Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Phương pháp sử dụng mô hình ngôn ngữ lớn trong phân tích phản hồi khảo sát giáo dục

Báo cáo giữa kỳ

Nhóm B3.31:
21120224 - Lều Huy Đức
21120175 - Tô Ngọc Hân
21120071 - Nguyễn Thị Thanh Hoa

Trường Đại học Khoa học Tự nhiên VNU-HCM
Khoa Công nghệ thông tin
Môn: Nhập môn xử lý ngôn ngữ tự nhiên

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

Giới thiệu

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Bài báo trình bày một phương pháp 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 Chain-of-Thought (CoT) Reasoning để đưa ra giải thích kết quả cho mô hình. Tuy nhiên báo cáo còn một số hạn chế như việc sử dụng dữ liệu trong một lĩnh vực khá nhỏ và ngôn ngữ chỉ là tiếng Anh.

Cách tiếp cận trước đây

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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 (unstructured text), 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,...

Phân loại nhiệm vụ

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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

- *Phân loại bình luận:* Xác định các bình luận là đơn nhãn (Single-label) hay đa nhãn (Multi-label). 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.

LLM và nghiên cứu liên quan

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Việc phân tích phản hồi giáo dục trích xuất các ý kiến khảo sát 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 (text annotation) và phân loại văn bản (text classification), với kết quả kết hợp dựa trên sự biến đổi của các lời nhắc (prompt), tập dữ liệu (dataset), 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 (text annotation) hoặc cung cấp giải thích bằng ngôn ngữ tự nhiên ở mức tiếp cận hoặc tương đương với con người.

Ý nghĩa và mục tiêu

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Trong bài báo các tác giả đã:

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

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

Dữ liệu khảo sát được sử dụng để đánh giá

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Dữ liệu khảo sát:

- 2500 phản hồi được chọn ngẫu nhiên từ cuộc khảo sát cuối khóa về các khóa học y sinh học.
- Các khóa học bao gồm gen học, miễn dịch học và dược học.
- Sử dụng một cuộc khảo sát cuối khóa duy nhất với nhiều loại câu hỏi.

Câu hỏi khảo sát: Bốn câu hỏi mở để thu thập ý kiến.

- "Mô tả những phần tốt nhất của khóa học."[Q1]
- "Những phần nào tăng cường sự hiểu biết về các khái niệm nhất?"[Q2]
- "Làm thế nào chúng ta có thể cải thiện khóa học?"[Q3]
- "Cung cấp bất kỳ đề xuất, ý kiến hoặc ý tưởng nào khác."[Q4]

Dữ liệu khảo sát được sử dụng để đánh giá

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Xử lý dữ liệu khảo sát:

- Sử dụng Qualtrics để thu thập phản hồi.
- Sử dụng Pandas 2.0.1 để xử lý dữ liệu, loại bỏ khoảng trắng không cần thiết và các phản hồi trống (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. Quá trình kiểm tra thủ công các phản hồi giúp đảm bảo chất lượng của dữ liệu và loại bỏ bất kỳ thông tin riêng tư hoặc nhạy cảm nào mà các phương pháp tự động có thể đã bỏ sót.
- 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.

Phát triển hệ thống gắn thẻ khóa học

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Xây dựng hệ thống nhãn dán: Tác giả đã phát triển bộ 71 nhãn cho khóa học trực tuyến, sau đó làm rõ thông qua các vòng kiểm thử và tối ưu hóa bằng sự đóng góp của chuyên gia.

- Mỗi phản hồi từ cuộc khảo sát có thể liên quan đến nhiều chủ đề, tạo nên một nhiệm vụ phân loại đa nhãn.
- Bốn chuyên gia trong lĩnh vực đã gắn nhãn cho một tập 2000 phản hồi, với sự đồng thuận ban đầu tương đối thấp.
- Các nhãn được kết hợp, và áp dụng các best practices để thu được một tập nhãn nhỏ.

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.

Hình: Final tags and descriptions

Phát triển hệ thống gắn thẻ khóa học

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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.

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

LLM processing

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

- 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 gọi hàm để tạo ra đầu ra có cấu trúc cho mọi nhiệm vụ một cách đáng tin cậy.

So sánh với các mô hình khác

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Đánh giá hiệu suất mô hình bằng cách so sánh nó với các nhãn do con người tạo ra và các mô hình khác như SetFit, SentenceTransformers và RoBERTa.

- *SetFit:* Sử dụng SentenceTransformers cho phân loại văn bản.
- *SentenceTransformers:* Biểu diễn câu trong không gian vector, thường được sử dụng cho nhiệm vụ ngôn ngữ tự nhiên.
- *RoBERTa:* Mô hình ngôn ngữ sử dụng kiến trúc Transformer, đào tạo trên dữ liệu Tweets để phân tích cảm tính.

