RESEARCH ARTICLE

KHẢO SÁT HỆ THỐNG KHUYẾN NGHỊ CÓ GIẢI THÍCH TỪ PHƯƠNG PHÁP ỨNG DỤNG ĐẾN THỬ THÁCH

Trần Nguyễn Minh Hiếu[1], Trần Văn Lăng[2,*], Nguyễn Quốc Huy[1]

[1] Trường Đại học Sài Gòn
[2] Trường Đại học Ngoại ngữ - Tin học TP.HCM
tnmhieu@sgu.edu.vn, langtv@huflit.edu.vn, nqhuy@sgu.edu.vn

TÓM TẮT— Hệ thống khuyến nghị có giải thích (Explainable Recommendation Systems – ERS) không chỉ cung cấp các đề xuất phù hợp mà còn kèm theo lời giải thích rõ ràng, giúp tăng tính minh bạch và độ tin cậy từ người dùng. Bài viết này khảo sát các phương pháp giải thích chính trong ERS, bao gồm giải thích dựa trên mô hình, giải thích sau khi ra kết quả và giải thích theo hướng người dùng, đồng thời phân tích các công trình tiêu biểu ứng dụng SHAP, LIME, PEPLER-D, GaVaMoE và G-Refer. Kết quả khảo sát chỉ ra nhiều thách thức quan trọng, như hạn chế trong mô hình hóa sở thích phức tạp, chi phí tính toán cao khi sử dụng LLMs, hiện tượng "ảo giác" trong giải thích, thiếu các tiêu chí đánh giá định lượng, cũng như rủi ro bảo mật dữ liệu cá nhân. Để khắc phục các thách thức, các hướng phát triển được đề xuất bao gồm: tối ưu hóa chi phí tính toán và khả năng mở rộng, đảm bảo tính nhất quán và chất lượng giải thích, cá nhân hóa lời giải thích theo người dùng, kết hợp đa phương pháp giải thích để tăng tính toàn diện, và xây dựng các cơ chế bảo mật và đạo đức trong hệ thống khuyến nghị có giải thích. Nghiên cứu này cung cấp cái nhìn tổng quan hệ thống ERS, định hướng cho các công trình tiếp theo nhằm nâng cao chất lượng và ứng dụng thực tiễn của ERS trong nhiều lĩnh vực.

Từ khóa— Hệ thống khuyến nghị có giải thích, giải thích dựa trên mô hình, giải thích sau khi ra kết quả, giải thích theo hướng người dùng.

I. GIỚI THIỆU

Hệ thống khuyến nghị (Recommender System – RS) đã xuất hiện từ giữa những năm 1990 nhằm giải quyết vấn đề quá tải thông tin trong môi trường số. Ngày nay, với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), các hệ thống RS đã có thể phân tích lượng dữ liệu khổng lồ từ hành vi và sở thích cá nhân để đưa ra các gợi ý phù hợp với từng người dùng. Hầu hết các công ty công nghệ lớn đều đã tích hợp RS vào sản phẩm của mình nhằm tối ưu hóa trải nghiệm người dùng: Amazon đề xuất sản phẩm dựa trên lịch sử mua sắm, Netflix và YouTube sử dụng RS để gợi ý nội dung phù hợp với sở thích cá nhân, trong khi Facebook ứng dụng RS để đề xuất bạn bè, trang theo dõi hoặc nội dung quan tâm [1].

Một nhánh quan trọng của RS là hệ thống khuyến nghị có giải thích ERS, được phát triển nhằm cung cấp sự minh bạch và tính giải thích trong quá trình ra quyết định của hệ thống. Không giống như các hệ thống RS truyền thống chỉ đơn thuần đưa ra khuyến nghị mà không giải thích lý do, ERS có khả năng cung cấp thông tin chi tiết về cách hệ thống tạo ra các đề xuất, giúp người dùng hiểu được cơ sở của các quyết định này. Điều này không chỉ giúp tăng cường sự tin tưởng của người dùng đối với hệ thống mà còn nâng cao mức độ chấp nhận và tương tác với các khuyến nghị được đưa ra. Đặc biệt, trong các lĩnh vực quan trọng như tài chính, y tế và giáo dục, sự minh bạch trong quyết định có thể ảnh hưởng trực tiếp đến hiệu quả sử dụng hệ thống, hỗ trợ người dùng đưa ra lựa chọn chính xác và hợp lý hơn [2].

Việc áp dụng ERS mang lại nhiều lợi ích đáng kể trong việc tối ưu hóa trải nghiệm người dùng cũng như cải thiện chất lượng khuyến nghị. Trước hết, tính minh bạch mà ERS mang lại giúp người dùng hiểu rõ lý do tại sao một sản phẩm hoặc dịch vụ được đề xuất, từ đó củng cố niềm tin vào hệ thống. Khi người dùng có thể nắm bắt cách thức hoạt động của hệ thống khuyến nghị, họ sẽ cảm thấy an tâm hơn khi đưa ra quyết định dựa trên các gợi ý này.

Bên cạnh đó, khả năng cá nhân hóa lời giải thích đóng vai trò quan trọng trong việc nâng cao trải nghiệm người dùng. Thay vì chỉ nhận các khuyến nghị một cách thụ động, người dùng có thể điều chỉnh và tinh chỉnh sở thích của mình dựa trên các yếu tố mà hệ thống đưa ra, từ đó nhận được những gợi ý sát với nhu cầu thực tế hơn. Điều này đặc biệt hữu ích trong các nền tảng thương mại điện tử, dịch vụ phát trực tuyến hoặc mạng xã hội, nơi người dùng mong muốn có sự linh hoạt trong việc lựa chọn nội dung.

Ngoài ra, ER cũng góp phần tăng cường tính công bằng và giảm thiểu sai lệch trong hệ thống khuyến nghị. Việc công khai các yếu tố ảnh hưởng đến quyết định giúp phát hiện và hạn chế tình trạng thiên vị trong mô hình, qua đó đảm bảo rằng các gợi ý không bị ảnh hưởng bởi những định kiến không mong muốn. Đặc biệt, trong các ứng dụng nhạy cảm như tuyển dụng, tài chính hay chăm sóc sức khỏe, tính minh bạch này giúp hạn chế các quyết định không công bằng hoặc sai lệch có thể gây hậu quả tiêu cực.

