# Bài 1: HiQA: A Hierarchical Contextual Augmentation RAG for Massive Documents QA(<https://arxiv.org/pdf/2402.01767.pdf> )

Bài viết giới thiệu về mô hình LLM hiện đang phổ biến, nhưng nó gặp các thách thức khác như vấn đề ảo giác, tính kịp thời của dữ liệu về vấn đề an ninh. Có một giải pháp cho vấn đề đó là dùng RAG tuy nhiên cách dùng RAG sẽ có hạn chế với các văn bản phi cấu trúc phức tạp như dưới dạng bảng, hình ảnh. Vì vậy nhóm đề xuất sử dụng PDFTriage để giải quyết các nhiệm vụ của QA bằng cách trích xuất và chuyển đổi chúng thành siêu dữ liệu có thể truy xuất được. Việc sử dụng PDF Triage có thể có thể được mô tả là một kỹ thuật phân vùng cứng. Chiến lược này tương đương với việc cắt tỉa và lựa chọn các tập hợp con trước khi truy xuất thông tin. Như là các biện pháp được thực hiện để tinh chỉnh độ chính xác của việc truy xuất bằng cách giảm kích thước của các phân đoạn. Tuy nhiên, trong các kịch bản liên quan đến các nhiệm vụ phức tạp như tìm kiếm tài liệu chéo hoặc so sánh nhiều sản phẩm, rủi ro thông tin quan trọng bị mất trước khi truy xuất trong các phương pháp phân vùng cứng.Vì những điều trên nhóm đề xuất tăng cương phân đoạn tài liệu bằng siêu dữ liêu, điều chỉnh sự phân bố🡪các phân đoạn sẽ có mối quan hệ với nhau. Bài báo nhắc tới việc trả lời câu hỏi nhiều tài liệu có nhiều thách thức như sự tương đồng giống nhau giữa các tài liệu, vấn đề về sự truy xuất trong tài liêu. 🡪 Bài báo đề xuất HiQA(Hierarchical

Contextual Augmentation RAG for Massive Documents QA)

Đề xuất HiQA (Hierarchical RAG tăng cường theo ngữ cảnh cho QA tài liệu lớn) là một giải pháp nhằm giải quyết thách thức của việc truy xuất kiến thức từ các tài liệu lớn mà không thể phân biệt được. Phương pháp này tích hợp phân tích và chuyển đổi tài liệu mới, sử dụng phương pháp tăng cường nhúng dựa trên siêu dữ liệu và cơ chế truy xuất Đa tuyến đường phức tạp để nâng cao độ chính xác và mức độ phù hợp của việc truy xuất kiến thức trong môi trường đa tài liệu.

Cụ thể, các đóng góp chính của bài viết này bao gồm:

Xác định vấn đề thực tế: Bài viết nhấn mạnh vấn đề quan trọng về tài liệu lớn không thể phân biệt được mà các phương pháp truy xuất thông thường gặp khó khăn trong việc giải quyết.

Đề xuất hệ thống HiQA: Sử dụng siêu dữ liệu và phương pháp phân vùng mềm để xử lý vấn đề tài liệu không thể phân biệt được, HiQA được đề xuất như một giải pháp hiệu quả

Phát hành tiêu chuẩn MasQA: Bao gồm nhiều loại tài liệu đa tài liệu và nhiều câu hỏi để tạo điều kiện thuận lợi cho việc nghiên cứu và đánh giá trong môi trường MDQA (Multilingual Document Question Answering).

Những công trình này đánh dấu sự tiến bộ trong việc xử lý các nhu cầu phức tạp của việc truy xuất kiến thức từ các tài liệu lớn và không phân biệt được, đồng thời mở ra cơ hội cho sự phát triển và hợp tác trong cộng đồng nghiên cứu.