Các chỉ số đánh giá

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

- Công cụ nghiên cứu:
  + Sử dụng Scikit-learn 1.2.0 để kiểm tra thống kê, numpy 1.23.5 và Pandas 2.0.1 để phân tích dữ liệu.
  + Sử dụng Trọng số (Weights) và Xu hướng (Biases) để theo dõi kết quả đánh giá mô hình.
- Phân loại đa nhãn: So sánh kết quả của mô hình với các nhãn do con người tạo ra. Sử dụng hai phương pháp để tạo các nhãn này từ nhiều chú thích:
  + *Hàng đồng thuận (consensus rows):* Giữ lại các câu trả lời mà cả bốn người chú thích đồng ý về các thẻ.
  + *Nhãn đồng thuận (consensus labels):* Giữ lại tất cả các câu trả lời nhưng chỉ chọn các nhãn có sự đồng ý của đa số.
- Các chỉ số và phương pháp đánh giá hiệu suất mô hình:
  + Trong phân loại đa nhãn: Sử dụng các số liệu: Jaccard similarity coefficient, average precision per tag, average recall per tag, macro và micro average precision, recall, F1 score across all tags, Hamming loss và subset accuracy.
  + Trong phân loại nhị phân: Sử dụng accuracy, precision, recall và F1 score.
  + Trong trích xuất: Tạo một phiếu tự đánh giá và so sánh kết quả của mô hình với chú thích của con người.
  + Trong phân tích cảm xúc: So sánh với kết quả của bộ phân loại tình cảm RoBERTa được đào tạo trên 124 triệu tweet và với kết quả từ người chú thích của con người.

Phương pháp cho các quy trình LLM

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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, tương tự như việc mô hình chủ đề.
  + 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, phân loại đa nhãn là bước quan trọng đầu tiên, phân loại các phản hồi thành các hạng mục liên quan.
- Phân tích tập trung:
  + Extraction: Trích xuất là bước quan trọng để tập trung vào một mục tiêu cụ thể trong phân tích.
  + Multi-class Classification: Để cung cấp đầu ra cho phân tích và đánh giá tiếp theo, phân loại đa lớp có thể được sử dụng sau khi trích xuất.
  + Sentiment Analysis: Áp dụng cuối cùng để xác định sự tích cực hoặc tiêu cực trong các đoạn văn.

Mặc dù các phản hồi của mô hình được hiển thị dưới dạng JSON, nhưng chỉ văn bản đầu ra có liên quan được hiển thị để đơn giản và rõ ràng.

Phương pháp cho các quy trình LLM

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Objective	Question	NLP Tasks	Notes
High-level initial analysis	What did students say (and how did they feel about the course)?	Multi-label classification, inductive thematic analysis, sentiment analysis	Depends on whether analysis is top-down (using pre-determined labels or areas of interest) or bottom-up (deriving themes from scratch based on student comments)
Answering a focused question	What did students say about x (particular focus)?	Extraction	Results are amenable to multi-class classification or inductive thematic analysis
Quantification of textual survey responses	How many comments were there on each aspect?	Classification (binary, multi-label, or multi-class)	Helps the person performing analysis find themes of greater importance

Hình: NLP tasks that may be used for analysis of textual survey responses.

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Ví dụ về quy trình làm việc

Phân tích chủ đề bằng phương pháp quy nạp(từ dưới lên)

Qua 3 bước LLM:

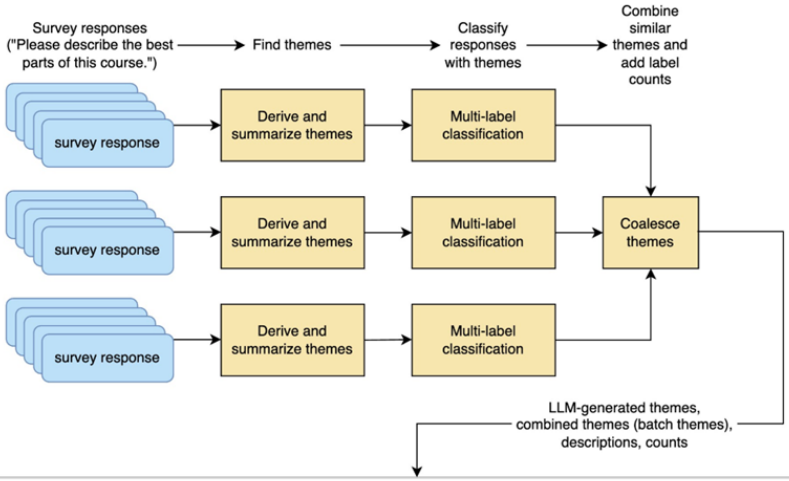
- Các chủ đề được xác định và tóm tắt cho từng lô nhận xét. Mỗi lô được thiết lập với kích thước phù hợp với ngữ cảnh của mô hình sử dụng.
- Phân loại các nhận xét bằng cách sử dụng các chủ đề được xác định trong bước thứ nhất.
- Hợp nhất Các chủ đề thành bộ chủ đề cuối cùng và số lượng nhãn được tổng hợp từ các chủ đề được kết hợp.

Trong phương pháp định tính, các bước 1 và 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.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Phân tích chủ đề bằng phương pháp quy nạp(từ dưới lên)