Có hai loại mô hình giải thích là mô hình nội tại và mô hình không phụ thuộc. Mô hình nội tại liên quan đến việc xây dựng các mô hình có khả năng giải thích ngay từ bên trong mô hình, có nghĩa là bản thân mô hình được xây dựng sao cho có thể cung cấp thông tin về lý do ra quyết định mà không cần thêm bất kỳ cơ chế nào bên ngoài.

* Corresponding Author

Khác với mô hình nội tại, phương pháp này không can thiệp vào quá trình ra quyết định của hệ thống mà chỉ cung cấp lời giải thích sau khi khuyến nghị đã được tạo ra [3].

Mặc dù các hệ thống ERS đã mang lại nhiều lợi ích trong việc cải thiện tính minh bạch và hiệu quả của hệ thống khuyến nghị, nhưng vẫn còn một số thách thức lớn cần được giải quyết. Một trong những vấn đề có thể kể đến là đánh giá chất lượng lời giải thích. Hiện nay, chưa có một tiêu chuẩn chung nào để xác định một lời giải thích là tốt hay không, và tiêu chí này có thể thay đổi tùy theo ngữ cảnh sử dụng. Điều này đặt ra yêu cầu cần có những phương pháp đánh giá khách quan, kết hợp giữa đánh giá tự động và phản hồi từ người dùng thực tế. Ngoài ra, việc đa dạng hóa các phương pháp giải thích cũng là một vấn đề quan trọng. Không phải người dùng nào cũng có cùng một cách tiếp cận hoặc mức độ hiểu biết về hệ thống, do đó cần có các phương pháp giải thích linh hoạt, phù hợp với từng đối tượng người dùng khác nhau. Một số người có thể thích lời giải thích trực quan bằng đồ thị hoặc hình ảnh, trong khi những người khác có thể cần các diễn giải bằng văn bản hoặc quy tắc logic rõ ràng. Các vấn đề liên quan đến quyền riêng tư và đạo đức cũng cần được xem xét khi triển khai hệ thống ERS. Việc cung cấp lời giải thích chi tiết về cách hệ thống hoạt động có thể vô tình tiết lộ thông tin cá nhân hoặc chiến lược kinh doanh của tổ chức. Điều này đòi hỏi sự cân bằng giữa tính minh bạch và bảo vệ quyền riêng tư, thông qua các cơ chế kiểm soát quyền truy cập và quản lý dữ liệu một cách hợp lý.

Trong bối cảnh hiện nay, việc phát triển các phương pháp hiệu quả nhằm cung cấp lời giải thích rõ ràng cho cả người dùng và nhà thiết kế hệ thống là một yêu cầu cấp thiết. Các hệ thống khuyến nghị có giải thích không chỉ giúp nâng cao tính minh bạch mà còn đóng vai trò quan trọng trong việc thu thập và xử lý phản hồi từ người dùng, góp phần cải thiện chất lượng và độ tin cậy của hệ thống. Tuy nhiên, như đã đề cập, vẫn tồn tại nhiều thách thức cần được giải quyết để tối ưu hóa hiệu quả của các phương pháp giải thích.

Để đi sâu vào phân tích lĩnh vực này, bài viết đã thực hiện một khảo sát tài liệu toàn diện. Quá trình tìm kiếm và lựa chọn tài liệu được thực hiện một cách có hệ thống, dựa trên các tiêu chí lựa chọn tài liệu chặt chẽ nhằm đảm bảo tính khoa học, độ tin cậy và cập nhật của thông tin.

Đầu tiên, tính chuyên môn và học thuật là ưu tiên hàng đầu. Bài viết tập trung vào các công trình khoa học uy tín, chủ yếu từ năm 2007 đến 2025, đã được duyệt và xuất bản tại các hội nghị hàng đầu trong cộng đồng khoa học máy tính và trí tuệ nhân tạo. Cụ thể, các tài liệu từ kỷ yếu của NeurIPS, KDD, SIGIR, AAAI, ICCV, cùng với các tạp chí chuyên ngành danh tiếng như IEEE Transactions on Knowledge and Data Engineering, User Modeling and User-Adapted Interaction, và ACM Transactions.

Thứ hai, để đảm bảo nội dung luôn cập nhật và nắm bắt được những xu hướng mới nhất, nghiên cứu đã bổ sung các bài báo tiền xuất bản từ arXiv. Điều này giúp bài viết phản ánh kịp thời những thông tin và phát triển tiên phong ngay khi chúng được chia sẻ rộng rãi trong cộng đồng khoa học.

Thứ ba, về phạm vi và nội dung, nghiên cứu đã tập trung vào các khía cạnh cốt lõi của lĩnh vực. Cụ thể các tài liệu được lựa chọn xoay quanh các phương pháp giải thích cho hệ thống khuyến nghị, làm rõ vai trò ngày càng tăng của mô hình ngôn ngữ lớn, phân tích cơ chế attention và các kỹ thuật giải thích mô hình hộp, cùng với các phương pháp kết hợp đa dạng khác.

Với mục tiêu phân tích chuyên sâu các khía cạnh trên, bài viết được tổ chức thành các phần sau:

Phần II: Trình bày các phương pháp phổ biến trong hệ thống khuyến nghị có giải thích;

Phần III: Tổng hợp các công trình nghiên cứu liên quan;

Phần IV: Dataset trong hệ thống khuyến nghị có giải thích;

Phần V: Đề xuất các hướng phát triển trong tương lai.

II. CÁC PHƯƠNG PHÁP GIẢI THÍCH TRONG HỆ THỐNG KHUYẾN NGHỊ

A. MODEL-BASED EXPLAINABILITY (GIẢI THÍCH DỰA TRÊN MÔ HÌNH)

Giải thích dựa trên mô hình (Model-Based Explainability) hướng đến việc thiết kế các mô hình với khả năng tự giải thích từ bên trong, những phương pháp này giúp "mở hộp đen" của mô hình và cung cấp các bằng chứng nội tại cho quyết định [2].

Giải thích dựa trên mô hình mang lại nhiều lợi ích quan trọng trong hệ thống khuyến nghị. Trước hết, phương pháp này đảm bảo tính minh bạch cao, khi lời giải thích được rút trích trực tiếp từ mô hình, giúp người dùng hiểu rõ cách hệ thống đưa ra khuyến nghị, từ đó tăng mức độ tin cậy và sự chấp nhận. Bên cạnh đó, lời giải thích phản ánh chính xác cách mô hình hoạt động, tránh được sai lệch có thể xảy ra. Ngoài ra, việc có được thông tin giải thích ngay từ bên trong hệ thống giúp các nhà phát triển dễ dàng điều chỉnh mô hình, từ đó cải thiện hiệu suất và chất lượng khuyến nghị.

