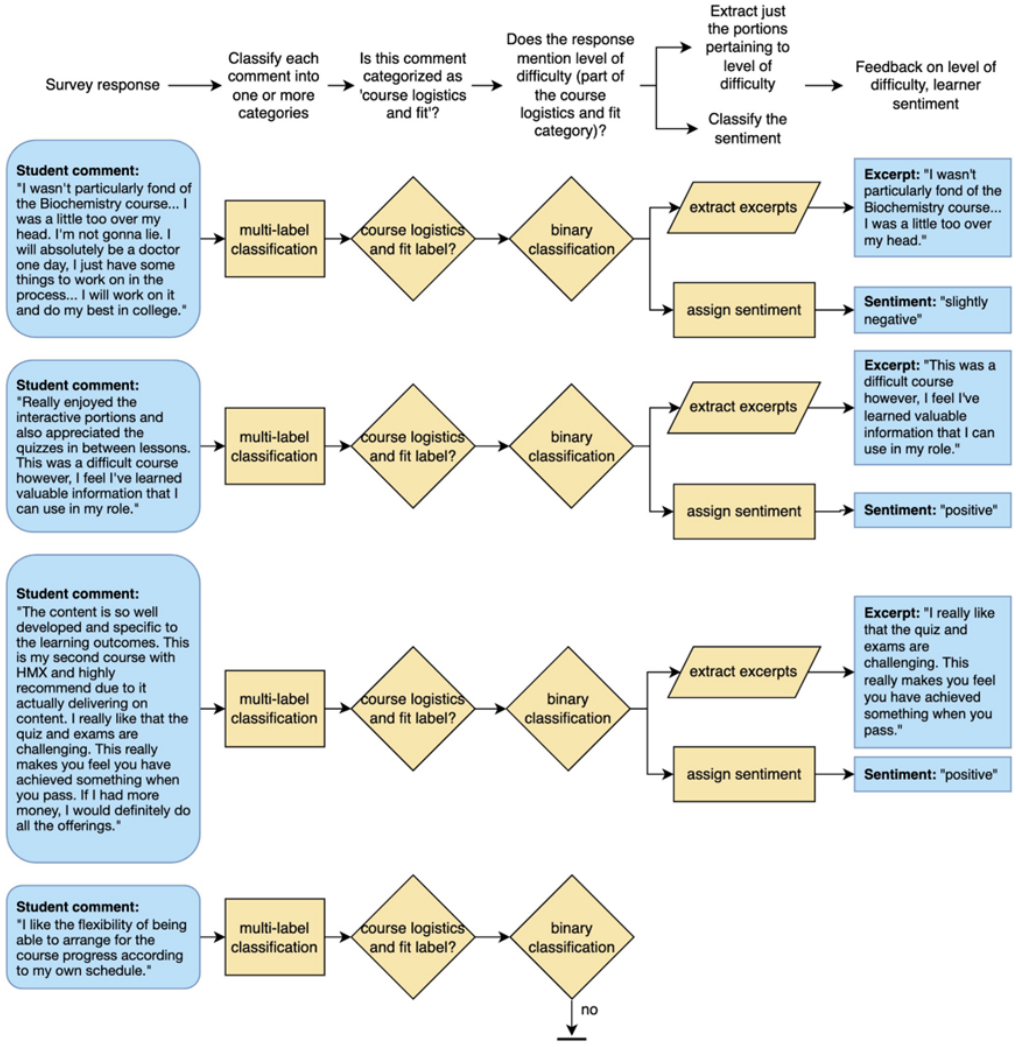


Hình II.5: Phản hồi về việc giảng dạy và giải thích (kết quả hiển thị bằng GPT-4). Chữ 'x' màu đỏ biểu thị lỗi do mô hình.

7.6 Ví dụ - Sinh viên cảm thấy thế nào về mức độ khó của khóa học?

Từ các phản hồi của khảo sát các câu hỏi liên quan đến khóa học miễn dịch (Q3 và Q4) qua các qui trình làm việc:

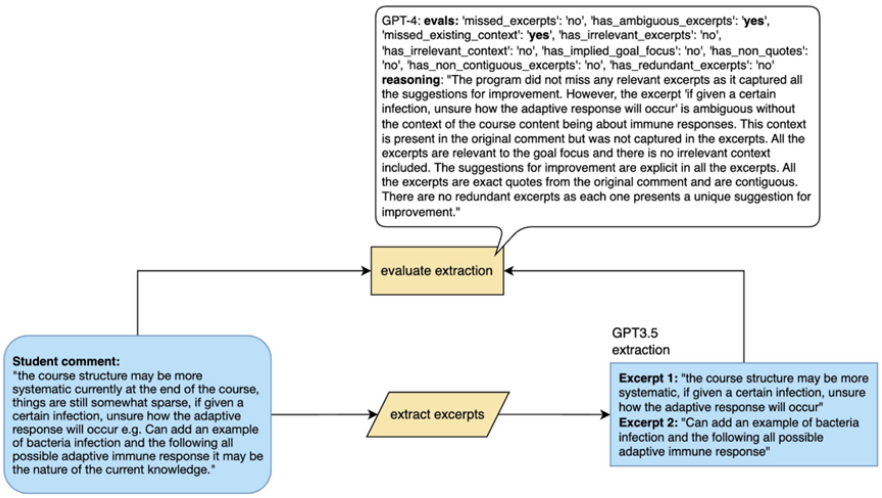
- Sử dụng nhãn có sẵn thông qua phân loại đa nhãn.
- Lọc ý kiến dưới nhãn course logistics and fit.
- Áp dụng bước phân loại nhị phân để chỉ lọc các bình luận có nội dung về độ khó của khóa học. Phân loại nhị phân giảm số lượng bình luận cần xử lý với nhiệm vụ trích xuất phức tạp hơn.
- Thực hiện phân tích cảm xúc và trích xuất ý kiến về độ khó của khóa học.



Hình II.6: Tìm phản hồi về mức độ khó (LLM: GPT-4).

8. Chain-of-Thought Reasoning

Để điều chỉnh đánh giá và đồng bộ kết quả với sở thích của con người, nhóm tác giả kiểm tra lý luận CoT cùng với kết quả đánh giá có cấu trúc cho bộ câu trả lời của khảo sát, và thực hiện sửa đổi các lời nhắc đánh giá theo kiểu lặp đi lặp lại. Một ví dụ về đầu ra CoT cho GPT-4 được hiển thị trong hình dưới đây. Khi lời nhắc được điều chỉnh dựa trên xác nhận của con người, kết quả đánh giá thay đổi một cách nhất quán, ngụ ý rằng lý luận CoT từ GPT-4 có thể hữu ích trong việc làm tinh chỉnh các đánh giá LLM.



Hình II.7: Ví dụ về lý luận GPT CoT để đánh giá quá trình trích xuất.

Chapter III

Kết quả

1. Đánh giá các nhiệm vụ NLP

1.1 Multi-label classification

Tính toán hệ số tương đồng Jaccard cho mỗi cặp annotator và GPT-4 để đánh giá độ đồng thuận.

Bảng III.1: Hệ số tương đồng Jaccard giữa human annotators and GPT-4.

	Annotator 1	**Annotator 2**	**GPT-4**	**Annotator 3**	**Annotator 4**
Annotator 1	-	81.27	80.18	83.37	82.35
Annotator 2	81.27	-	79.40	80.84	78.42
GPT-4	80.18	79.40	-	80.74	78.22
Annotator 3	83.37	80.84	80.74	-	81.18
Annotator 4	82.35	78.42	78.22	81.18	-

Kết quả:

- Hệ số tương đồng Jaccard trung bình giữa bốn annotator với nhau là 81.24%, đồng nghĩa với việc công việc này khó khăn đối với người đánh giá chuyên nghiệp.
- GPT-4 có hệ số tương đồng Jaccard trung bình với annotator là 80.60%.
- Kết quả cho thấy GPT-4 và con người có độ đồng thuận khá cao, đặc biệt là trong bối cảnh nhiệm vụ khó khăn như vậy.

Thực hiện phân loại đa nhãn bằng cả GPT và SetFit để so sánh hiệu suất.

Bảng III.2: Kết quả đánh giá về hàng đồng thuận với sự đồng ý trên tất cả các tags.

			Macro Average			**Micro Average**		
Model	**Jaccard**	**Average precision**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
GPT-4	92.97	93.91	89.88	90.59	89.78	93.66	93.26	93.46
GPT-3.5	72.61	74.79	69.34	82.18	72.63	72.36	84.48	77.96
SetFit	73.86	78.01	84.37	57.59	66.85	91.92	71.43	80.39

Bảng III.3: Kết quả đánh giá trên các hàng dùng nhãn đồng thuận

			Macro Average			Micro Average		
Model	**Jaccard**	**Average precision**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
GPT-4	80.17	81.53	73.91	88.38	79.69	78.32	89.70	83.63
GPT-3.5	63.00	65.18	60.42	79.79	65.75	60.31	83.45	70.02
SetFit	62.72	67.52	73.22	53.08	59.61	79.40	65.14	71.57

Đối với đánh giá hàng đồng thuận, kết quả zero-shot cho GPT-4 đạt được mong đợi của các phân loại tinh chỉnh. GPT-4 có hiệu suất tương tự như mô hình fine-tuned.