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

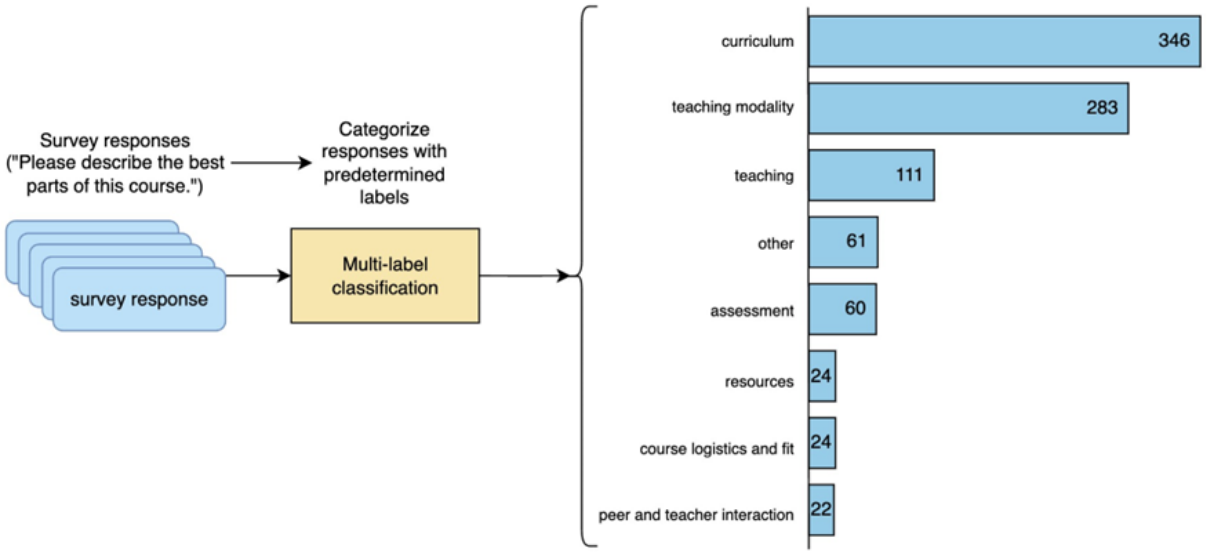
Hình: Quy trình lấy chủ đề từ nhận xét của sinh viên (kết quả hiển thị bằng GPT-4)

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Phân tích bằng cách phân loại các nhận xét của sinh viên (từ trên xuống)

Phân loại đa nhãn của các câu câu trả lời cho khảo sát, sử dụng những nhãn được xác định trước và phát triển cho đề tài đã được thử nghiệm trên 625 câu nhận xét từ câu hỏi Q1 (“Nêu những điểm tốt nhất trong khóa học?”) kết quả nhận được như hình hình dưới đây. 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.



Hình: Phân loại đa nhãn từ nhận xét (kết quả được hiển thị bằng GPT-4).

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

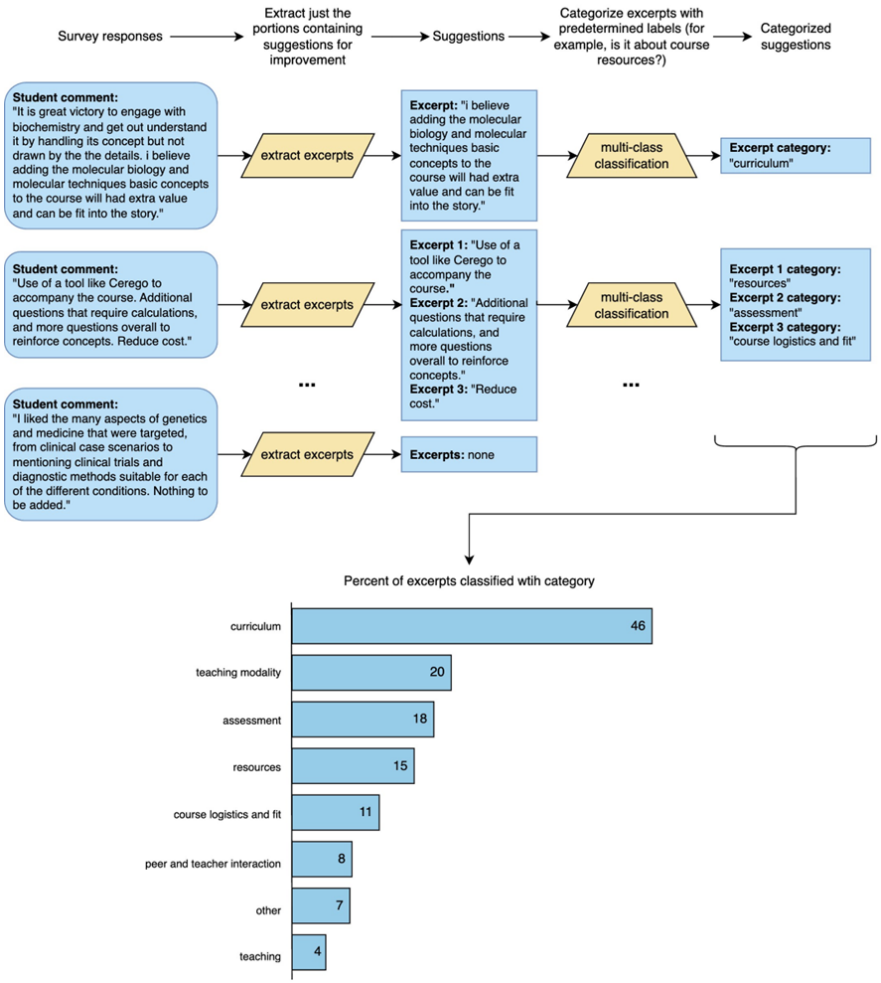
Cải tiến từ nhận xét của sinh viên

Quy trình thực hiện tìm kiếm và định lượng các cách cải thiện khóa học gồm: trích xuất các trích đoạn có liên quan, phân loại nhiều lớp dựa vào nhãn để tạo định lượng cũng như định tuyến cho những nhận xét cho các bên 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. 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.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Cải tiến từ nhận xét của sinh viên



Hình: Quy trình tìm kiếm đề xuất cải tiến (kết quả hiển thị bằng GPT-4).