Tuy nhiên, phương pháp này cũng tồn tại một số hạn chế nhất định. Một trong những nhược điểm lớn là hạn chế về khả năng áp dụng, khi các mô hình phức tạp như mạng nơ-ron sâu hay ma trận phân rã khó có thể cung cấp lời

giải thích trực tiếp do bản chất "hộp đen" của chúng. Hơn nữa, để đảm bảo khả năng diễn giải, các mô hình có thể phải được thiết kế đơn giản hơn, điều này có thể làm giảm độ chính xác so với các phương pháp tối ưu hóa hoàn toàn theo hiệu suất. Cuối cùng, phương pháp này đòi hỏi thiết kế mô hình đặc thù ngay từ đầu, khiến việc triển khai trở nên khó khăn hơn.

Có nhiều hướng tiếp cận theo phương pháp này, có thể kể đến như sau:

1. FEATURE IMPORTANCE ANALYSIS (MỨC ĐỘ QUAN TRỌNG CỦA CÁC ĐẶC TRƯNG)

Phương pháp được sử dụng để đo lường và đánh giá mức độ đóng góp của từng đặc trưng (feature) vào kết quả dự đoán của mô hình học máy. Nói cách khác, nó giúp xác định những đặc trưng nào ảnh hưởng mạnh nhất đến quyết định của mô hình.

Mục đích của phương pháp này là phân tích tầm quan trọng của các đặc trưng nhằm hiểu rõ hơn về quá trình ra quyết định của mô hình, từ đó giúp cải thiện tính minh bạch và khả năng giải thích.

Một trong những kỹ thuật tiêu biểu được áp dụng trong phương pháp này là SHAP (Shapley Additive Explanations) [4]. SHAP sử dụng giá trị Shapley, xuất phát từ lý thuyết trò chơi hợp tác, để định lượng một cách công bằng đóng góp của từng đặc trưng vào kết quả dự đoán của mô hình, từ đó cung cấp một lời giải thích có cơ sở lý thuyết vững chắc.

Giá trị Shapley của một đặc trưng được tính bằng cách xác định đóng góp cận biên trung bình của đặc trưng đó đối với dự đoán của mô hình, tính qua tất cả các cách sắp xếp có thể của các đặc trưng. Công thức chung để tính giá trị Shapley cho đặc trưng i trong tập các đặc trưng N như sau:

$$\emptyset_i = \sum_{S \in N\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Ở đây:

- N là tập hợp tất cả các đặc trưng,
- S là các tập con của N không chứa đặc trưng i,
- $f(S)$ là giá trị dự đoán của mô hình khi chỉ sử dụng tập các đặc trưng S,
- $f(S \cup \{i\})$ là giá trị dự đoán khi thêm đặc trưng i vào tập S,
- $|S|!$ và $(|N|-|S|-1)!$ là các hệ số trọng số để đảm bảo rằng mỗi cách sắp xếp các đặc trưng được tính đến một cách công bằng,
- $|N|!$ là giai thừa của tổng số đặc trưng.

Ý nghĩa của công thức là đối với mỗi tập S không chứa đặc trưng i, tính sự thay đổi trong dự đoán khi thêm i vào S, sau đó trung bình các đóng góp này qua tất cả các tập con S, với trọng số tương ứng với số thứ tự có thể của S. Qua đó, giá trị Shapley thể hiện mức độ đóng góp trung bình của đặc trưng i vào dự đoán của mô hình.

Kỹ thuật này hoạt động bằng cách tính toán đóng góp trung bình của mỗi đặc trưng đối với kết quả dự đoán, thông qua việc xem xét tất cả các tổ hợp con của các đặc trưng. Nhờ đó, SHAP cung cấp khả năng phân tích cục bộ, tức là giải thích chi tiết từng dự đoán bằng cách phân chia kết quả thành các phần đóng góp của từng đặc trưng sao cho tổng các giá trị này khớp với dự đoán ban đầu.

Hơn nữa, SHAP thỏa mãn các tính chất lý thuyết quan trọng như tính chính xác cục bộ, tính phụ thuộc đơn giản và tính nhất quán, điều này làm cho lời giải thích trở nên đáng tin cậy và có cơ sở lý thuyết vững chắc. Ưu điểm của SHAP nằm ở khả năng cung cấp một cách công bằng và nhất quán để phân chia đóng góp của từng đặc trưng, khả năng ứng dụng rộng rãi trên nhiều loại mô hình khác nhau.

Tuy nhiên, SHAP cũng đối mặt với một số thách thức, chẳng hạn như độ phức tạp tính toán cao do cần xem xét tất cả các tổ hợp con của đặc trưng, khó khăn trong việc xử lý các đặc trưng có mối tương quan cao, và thậm chí là việc chuyển đổi các giá trị định lượng thành lời giải thích dễ hiểu cho người dùng.

2. ATTENTION-BASED (MÔ HÌNH CHÚ Ý)

Cơ chế "attention" là một kỹ thuật giúp mô hình tập trung vào các phần quan trọng của dữ liệu đầu vào khi thực hiện tác vụ, thay vì xử lý toàn bộ dữ liệu một cách đồng đều. Điều này tương tự như cách con người tập trung vào những thông tin cần thiết khi đọc hoặc quan sát, bỏ qua những chi tiết không quan trọng [5].

Trong ngữ cảnh hệ thống khuyến nghị có giải thích, cách tiếp cận này không chỉ cải thiện hiệu suất dự đoán mà còn cung cấp thông tin minh bạch, cho thấy những phần dữ liệu nào đã được ưu tiên khi đưa ra đề xuất.

Các trọng số attention có thể được trực quan hóa (ví dụ: dưới dạng heatmap) để chỉ ra các yếu tố như lịch sử tương tác, đánh giá của người dùng, hay các đặc điểm sản phẩm nào đóng góp nhiều nhất vào quyết định đưa ra khuyến nghị. Điều này giúp "mở hộp đen" của mô hình học sâu và cung cấp lời giải thích trực quan cho người dùng.

Mặc dù được trực quan hóa, nhưng các trọng số attention không trực tiếp phản ánh lý do ra quyết định theo cách mà phụ thuộc nhiều vào cấu trúc dữ liệu và việc hiệu chỉnh mô hình.