Kết quả:

A screenshot of a document

Description automatically generated

A table with numbers and text

Description automatically generated

Bài 2. Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition(<https://arxiv.org/pdf/2401.12599.pdf>)

LLM được đào tạo dựa trên nguồn dữ liệu từ các nguồn internet ví dụ như web,sách, tin tức . Tuy nhiên khi sử dụng các kiến thức liên quan đến vấn đề về miền chẳng hạn như miền kiến thức về tập tin của mình đưa vào thì tập dữ liệu này mang tính riêng tư và ko phải 1 phần được train của LLM. Giải quyết vấn đề này ta có 1 cách tiếp cận đó là RAG. RAG trả lời câu hỏi theo 4 bước là người dùng đề xuất một truy vấn, hệ thống lấy nội dung liên quan từ cơ sở tri thức riêng tư, kết hợp nó với truy vấn của người dùng dưới dạng bối cảnh và cuối cùng yêu cầu LLM đưa ra câu trả lời. Tuy nhiên, quá trình truy xuất từ các tệp PDF gặp nhiều thách thức. Các vấn đề thường gặp bao gồm trích xuất văn bản không chính xác và sự lộn xộn trong mối quan hệ hàng-cột của các bảng bên trong tập tin PDF. Vì vậy, trước RAG, chúng ta cần chuyển đổi các tài liệu lớn thành nội dung có thể truy xuất được. Các chuyển đổi bao gồm một số bước:

• Phân tích và phân chia tài liệu. Nó liên quan đến việc trích xuất các đoạn văn, bảng biểu và nội dung khác các khối, sau đó chia nội dung được trích xuất thành các phần để truy xuất tiếp theo.

• Nhúng. Nó chuyển đổi các đoạn văn bản thành các vectơ có giá trị thực và lưu trữ chúng trong cơ sở dữ liệu. Vì mỗi bước này có thể dẫn đến mất thông tin nên tổn thất gộp có thể đáng kể. ảnh hưởng đến hiệu quả phản hồi của RAG.

Bài viết này chủ yếu giải quyết câu hỏi liệu chất lượng phân tích và phân đoạn PDF có tốt không? ảnh hưởng đến kết quả của RAG. Chúng ta sẽ khám phá những thách thức, phương pháp và trường hợp thực tế những nghiên cứu liên quan đến vấn đề này.

Thiết lập đánh giá:

A paper with text and images

Description automatically generated

Kết quả so sánh

A table with numbers and text

Description automatically generated

Bài 3: RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots∗ (<https://arxiv.org/pdf/2403.01193.pdf> )

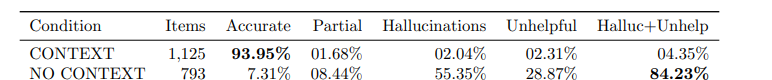
Từ khi ra mắt vào tháng 11 năm 2022, ChatGPT và các hệ thống tương tự đã có ảnh hưởng lớn đến nhận thức của công chúng về trí tuệ nhân tạo (AI). Những mô hình này có khả năng tạo ra các phản hồi mạch lạc và phù hợp với ngữ cảnh. Tuy nhiên, mặc dù có hiệu suất vượt trội, những mô hình này vẫn gặp phải vấn đề ảo giác, trong đó thông tin tạo ra có vẻ đáng tin cậy nhưng không chính xác.

Một ví dụ cụ thể về hệ lụy của vấn đề này là vụ kiện vào tháng 6 năm 2023, khi một luật sư sử dụng ChatGPT để thực hiện nghiên cứu pháp lý và vô tình trích dẫn những trường hợp không tồn tại. Mặc dù tòa án không phát hiện lỗi trong trường hợp này, nhưng nó làm nổi bật nhu cầu hạn chế ảo giác trong các mô hình AI để ngăn chặn những vấn đề tương tự trong tương lai. Tin tức về vấn đề này đã lan truyền rộng rãi trên mạng xã hội và được các tờ báo lớn như New York Times đưa tin.

Đào tạo người dùng để áp dụng phương pháp tư duy phê phán đối với đầu ra của các mô hình AI có vẻ là một giải pháp hợp lý. Tuy nhiên, nó không giải quyết được hoàn toàn vấn đề về tin tưởng và tương tác của con người với các hệ thống này. Một số nghiên cứu gần đây đã chỉ ra những hạn chế của việc này và cần phải có những phương pháp khác để giảm ảo giác trong các mô hình AI.

Thế hệ Tăng cường Truy xuất (RAG) đã được chứng minh là có thể giảm ảo giác bằng cách cung cấp ngữ cảnh bổ sung cho lời nhắc. Tuy nhiên, RAG vẫn đối mặt với những hạn chế khi xử lý thông tin không chính xác so với thông tin được đào tạo. Việc cung cấp ngữ cảnh có thể cải thiện khả năng tạo ra phản hồi liên quan nhưng vẫn cần phải giải quyết các vấn đề về tính chính xác.