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Ví dụ về quy trình làm việc

Ví dụ - Sinh viên quan tâm đến nội dung hoặc chủ đề nào khác?

Quy trình:

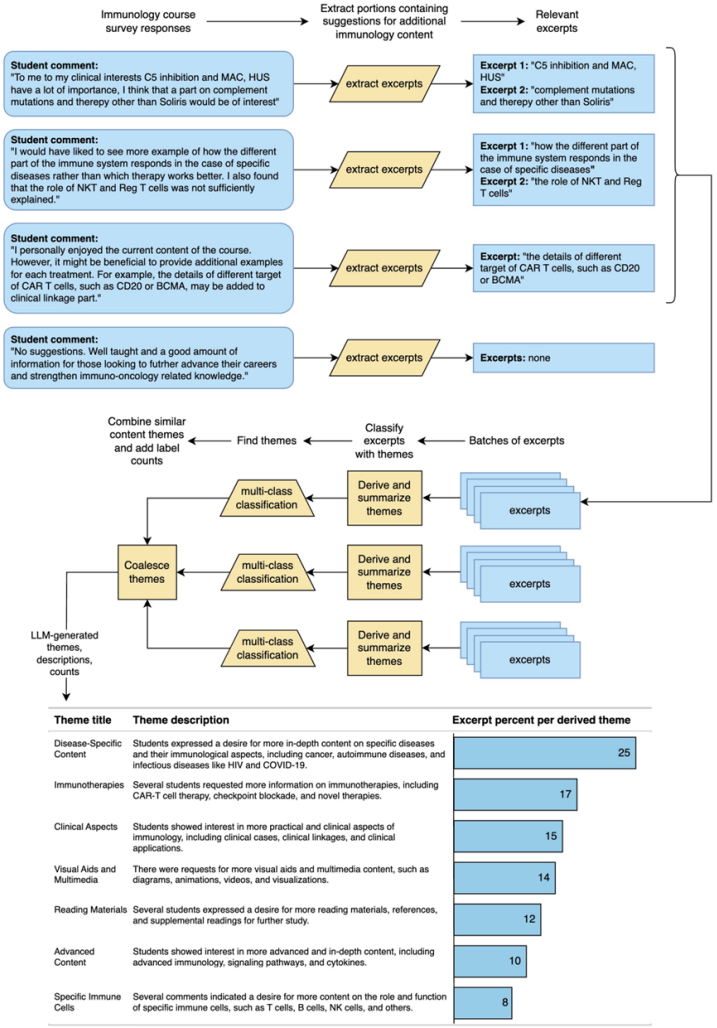
- Khảo sát khóa học về miễn dịch: Tập trung vào câu hỏi liên quan đến khóa học miễn dịch (Q3 và Q4).
- Trích xuất các phần chỉ chứa gợi ý nội dung hoặc chủ đề mới từ câu phản hồi của sinh viên. Các chủ đề gợi ý nội dung sau đó được rút ra và tóm tắt từ các đoạn trích; điều này được thực hiện theo lô nếu chúng không thể phù hợp trong một lời nhắc duy nhất cho LLM.
- 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.

GPT-4 có khả năng tìm thấy gợi ý nội dung, ngay cả trong lĩnh vực y sinh, nhờ vào đa dạng và lượng dữ liệu đào tạo trước của mô hình.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Ví dụ - Sinh viên quan tâm đến nội dung hoặc chủ đề nào khác?



Hình: : Tìm gợi ý cho nội dung khóa học miễn dịch từ nhận xét (kết quả hiển thị bằng GPT-4).

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Ví dụ về quy trình làm việc

Ví dụ - Sinh viên phản hồi về giảng dạy và giải thích như thế nào?

Quy trình:

- Sử dụng phân loại đa nhãn như một bước khởi đầu, sử dụng các nhãn có sẵn đã được phát triển.
- 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.
- Định rõ mục tiêu thông qua zero-shot prompting để giảm thiểu lỗi.