3. LATENT FACTOR INTERPRETATION IN MATRIX FACTORIZATION (DIỄN GIẢI YẾU TỐ TIỀM ẨN TRONG PHÂN TÍCH MA TRẬN)

Matrix Factorization, hay còn gọi là phân rã ma trận, là một kỹ thuật trong học máy được sử dụng rộng rãi trong các hệ thống khuyến nghị, đặc biệt là trong lọc cộng tác. Một ma trận đại diện cho tương tác giữa người dùng và sản phẩm (ví dụ: điểm đánh giá, lượt xem, lượt mua) được phân rã thành hai ma trận yếu hơn, thường là ma trận biểu diễn các đặc trưng tiềm ẩn của người dùng và một ma trận biểu diễn các đặc trưng tiềm ẩn của sản phẩm.

Trong bài báo "Matrix Factorization Techniques for Context-Aware Collaborative Filtering Recommender Systems: A Survey" của tác giả Mohamed Hussein Abdi cùng cộng sự [6] cho thấy rằng các kỹ thuật Matrix Factorization không chỉ đóng vai trò chính trong việc dự đoán khuyến nghị mà còn là công cụ hữu hiệu để "mở hộp đen" của hệ thống thông qua việc giải thích các latent factors – từ đó cung cấp cho người dùng lý do cụ thể và có căn cứ cho các đề xuất được đưa ra.

Phương pháp phân rã ma trận giúp hệ thống học được các latent factors – tức là các đặc trưng tiềm ẩn mô tả sở thích của người dùng và các thuộc tính của sản phẩm. Khi các yếu tố này có thể liên kết với các đặc tính dễ hiểu như thể loại phim, phong cách mua sắm, hoặc các đặc trưng ngữ cảnh, chúng trở thành cơ sở để giải thích rằng "khuyến nghị này được đưa ra vì bạn có sở thích cao về [yếu tố cụ thể]".

Trong bối cảnh hệ thống khuyến nghị có ngữ cảnh, ma trận tương tác được mở rộng (thường dưới dạng tensor) để bao gồm thêm thông tin như thời gian, địa điểm hay trạng thái cảm xúc. Việc phân rã các tensor này giúp trích xuất ra các latent factors gắn liền với ngữ cảnh, từ đó cung cấp lời giải thích rằng "sản phẩm này phù hợp với bạn vào thời điểm/địa điểm hiện tại vì nó đáp ứng nhu cầu theo bối cảnh của bạn".

Bài báo cũng nêu ra các biến thể của phân rã ma trận (như non-negative matrix factorization hay các ràng buộc bổ sung) nhằm mục tiêu làm cho các latent factors trở nên trực quan và dễ diễn giải hơn. Khi các latent factors được thiết kế để tương quan trực tiếp với các đặc trưng đã biết của dữ liệu, hệ thống có thể giải thích rằng "lời khuyến nghị xuất phát từ những đặc trưng cụ thể mà bạn quan tâm".

Mặc dù việc phân rã ma trận cho phép truy xuất các latent factors hữu ích, nhưng bài báo cũng chỉ ra rằng việc giải thích các yếu tố này vẫn gặp khó khăn do chúng không luôn dễ ánh xạ với các thuộc tính trực quan. Điều này đặt ra thách thức cho các nghiên cứu nhằm cải tiến khả năng diễn giải của các latent factors thông qua các kỹ thuật ràng buộc hoặc kết hợp với dữ liệu ngoại lai.

B. POST-HOC EXPLAINABILITY (GIẢI THÍCH SAU KHI RA KẾT QUẢ)

Phương pháp giải thích sau khi có kết quả (Post-Hoc Explainability) là một cách tiếp cận trong hệ thống khuyến nghị nhằm cung cấp lời giải thích sau khi hệ thống đã đưa ra khuyến nghị cho người dùng. Thay vì thiết kế mô hình với khả năng tự giải thích ngay từ đầu, phương pháp này phân tích và diễn giải kết quả từ các mô hình đã có, đặc biệt là các mô hình phức tạp như deep learning hay matrix factorization [7].

Phương pháp Post-Hoc Explainability mang lại nhiều lợi ích đáng kể trong hệ thống khuyến nghị. Một trong những ưu điểm quan trọng là khả năng áp dụng cho nhiều mô hình phức tạp mà không cần thay đổi cấu trúc bên trong, giúp duy trì tính linh hoạt và tiết kiệm chi phí triển khai. Hơn nữa, phương pháp này có thể dễ dàng tích hợp vào các hệ thống khuyến nghị đã có sẵn, cho phép cung cấp lời giải thích mà không ảnh hưởng đến hiệu suất của mô hình gốc. Ngoài ra, nó còn tạo ra các lời giải thích linh hoạt, có thể tùy chỉnh theo nhu cầu của người dùng, giúp cải thiện trải nghiệm và mức độ tin cậy đối với hệ thống.

Tuy nhiên, phương pháp này cũng tồn tại một số hạn chế nhất định. Do hoạt động như một lớp diễn giải bên ngoài, Post-Hoc Explainability không phản ánh hoàn toàn chính xác cách mô hình gốc thực sự đưa ra quyết định, đặc biệt khi sử dụng các mô hình thay thế để giải thích. Một số kỹ thuật như phân tích dữ liệu phản thực có thể trở nên khó hiểu đối với người dùng không có chuyên môn, gây cản trở trong việc tiếp nhận thông tin. Hơn nữa, nếu lời giải thích không thực sự khớp với logic của hệ thống khuyến nghị, có thể dẫn đến sai lệch, làm giảm tính chính xác và hiệu quả của quá trình giải thích.

Có nhiều hướng tiếp cận theo phương pháp này, có thể kể đến như sau:

1. RULE-BASED EXPLANATIONS (MÔ HÌNH GIẢI THÍCH DỰA TRÊN QUY TẮC)

Mô hình dựa trên quy tắc là những hệ thống giải thích dựa trên các quy tắc logic rõ ràng, thường được biểu diễn dưới dạng các câu "nếu – thì". Các quy tắc này được xây dựng từ dữ liệu hoặc được xác định bởi chuyên gia, giúp xác định các điều kiện cụ thể dẫn đến một dự đoán hoặc đề xuất trong hệ thống khuyến nghị [2].

Ưu điểm của mô hình rule-based là tính giải thích cao, các quy tắc được xây dựng theo cách mà con người có thể hiểu được, do đó giúp mở "hộp đen" của các hệ thống khuyến nghị phức tạp. Tuy nhiên, cũng có hạn chế là khả năng mở rộng kém khi dữ liệu trở nên quá phức tạp và tương tác giữa các đặc trưng không được thể hiện đầy đủ

qua các quy tắc đơn giản. Điều này có thể dẫn đến việc tạo ra số lượng lớn các quy tắc, làm giảm tính tổng quát của mô hình.