Các mô hình có điểm mạnh và điểm yếu riêng. Với SetFit có độ chính xác tương đối cao nhưng thu hồi thấp hơn, và GPT-3.5 ngược lại.

1.2 Binary classification

Phân loại xem 1250 ý kiến có chứa 'đề xuất cải tiến' hay không và so sánh với một đánh giá viên chuyên gia.

Bảng III.4: Kết quả nhiệm vụ phân loại nhị phân.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
GPT-4	95.20	96.20	95.39	95.79
GPT-3.5	90.16	89.01	93.35	91.14

Kết Quả: Cả hai mô hình GPT-4 và GPT-3.5 đều có hiệu suất tốt với độ chính xác cao. Phân loại nhị phân có thể được coi là đơn giản nhất trong số các nhiệm vụ NLP.

1.3 Extraction

Sử dụng 'đề xuất cải tiến' làm mục tiêu trích xuất và đánh giá độ chính xác của mô hình

Model	Missed Excerpts	Ambiguous Excerpts	Missed Existing Context	Irrelevant Excerpts	Implied Goal Focus	Non Quotes	Redundant Excerpts	Hallucinations
GPT-4	2.37	4.61	0.28	0.14	3.07	0.00	0.28	0.00
GPT-3.5	7.82	4.75	0.84	0.84	2.79	6.01	2.79	3.91

Hình III.1: Tỷ lệ lỗi (%) trích xuất cho 'đề xuất cải tiến' từ các nhận xét được phân loại là có chứa 'đề xuất cải tiến'

Kết quả:

- GPT-4 bao gồm một số ambiguous excerpts (trích đoạn mơ hồ). Tuy nhiên, những điều đó phổ biến nhất là do thiếu ngữ cảnh trong chính nhận xét, thay vì mô hình không làm được trích xuất bối cảnh đó.

- GPT-4 làm theo hướng dẫn rất chặt chẽ, kết quả không chứa hallucinations=0. Ngược lại, dầu ra của GPT-3.5 chứa hallucinations =3.91% và chỉnh sửa nhận xét với tỷ lệ 6%. .

- GPT-3.5 cũng bỏ lỡ các trích đoạn có liên quan hơn đáng kể so với GPT-4 (missed excerpts). Điều chỉnh prompt tuning có thể làm giảm tỷ lệ các lỗi này.

1.4 Sentimet analysis

Phân loại ý kiến về đề xuất và cải tiến thành 'negative', 'slightly negative', 'neutral', 'slightly positive', 'positive'.

Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
GPT-4	80.86	82.65	80.28	80.78
GPT-3.5	65.17	73.68	66.44	64.88
twitter-roberta-base-sentiment-latest	66.69	71.38	64.86	61.10

Hình III.2: Hiệu suất phân loại cảm xúc

Kết Quả: GPT-4 có hiệu suất cao hơn so với GPT-3.5 và một mô hình khác, nhất là trong việc phân loại 'tiêu cực'.Tuy nhiên, Kết quả đều thấp so với những gì đã thấy với mô hình được fine-tune trên các bộ dữ liệu trong miền, cho thấy tình cảm được thể hiện trong phản hồi khóa học của sinh viên có thể khác với phạm vi tình cảm được thể hiện trong dữ liệu đào tạo Internet của các mô hình này.

2. Chi phí và thời gian của LLM

- Thời gian cho các cuộc gọi mô hình của GPT-4 chậm hơn các mô hình OpenAI,GPT-4 có thời gian chạy khoảng 10 giây cho 100 nhận xét với hầu hết các nhiệm vụ.

- Chi phí sử dụng API OpenAI cho GPT-4 và GPT-3.5 phụ thuộc vào số lượng prompt tokens và số lượng tokens thông báo hoàn thành.

Task	GPT-4	GPT-3.5
binary classification	$0.93	$0.04
multi-label classification	$2.63	$0.12
multi-class classification	$2.13	$0.10
text extraction	$1.10	$0.05
text extraction evaluation	$3.01	$0.13
sentiment analysis	$1.17	$0.05
inductive thematic analysis	$0.13	$0.006

Hình III.3: Chi phí trên 100 bình luận cho GPT-4 và GPT-3.5.

Chapter IV

Kết luận

- Nghiên cứu sử dụng nhiệm vụ và dữ liệu thực tế với mục tiêu đánh giá phản hồi giáo dục phi cấu trúc. GPT-4 thể hiện hiệu suất không kém con người trong phân loại đa nhãn và suy luận, vượt trội so với các mô hình như SetFit. Tuy nhiên, đối với công việc kết hợp nhiều nhiệm vụ, cần đảm bảo hiệu suất đáng tin cậy trên từng nhiệm vụ để tránh lỗi tích tụ trong kết quả cuối cùng.