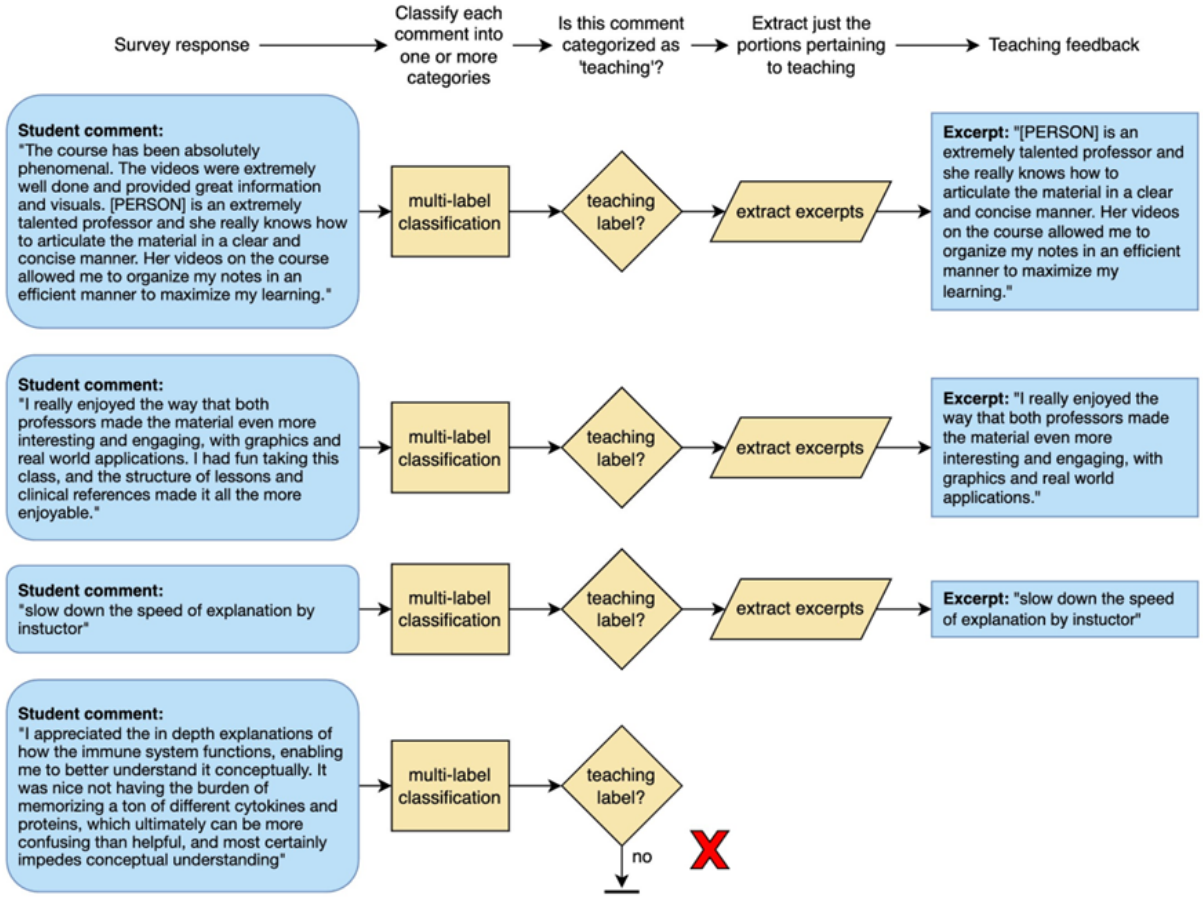
Bước trích xuất làm tinh tế thông tin, tuy nhiên có thể xuất hiện lỗi (như lọc ra các bình luận liên quan mặc dù chúng có tham chiếu đến chất lượng giải thích).

Quy trình giúp lọc và phân tích các ý kiến chỉ liên quan đến giảng dạy, đồng thời giải quyết các lỗi tiềm ẩn.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Ví dụ - Sinh viên phản hồi về giảng dạy và giải thích như thế nào?



Hình: Phản hồi về giảng dạy và giải thích. Dấu 'x' màu đỏ cho biết lỗi của mô hình.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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

Quy trình:

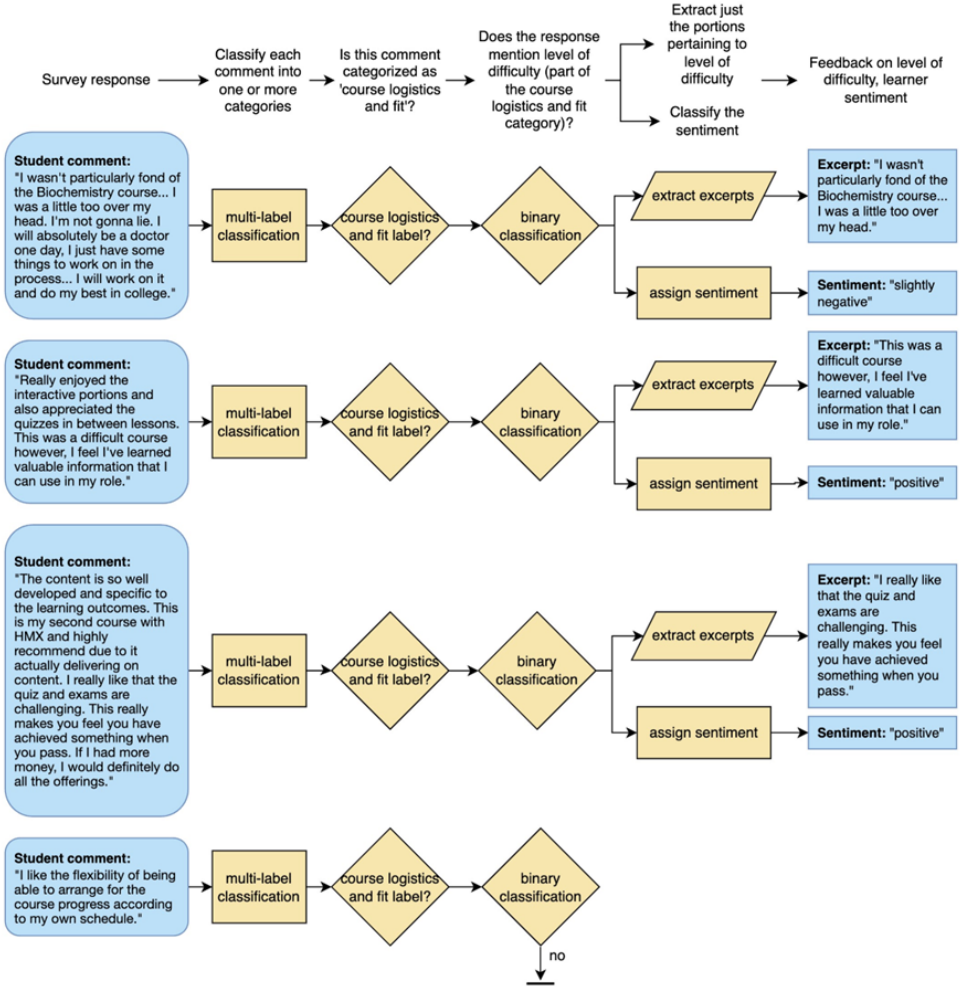
- Sử dụng nhãn có sẵn thông qua phân loại đa nhãn.
- Lọc ý kiến dưới nhãn 'logistics và fit'.
- Áp dụng bước phân loại nhị phân để chỉ lọc các bình luận có nội dung về độ khó của khóa học.
- Thực hiện phân tích cảm xúc và trích xuất ý kiến về độ khó của khóa học.