Bài báo "A Survey of Explanations in Recommender Systems" của đồng tác giả Nava Tintarev và Judith Masthoff [8], đã nêu lên giải thích dựa trên quy tắc được xem là một trong những hướng tiếp cận chủ đạo nhằm làm rõ cơ sở logic đằng sau các khuyến nghị. Các quy tắc có thể được xây dựng dựa trên dữ liệu lịch sử, khai thác các mẫu mối liên hệ hoặc dựa trên kiến thức chuyên môn của từng vùng miền. Công trình nghiên cứu nhấn mạnh rằng giải thích dựa trên quy tắc mang lại tính minh bạch và dễ hiểu cao vì chúng phản ánh trực tiếp các mối quan hệ logic đã được định nghĩa sẵn. Tuy nhiên, họ cũng lưu ý rằng khi hệ thống trở nên phức tạp, các quy tắc đơn giản có thể không đủ diễn tả đầy đủ các mối liên hệ giữa các đặc trưng, từ đó có thể cần kết hợp với các phương pháp giải thích khác được cung cấp cái nhìn toàn diện hơn.

2. COUNTERFACTUAL EXPLANATIONS (MÔ HÌNH GIẢI THÍCH PHẢN THỰC)

Counterfactual Explanations là một cơ chế giải thích trong Explainable AI, tập trung vào việc chỉ ra những thay đổi cần thiết đối với đầu vào để thay đổi dự đoán của mô hình. Nói cách khác, nó trả lời câu hỏi: "Nếu đầu vào khác đi như thế nào thì kết quả dự đoán sẽ khác?" [9].

Counterfactual explanations đưa ra các kịch bản "nếu – thì", ví dụ: "Nếu giá trị của đặc trưng X tăng lên 10% hoặc đặc trưng Y giảm đi, thì kết quả dự đoán sẽ thay đổi từ A sang B". Điều này giúp người dùng hiểu được yếu tố nào đóng vai trò then chốt trong quyết định của mô hình.

Phương pháp này thường được sử dụng để làm sáng tỏ ranh giới giữa các lớp dự đoán trong các bài toán phân loại và giúp người dùng nhận diện các đặc trưng có ảnh hưởng mạnh mẽ. Nó cũng giúp đưa ra các khuyến nghị cải thiện, chẳng hạn như "để đạt được kết quả mong muốn, bạn cần thay đổi đặc trưng này như thế nào".

Counterfactual explanations mang lại những lợi ích quan trọng khi cung cấp một cách tiếp cận trực quan, giúp người dùng dễ dàng nhận thấy rằng sự thay đổi cụ thể của một hoặc vài đặc trưng đầu vào có thể dẫn đến sự thay đổi đáng kể trong dự đoán của mô hình. Nhờ đó, nó giúp làm sáng tỏ các yếu tố then chốt quyết định kết quả, hỗ trợ việc hiểu và cải tiến hệ thống. Tuy nhiên, việc tạo ra các kịch bản phản thực khả thi và có ý nghĩa thực tiễn vẫn gặp nhiều khó khăn, đặc biệt là khi các đặc trưng có mối tương quan chặt chẽ, khiến việc điều chỉnh một yếu tố có thể gây ảnh hưởng đến các yếu tố khác. Hơn nữa, đảm bảo rằng các lời giải thích này thực sự phản ánh chính xác cơ chế ra quyết định của mô hình là một thách thức lớn, đòi hỏi sự nghiên cứu và cải tiến liên tục.

3. SURROGATE MODELS (MÔ HÌNH THAY THẾ)

Surrogate Models, hay còn gọi là mô hình thay thế [9], là các mô hình đơn giản được xây dựng để xấp xỉ hành vi của một mô hình phức tạp (thường là "hộp đen") trong một vùng lân cận của đầu vào cụ thể. Ý tưởng cơ bản là, thay vì cố gắng hiểu toàn bộ cấu trúc và cơ chế hoạt động của một mô hình phức tạp, chúng ta có thể huấn luyện một mô hình đơn giản, ví dụ như mô hình tuyến tính hay cây quyết định, để mô phỏng hành vi của mô hình gốc xung quanh một điểm dữ liệu cụ thể. Qua đó, các hệ số của mô hình thay thế có thể được sử dụng để giải thích tác động của các đặc trưng đối với dự đoán của mô hình phức tạp.

Có hai loại mô hình thay thế chính:

- Mô hình thay thế toàn cục (Global Surrogate Models): Cố gắng xấp xỉ toàn bộ mô hình gốc trên không gian đầu vào.
- Mô hình thay thế cục bộ (Local Surrogate Models): Chỉ tập trung vào việc mô phỏng hành vi của mô hình gốc trong một vùng nhỏ quanh một điểm dữ liệu cụ thể.

C. USER-CENTRIC EXPLAINABILITY (GIẢI THÍCH THEO HƯỚNG NGƯỜI DÙNG)

Giải thích theo hướng người dùng tập trung vào nhu cầu và sự hiểu biết của người dùng. Thay vì chỉ cung cấp một lời giải thích cố định, phương pháp này điều chỉnh lời giải thích dựa trên đặc điểm cá nhân của từng người dùng, giúp họ dễ dàng hiểu và tin tưởng hơn vào hệ thống khuyến nghị [8].

Nhờ vào khả năng cá nhân hóa lời giải thích theo nhu cầu của từng cá nhân, phương pháp này giúp tăng cường sự hiểu biết và tin cậy của người dùng đối với hệ thống. Đồng thời, cải thiện trải nghiệm người dùng khi cho phép họ tự điều chỉnh các yếu tố ảnh hưởng đến khuyến nghị, tạo cảm giác kiểm soát tốt hơn. Đặc biệt, trong các lĩnh vực như tài chính, y tế và giáo dục, phương pháp này đóng vai trò quan trọng trong việc hỗ trợ người dùng đưa ra quyết định chính xác và phù hợp hơn.