- Trong nhóm mô hình cao nhất, kỹ thuật gợi ý và điều chỉnh gợi ý đều ảnh hưởng đáng kể và có sự tương tác với mô hình được điều chỉnh. Mặc dù nghiên cứu tập trung vào gợi ý không giám sát trong giáo dục, kết quả chỉ ra rằng việc sử dụng gợi ý vài lần có thể cải thiện hiệu suất, đặc biệt là trong nhiệm vụ phân tích tâm trạng. Đối với các nhiệm vụ khó xác định, việc sử dụng ví dụ vài lần cũng có thể hữu ích trong quá trình hiệu chuẩn mô hình.

- Nghiên cứu trình bày ví dụ về quy luận suy nghĩ theo chuỗi của GPT-4, cho thấy sự rõ ràng và nhất quán của quy luận với nhãn hoặc kết quả đánh giá. Thay đổi gợi ý ảnh hưởng đến kết quả quy luận, và GPT-4 có hiệu suất cao trong việc đánh giá quy luận nguyên nhân. Mặc dù có nghi ngờ về cách mô hình đạt được câu trả lời, nhưng quy luận hợp lý có thể tăng cường tin tưởng và giảm thiểu cảm nhận về các mô hình như những hộp đen. Tuy nhiên, yêu cầu con người cung cấp lý do nhất quán cho mỗi quyết định là không khả thi với các bộ dữ liệu quy mô lớn.

- Nghiên cứu chỉ ra rằng GPT-4 có hiệu suất cao trên các nhiệm vụ cụ thể từ dữ liệu thực tế, làm cho nó là lựa chọn đáng tin cậy cho quy trình đa bước. Tuy nhiên, với sự tiến bộ nhanh chóng của các mô hình khác, quy trình này có thể trở nên phổ biến hơn trong tương lai. GPT-4 mở ra khả năng mở rộng cho nhiều loại học và khảo sát thông qua điều chỉnh nhỏ về danh mục và gợi ý.

Chapter V

Hạn chế và nghiên cứu trong tương lai

1. Hạn chế

- Phạm vi dữ liệu hẹp: Nghiên cứu chỉ sử dụng dữ liệu từ lĩnh vực cụ thể là các khóa học y sinh trực tuyến và bằng tiếng Anh. Điều này có thể giới hạn độ chủ động và đa dạng của mô hình, không thể chắc chắn rằng các kết quả có thể được tổng quát hóa cho các lĩnh vực khác hoặc ngôn ngữ khác.

- Thiếu sử dụng các kỹ thuật prompting: Mặc dù tác giả đề cập đến khả năng tăng hiệu suất thông qua các kỹ thuật prompting như self-consistency, reflection, và few-shot learning, nhưng các kỹ thuật này không được áp dụng trong nghiên cứu. Điều này có thể bỏ lỡ cơ hội tối ưu hóa hiệu suất của mô hình và chưa thể đánh giá đầy đủ ảnh hưởng của các kỹ thuật này.

- Các thử nghiệm được thực hiện với các tham số được điều chỉnh, bao gồm nhiệt độ và các tham số khác của mô hình.

- So sánh giới hạn: Nghiên cứu chỉ so sánh với mô hình SetFit và RoBERTa, hạn chếtrong việc khám phá các mô hình mới của OpenAI như Claude v2, Command và Llama 2. Điều này Có thể không phản ánh đầy đủ sự tiến bộ của mô hình so với các xu hướng và tiêu chuẩn mới trong lĩnh vực.

2. Hướng phát triển trong tương lai

- Sử dụng mô hình mã nguồn mở: sử dụng các mô hình mã nguồn mở để đảm bảo tính ổn định và khả năng kiểm soát cao. Cộng đồng có thể dễ dàng đóng góp và phát triển mô hình, đồng thời giảm rủi ro từ sự thay đổi không kiểm soát.

- Tối ưu hóa sử dụng các kỹ thuật prompting: Khám phá và triển khai các kỹ thuật prompting như self-consistency, reflection và few-shot learning để tối ưu hóa hiệu suất của mô hình.

- Mở rộng phạm vi so sánh: Đề xuất mở rộng so sánh với các mô hình OpenAI mới như Claude v2, Command và Llama 2.

- Tăng cường khả năng tổng hợp thông tin qua agent: Nghiên cứu có thể tập trung vào việc phát triển khả năng tổng hợp thông tin của agent để đáp ứng mục tiêu phân tích từ khảo sát. Từ đó có thể cung cấp công cụ mạnh mẽ hơn cho người phân tích để đạt được kết quả mong muốn và tối ưu hóa quy trình phân tích.