Phân loại nhị phân giảm số lượng bình luận cần xử lý với nhiệm vụ trích xuất phức tạp hơn. Kết quả của phân tích cảm xúc cho thấy Phân tích cảm xúc có thể là một thách thức trong lĩnh vực học trực tuyến y sinh do bối cảnh phức tạp của khóa học và nhiều chủ đề trong một ý kiến có thể khác biệt đáng kể so với dữ liệu đào tạo trước công bố của mô hình.

Ví dụ về quy trình làm việc

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

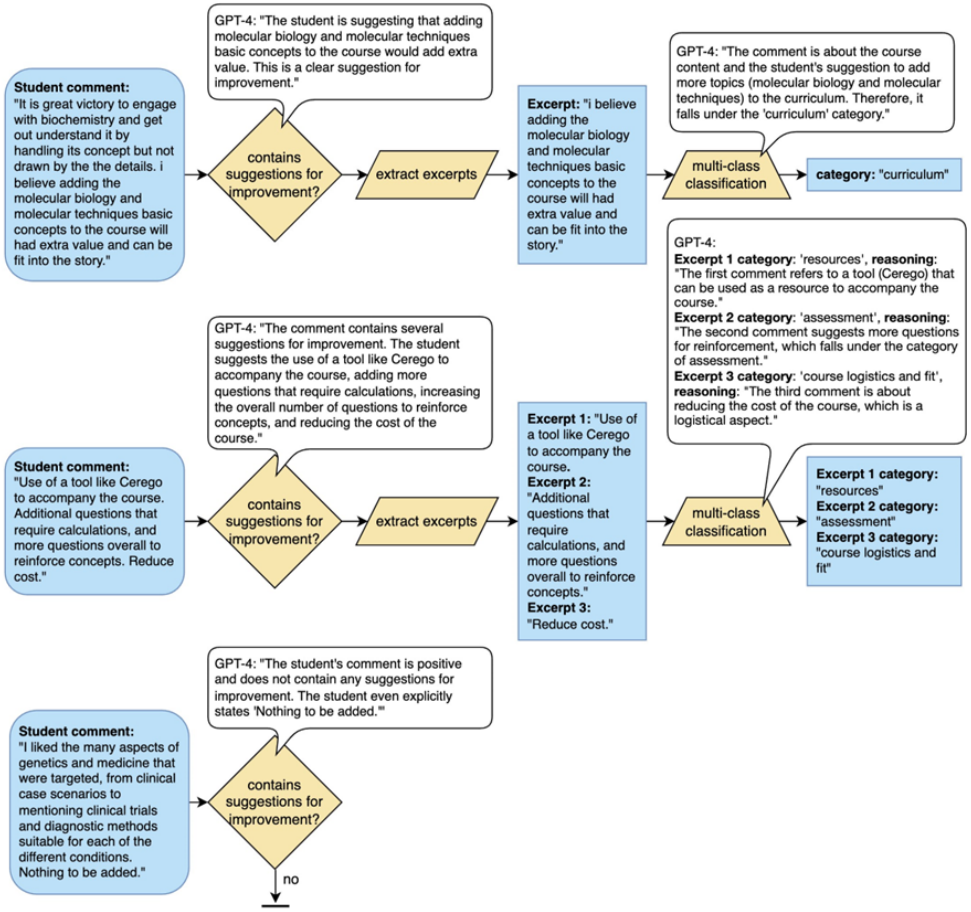
Ví dụ - Sinh viên cảm thấy thế nào về độ khó của khóa học?