Tuy nhiên, phương pháp này cũng tồn tại một số hạn chế đáng kể. Để cá nhân hóa lời giải thích, hệ thống cần thu thập nhiều thông tin về hành vi và sở thích cá nhân, gây ra những lo ngại về quyền riêng tư. Bên cạnh đó, việc tạo ra lời giải thích theo từng người dùng đòi hỏi chi phí tính toán cao hơn so với các phương pháp giải thích cố định. Hơn nữa, do mỗi người có cách hiểu khác nhau, việc đánh giá chất lượng lời giải thích trở nên khó khăn, khiến hệ thống gặp thách thức trong việc tối ưu hóa trải nghiệm cho mọi đối tượng.

Có nhiều cách tiếp cận của phương pháp này, có thể kể đến như sau:

1. VISUALIZATION-BASED EXPLANATION (GIẢI THÍCH DỰA TRÊN MÔ HÌNH TRỰC QUAN)

Giải thích trực quan đóng vai trò quan trọng trong việc tăng cường tính minh bạch và khả năng hiểu của người dùng đối với hệ thống khuyến nghị. Theo Mohamed Amine Chatti cùng cộng sự [10], các phương pháp giải thích trực quan có thể được phân loại theo bốn chiều chính: mục tiêu (aim), phản ánh mục đích của giải thích như nâng cao sự tin cậy, hiệu quả, hay thuyết phục người dùng; phạm vi (scope), bao gồm giải thích đầu vào (input), quy trình xử lý (process) hoặc đầu ra (output); phương pháp (method), phân theo kỹ thuật khuyến nghị được sử dụng như content-based, collaborative filtering, hoặc knowledge-based; và định dạng (format), tập trung vào cách thức trình bày giải thích như biểu đồ (bar chart), bản đồ nhiệt (heatmap), đám mây từ (tag cloud) hoặc giao diện tương tác (interactive visualization). Tác giả cũng nhấn mạnh rằng các thiết kế giải thích trực quan cần đảm bảo tính dễ hiểu, tính nhất quán và phù hợp với ngữ cảnh của người dùng, nhằm nâng cao mức độ tin cậy và trải nghiệm tổng thể khi sử dụng hệ thống khuyến nghị.

2. PERSONALIZED EXPLANATION (GIẢI THÍCH CÁ NHÂN HÓA)

Personalized Explanation trong hệ thống khuyến nghị có giải thích đề cập đến việc tạo ra các lời giải thích được cá nhân hóa dựa trên đặc điểm, sở thích và hành vi của từng người dùng. Thay vì cung cấp một lời giải thích chung chung cho tất cả mọi người, phương pháp này tận dụng dữ liệu cá nhân (ví dụ: lịch sử tương tác, sở thích, thông tin nhân khẩu học, và ngữ cảnh sử dụng) để xây dựng lời giải thích cụ thể, có ý nghĩa và dễ hiểu cho từng cá nhân [11].

Các cơ chế của personalized explanation thường bao gồm:

- Tùy chỉnh nội dung giải thích: Chọn lọc các đặc trưng quan trọng dựa trên hồ sơ người dùng để giải thích vì sao một đề xuất lại được đưa ra.
- Học prompt cá nhân hóa: Trong các hệ thống sử dụng các mô hình ngôn ngữ lớn (LLMs), các prompt đầu vào được điều chỉnh dựa trên thông tin người dùng để kích hoạt các lời giải thích phù hợp với ngữ cảnh cá nhân.
- Kết hợp các kỹ thuật giải thích: Các phương pháp như attention-based hay latent factor analysis có thể được sử dụng để xác định những yếu tố quan trọng nhất đối với từng người dùng và tích hợp chúng vào lời giải thích.

3. INTERACTIVE EXPLANATIONS (GIẢI THÍCH TƯƠNG TÁC)

Interactive Explanations là phương pháp giải thích được thiết kế theo hướng tương tác, cho phép người dùng không chỉ nhận được lời giải thích tĩnh mà còn có thể tương tác để khám phá thêm các chi tiết hoặc điều chỉnh thông tin giải thích theo nhu cầu cá nhân [9].

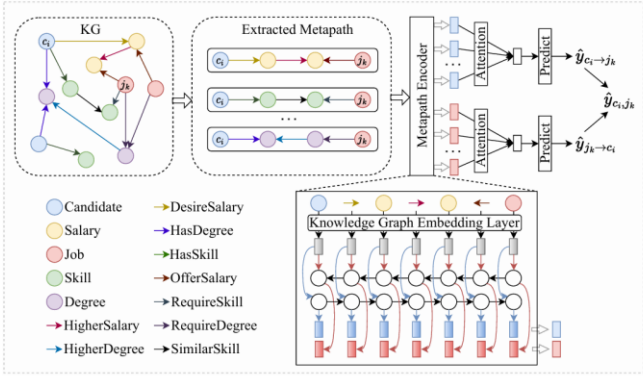
Thay vì chỉ nhận một lời giải thích duy nhất, hệ thống sẽ cung cấp các công cụ tương tác như giao diện trực quan, biểu đồ có thể nhấp chuột, hoặc hệ thống đối thoại, giúp người dùng có thể đặt câu hỏi, trao đổi và điều chỉnh các thành phần của lời giải thích. Nhờ đó, người dùng có thể khám phá chi tiết từng yếu tố, hiểu rõ hơn về cơ sở ra quyết định của mô hình, từ đó nắm bắt mối liên hệ giữa các đặc trưng và kết quả dự đoán. Mục tiêu của phương pháp này là tăng cường tính minh bạch, xây dựng niềm tin và tạo điều kiện cho việc cá nhân hóa lời giải thích theo từng ngữ cảnh và nhu cầu cụ thể của người dùng.

III. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

A. MODEL-BASED EXPLAINABILITY (GIẢI THÍCH DỰA TRÊN MÔ HÌNH)

Trong phương pháp giải thích này, có thể nói đến bài báo "Knowledge-Aware Explainable Reciprocal Recommendation" của tác giả Kai-Huang La cùng cộng sự [12], công trình nghiên cứu tập trung vào việc phát triển hệ thống khuyến nghị đối ứng (Reciprocal Recommender Systems - RRS), được ứng dụng rộng rãi trên các nền tảng trực tuyến như hẹn hò và tuyển dụng. RRS nhằm đáp ứng đồng thời nhu cầu của cả hai bên trong quá trình khuyến nghị. Tuy nhiên, do tính chất đặc thù của nhiệm vụ này, dữ liệu tương tác thường khan hiếm hơn so với các nhiệm vụ khuyến nghị khác. Các phương pháp hiện tại chủ yếu dựa trên khuyến nghị theo nội dung, nhưng thường mô hình hóa thông tin văn bản từ một góc nhìn thống nhất, dẫn đến khó khăn trong việc nắm bắt ý định riêng biệt của mỗi bên, hạn chế hiệu suất và thiếu tính giải thích.