Trong bài viết này, chúng tôi điều tra hiệu quả của việc sử dụng ngữ cảnh như được áp dụng trong RAG để giảm ảo giác và cải thiện chất lượng của đầu ra so với việc không sử dụng ngữ cảnh. Chúng tôi cũng xem xét những trường hợp mà việc cung cấp ngữ cảnh không đạt hiệu quả mong đợi, không chỉ về chất lượng mà còn về việc giảm ảo giác.



Với những kết quả này và các kết quả tương tự (ví dụ: [8, 4, 5]), chúng tôi có thể tự tin nói rằng RAG có thể cải thiện đáng tin cậy

tính chính xác của thông tin được cung cấp bởi LLM. Ở nhiều khía cạnh, vấn đề một lần nữa quay trở lại việc đảm bảo

chất lượng của kết quả tìm kiếm được cung cấp cho LLM.

Nhưng không phải trong mọi trường hợp. Trong [4], hiệu suất RAG được đánh giá dựa trên Độ bền tiếng ồn, Âm tính

Từ chối, tích hợp thông tin và tính mạnh mẽ phản thực tế. Trong mỗi trường hợp, kết quả RAG là

tốt hơn đáng kể so với kết quả không tăng cường, nhưng chúng vẫn không phải lúc nào cũng thể hiện chính xác dữ liệu

trong văn bản được truy xuất.

Bài 4: Financial Report Chunking for Effective Retrieval Augmented Generation(<https://arxiv.org/pdf/2402.05131.pdf>)

Trong thời đại của trí tuệ nhân tạo đang phát triển nhanh chóng, việc hiểu và xử lý tài liệu đã trở thành một thách thức quan trọng. Để đáp ứng nhu cầu này, các phương pháp hiện đại kết hợp cả thị giác máy tính và xử lý ngôn ngữ tự nhiên để phân tích và xác định các thành phần khác nhau trong một tài liệu. Trong bối cảnh này, các mô hình ngôn ngữ lớn (LLM) như GPT-4 đã đem lại những tiến bộ đáng kể trong việc hiểu và tạo ra ngôn ngữ tự nhiên, mở ra nhiều ứng dụng mới.

Tuy nhiên, các LLM thường gặp hạn chế do phụ thuộc vào kích thước và chất lượng dữ liệu mà chúng được huấn luyện. Một trong những hạn chế đáng chú ý là cửa sổ ngữ cảnh hạn chế của chúng, điều này ảnh hưởng đến khả năng hiểu sâu rộng của chúng về nội dung tài liệu lớn.

Gần đây, phương pháp Truy xuất Tăng cường (RAG) đã được phát triển để giải quyết vấn đề này bằng cách phục hồi thông tin thực tế từ LLM để tạo ra các câu trả lời chính xác hơn và đáng tin cậy hơn. Trong RAG, thông tin từ người dùng được sử dụng để truy xuất các phân đoạn tài liệu, và kết quả truy xuất này được kết hợp với LLM để tạo ra câu trả lời. Điều này giúp hạn chế các câu trả lời dựa trên dữ liệu không chính xác hoặc thiếu ngữ cảnh.

Trong nghiên cứu này, chúng tôi tập trung vào việc phân đoạn các báo cáo tài chính từ Ủy ban Chứng khoán và Giao dịch Hoa Kỳ (SEC), bao gồm các loại báo cáo như 10-K, 10-Q và 8-K. Việc này không chỉ giúp chúng tôi hiểu rõ hơn về sức khỏe tài chính và hoạt động của các công ty mà còn cung cấp cái nhìn sâu sắc về cách xử lý tài liệu và truy xuất thông tin hiệu quả từ chúng.

Thông qua việc nghiên cứu về cách phân đoạn và xử lý tài liệu, chúng tôi hy vọng sẽ đóng góp vào việc phát triển các phương pháp hiểu tài liệu hiệu quả hơn, đặc biệt là trong ngữ cảnh của việc sử dụng RAG và các phương pháp tương tự. Điều này có thể giúp cải thiện tính chính xác và đáng tin cậy của các hệ thống AI trong việc xử lý và hiểu biết văn bản phức tạp.

A screenshot of a document

Description automatically generated