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

Chain-of-Thought Reasoning

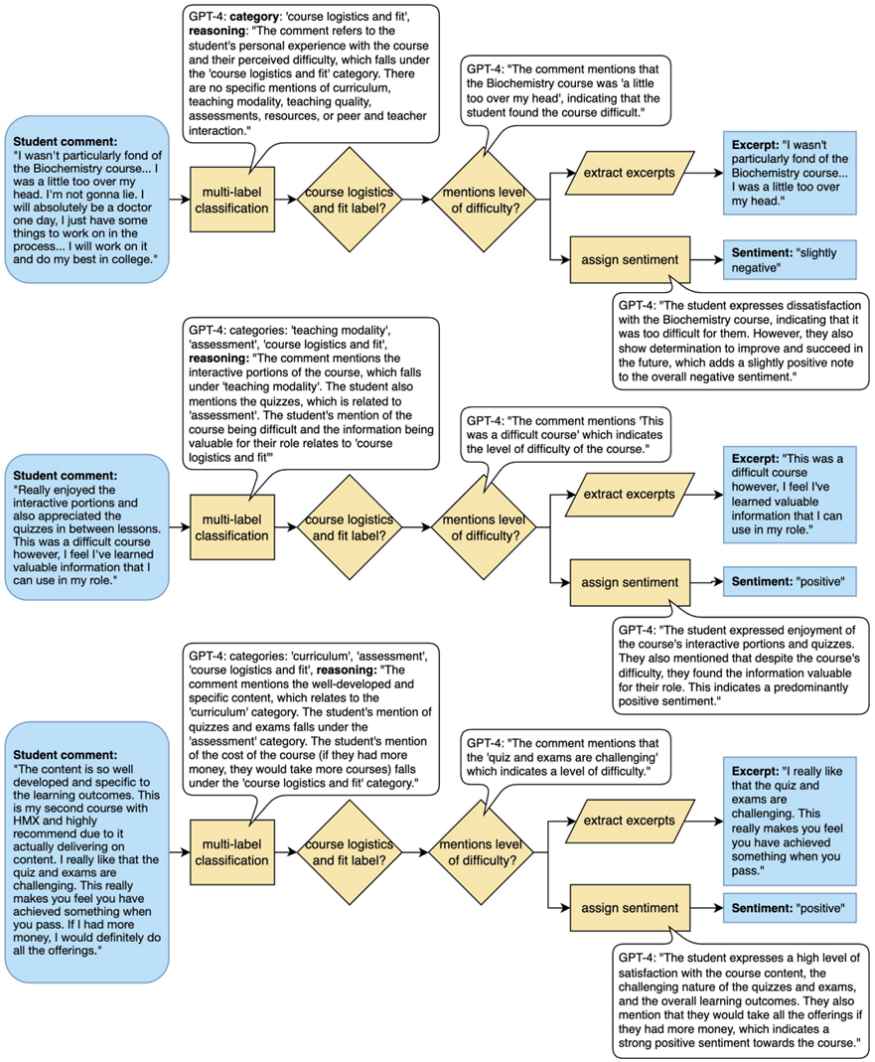
Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương



Hình: Ví dụ về lý luận GPT-4 CoT cho phân loại nhị phân và phân loại đa lớp liên quan đến nhiệm vụ tìm kiếm đề xuất cải tiến.

Chain-of-Thought Reasoning

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

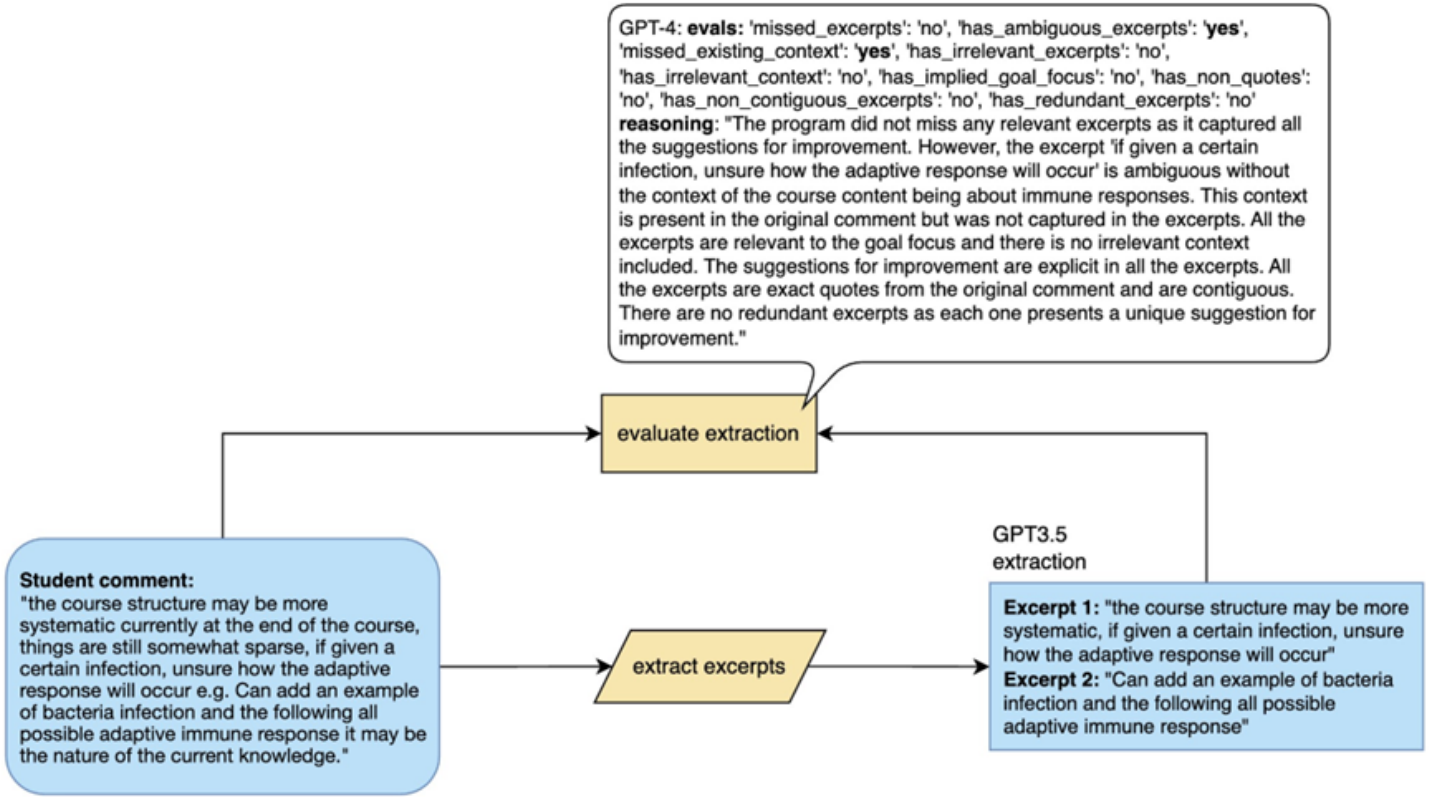


Hình: Ví dụ về lý luận GPT-4 CoT để phân loại đa nhãn, phân loại nhị phân và phân tích cảm xúc liên quan đến nhiệm vụ tìm hiểu cảm nhận của sinh viên về mức độ khó của khóa học.

Chain-of-Thought Reasoning

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

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

Đánh giá các nhiệm vụ NLP

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Multi-label classification metrics

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

	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	-

Hình: Hệ số tương đồng Jaccard giữa human annotators and GPT-4.

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.

Đánh giá các nhiệm vụ NLP

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Multi-label classification metrics

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

			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	84.48f	77.96
SetFit	73.86	78.01	84.37	57.59	66.85	91.92	71.43	80.39

Hình: 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	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

Hình: Kết quả đánh giá trên các hàng dùng nhãn đồng thuận.

Đánh giá các nhiệm vụ NLP

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Binary classification metrics

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

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

Hình: Hiệu suất nhiệm vụ phân loại nhị phân.

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

Đánh giá các nhiệm vụ NLP

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Extraction evaluation

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: Tỉ lệ lỗi (%) của extraction for 'đề xuất cải tiến' từ các nhận xét được phân loại là có chứa 'đề xuất cải tiến' .

Kết quả: GPT-4 có độ chính xác cao và không chứa hallucinations trong kết quả. GPT-3.5 có tỷ lệ hallucinations và chỉnh sửa văn bản, đặt ra nhu cầu cẩn trọng khi sử dụng nó cho trích xuất.

Đánh giá các nhiệm vụ NLP

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Sentiment analysis metrics

Phân loại ý kiến về đề xuất và cải tiến thành 'tích cực', 'hơi tích cực', 'trung lập', 'hơi tiêu cực', và 'tiêu cực'.

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: Kết quả phân lọai nhận xét

Kết quả: GPT-4 có hiệu suất cao hơn so với GPT-3.5 và một mô hình khác.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.

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Chi phí và thời gian của LLM

GPT-4 có thời gian chạy khoảng 10 giây cho 100 ý kiến với hầu hết các nhiệm vụ. Tổng cộng, các mô hình GPT-4 và GPT-3.5 có hiệu suất tốt trong nhiều 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: Chi phí trên 100 nhận xét cho GPT-4 và GPT-3.5

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Kết luận

Kết quả của các LLM riêng lẻ thay đổi đáng kể và GPT-4 hoạt động ngang bằng với con người ngay cả với zero-shot prompting.

Kết quả LLM phụ thuộc rất nhiều vào prompt. Điều này đề cập đến sự quan trọng của việc xem xét prompt như một yếu tố quan trọng khi đánh giá hiệu suất của các mô hình.

Kiểm tra quy luật suy nghĩ theo chuỗi (chain-of-thought reasoning) của một mô hình có thể có nhiều ứng dụng. Điều này mở ra khả năng sử dụng thông tin về quy luật suy nghĩ để cải thiện hiểu biết và tin tưởng trong quá trình đánh giá kết quả của mô hình.

Các mô hình ngôn ngữ lớn (LLMs) tạo nên cốt lõi của một phương pháp tiếp cận linh hoạt và có thể kết hợp. GPT-4 được nhấn mạnh là mô hình xuất sắc, có khả năng hoạt động hiệu quả trên nhiều nhiệm vụ và cung cấp cơ sở cho sự linh hoạt và kết hợp trong ứng dụng thực tế.

Mục lục

1 Giới thiệu bài báo
- Giới thiệu
- Cách tiếp cận trước đây
- Phân loại nhiệm vụ
- LLM và nghiên cứu liên quan
- Ý nghĩa và mục tiêu

2 Dữ liệu Khảo sát và Hệ thống Gắn thẻ
- Dữ liệu khảo sát được sử dụng để đánh giá
- Phát triển hệ thống gắn thẻ khóa học

3 Phương pháp áp dụng LLM vào bài toán
- LLM processing
- So sánh với các mô hình khác
- Các chỉ số đánh giá
- Phương pháp cho các quy trình LLM
- Ví dụ về quy trình làm việc
- Chain-of-Thought Reasoning
- Đánh giá các nhiệm vụ NLP
- Chi phí và thời gian của LLM

4 Kết luận

5 Hạn chế và nghiên cứu trong tương lai
- Hạn chế trong nghiên cứu
- Hướng phát triển tương lai

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Hạn chế trong nghiên cứu

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.

So sánh giới hạn: Nghiên cứu chỉ so sánh với mô hình SetFit và RoBERTa, hạn chế trong 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.

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Hướng phát triển 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.

Giới thiệu bài báo
Giới thiệu
Cách tiếp cận trước đây
Phân loại nhiệm vụ
LLM và nghiên cứu liên quan
Ý nghĩa và mục tiêu
Dữ liệu Khảo sát và Hệ thống Gắn thẻ
Dữ liệu khảo sát được sử dụng để đánh giá
Phát triển hệ thống gắn thẻ khóa học
Phương pháp áp dụng LLM vào bài toán
LLM processing
So sánh với các mô hình khác
Các chỉ số đánh giá
Phương

Thank you for your attention!