Để khắc phục những hạn chế này, bài báo đề xuất một hệ thống khuyến nghị đối ứng có khả năng giải thích và nhận thức kiến thức (Knowledge-Aware Explainable Reciprocal Recommender System - KAERR) (Hình 1). Hệ thống này mô hình hóa các đường dẫn siêu dữ liệu (metapath) giữa hai bên một cách độc lập, xem xét quan điểm và yêu cầu riêng của từng bên. Các metapath khác nhau được kết hợp thông qua cơ chế dựa trên sự chú ý, trong đó trọng số chú ý tiết lộ sở thích từ hai góc nhìn và cung cấp giải thích cho các khuyến nghị đối với cả hai bên. Các thí nghiệm trên hai tập dữ liệu thực tế từ các bối cảnh khác nhau cho thấy mô hình đề xuất vượt trội so với các phương pháp tiên tiến hiện tại, đồng thời cung cấp lý do thuyết phục cho các khuyến nghị đối với cả hai bên.



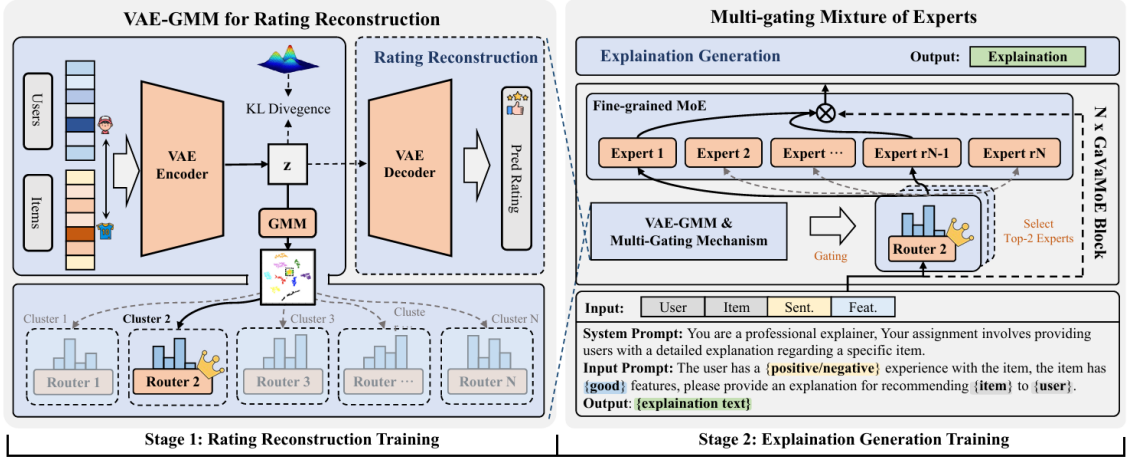
Hình 1. *Mô hình hệ thống KAERR*

Tuy nhiên, bài báo cũng chỉ ra một số thách thức cần được giải quyết. Một trong những thách thức chính là việc phát triển một phương pháp có thể vượt qua sự khan hiếm của các tín hiệu song phương và phù hợp với bối cảnh tương hỗ. Điều này đòi hỏi phải mô hình hóa thông tin văn bản từ hai phía một cách độc lập, thay vì từ một góc nhìn thống nhất, để nắm bắt được ý định riêng biệt của mỗi bên.

Những tiến bộ vượt bậc trong các mô hình LLMs đã mở ra những hướng tiếp cận mới nhằm tạo ra các giải thích giống con người trong hệ thống khuyến nghị, từ đó thúc đẩy sự phát triển của các hệ thống khuyến nghị có giải thích dựa trên LLMs. Một trong những công trình tiêu biểu là bài báo "Language Models are Few-Shot Learners" của Tom B. Brown cùng cộng sự [13], giới thiệu GPT-3 – mô hình ngôn ngữ lớn với 175 tỷ tham số, được xây dựng dựa trên kiến trúc Transformer. Công trình này không chỉ đánh dấu bước tiến quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên mà còn mở ra nhiều tiềm năng ứng dụng LLMs trong các hệ thống AI giải thích được. Tuy nhiên, việc ứng dụng LLMs vẫn tồn tại nhiều thách thức, đặc biệt liên quan đến hiệu năng, đạo đức và khả năng giải thích. Nhằm khắc phục các hạn chế này, nhiều nghiên cứu đã được thực hiện để nâng cao độ tin cậy và khả năng ứng dụng của các mô hình LLMs trong thực tiễn, tiêu biểu có thể kể đến như sau:

Bài báo “GaVaMoE: Gaussian-Variational Gated Mixture of Experts for Explainable Recommendation” của tác giả Fei Tang cùng cộng sự [14]. Nghiên cứu đã chỉ ra các hệ thống khuyến nghị dựa trên mô hình LLMs hiện tại gặp khó khăn trong các việc là mô hình hóa sở thích phức tạp giữa người dùng và sản phẩm, cá nhân hóa các giải thích cho từng người dùng cụ thể, xử lý tình huống khi dữ liệu tương tác giữa người dùng và sản phẩm bị thiếu hụt. Công trình nghiên cứu đã đề xuất khung làm việc GaVaMoE (Hình 2)+ để giải quyết các vấn đề trên với hai thành phần chính là:

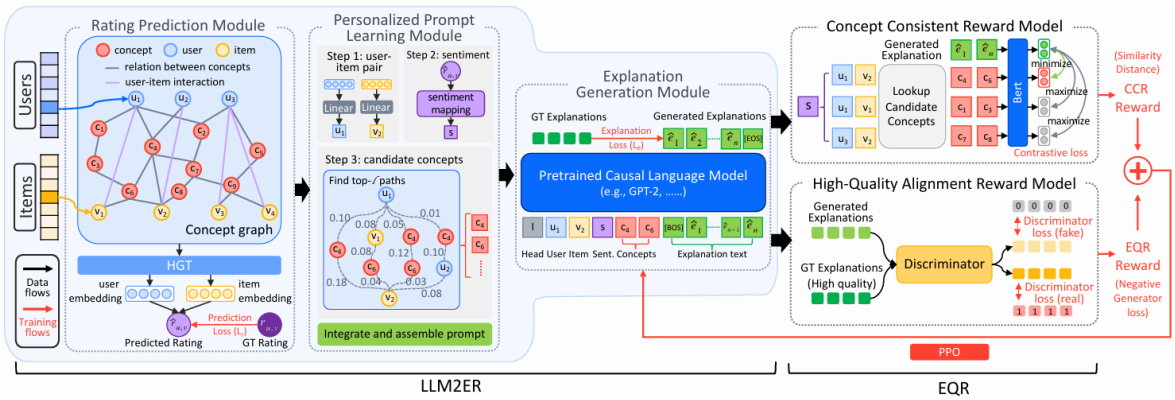
- Mô-đun tái tạo đánh giá: Sử dụng bộ tự mã hóa biến phân (VAE) kết hợp với Mô hình hỗn hợp Gaussian (GMM) để nắm bắt các sở thích phức tạp giữa người dùng và sản phẩm. VAE giúp mô hình hóa các yếu tố tiềm ẩn trong tương tác người dùng-sản phẩm, trong khi GMM phân cụm người dùng có hành vi tương tự, tạo thành cơ chế đa cổng được tiền huấn luyện.
- Tập hợp các mô hình chuyên gia chi tiết: Kết hợp với cơ chế đa cổng, các mô hình này tạo ra các giải thích được cá nhân hóa cao. Mỗi cụm người dùng tương ứng với một cổng trong cơ chế đa cổng, định tuyến cặp người dùng-sản phẩm đến mô hình chuyên gia phù hợp, giúp giảm thiểu vấn đề thiếu dữ liệu bằng cách tận dụng sự tương đồng giữa người dùng.



Hình 2. *Mô hình GaVaMoE*

Mặc dù mô hình GaVaMoE đã cải thiện đáng kể chất lượng giải thích và tính cá nhân hóa trong hệ thống khuyến nghị, nhưng vẫn còn một số thách thức cần được giải quyết. Trước hết, việc tích hợp hai mô hình VAE và GMM cùng với cơ chế đa cổng làm tăng độ phức tạp tính toán, đòi hỏi tài nguyên lớn hơn và có thể ảnh hưởng đến khả năng mở rộng của hệ thống. Bên cạnh đó, dù mô hình đã cho kết quả khả quan trên ba tập dữ liệu thực tế, nhưng khả năng tổng quát hóa vẫn là một vấn đề khi áp dụng vào các lĩnh vực hoặc tập dữ liệu có đặc điểm khác biệt đáng kể so với dữ liệu huấn luyện. Cuối cùng, mặc dù các giải thích do mô hình tạo ra mang tính cá nhân hóa cao, nhưng việc đảm bảo rằng chúng thực sự dễ hiểu và hữu ích đối với người dùng cuối vẫn là một thách thức quan trọng cần tiếp tục nghiên cứu.

Bài báo "Fine-Tuning Large Language Model Based Explainable Recommendation with Explainable Quality Reward" của tác giả Mengyuan Yang cùng cộng sự [15]. Công trình nghiên cứu đã đề xuất một mô hình mới, gọi là LLM2ER-EQR, nhằm cải thiện chất lượng giải thích trong hệ thống khuyến nghị dựa trên các mô hình LLMs. Các hệ thống gợi ý truyền thống thường gặp phải ba vấn đề chính: thiếu tính cá nhân hóa, thiếu nhất quán và dữ liệu giải thích có chất lượng kém. Để khắc phục, nhóm tác giả đã phát triển hai mô hình và áp dụng phương pháp học tăng cường để tinh chỉnh mô hình LLM2ER ban đầu, tạo ra LLM2ER-EQR (Hình 3).

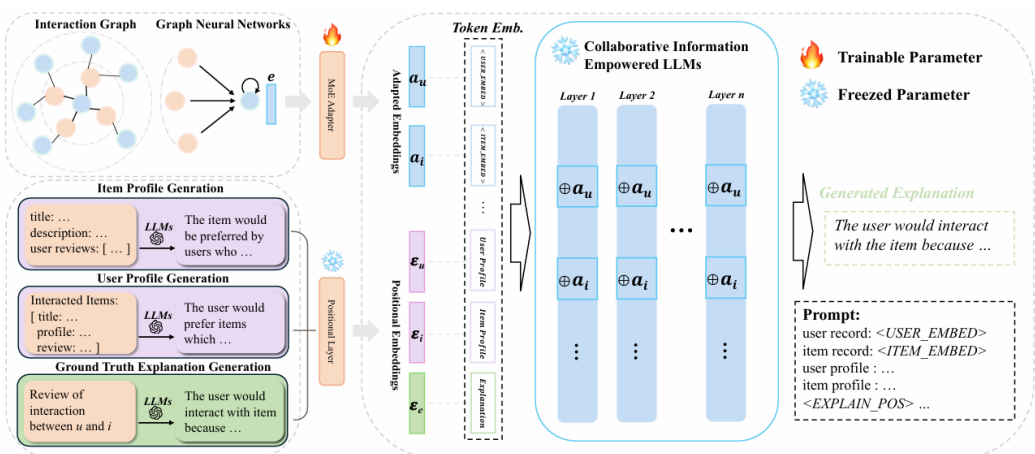


Hình 3. *Mô hình LLM2ER-EQR*

Công trình nghiên cứu đã đạt được những cải tiến đáng kể trong việc nâng cao chất lượng giải thích trong hệ thống khuyến nghị dựa trên mô hình LLMs. Cụ thể, những đóng góp chính bao gồm: giúp tạo ra các giải thích cá nhân hóa, thông tin và nhất quán, ngay cả khi dữ liệu thử nghiệm có chất lượng không cao, tinh chỉnh mô hình LLM, cải thiện độ chính xác và chất lượng của các giải thích được tạo.

Tuy nhiên, nghiên cứu cũng chỉ ra một số hạn chế và đề xuất hướng phát triển trong tương lai. Trước hết, chất lượng dữ liệu đầu vào vẫn là một yếu tố quan trọng, dù mô hình có khả năng xử lý dữ liệu chất lượng thấp, nhưng việc cải thiện dữ liệu sẽ giúp nâng cao hiệu suất của hệ thống. Bên cạnh đó, cần có thêm các nghiên cứu và thử nghiệm trong môi trường thực tế để đánh giá hiệu quả và khả năng triển khai của mô hình trong các hệ thống khuyến nghị đa dạng. Cuối cùng, việc tinh chỉnh các mô hình LLMs đòi hỏi tài nguyên tính toán đáng kể, do đó, nghiên cứu các phương pháp tối ưu hóa và giảm thiểu chi phí tính toán là cần thiết để đảm bảo khả năng mở rộng và ứng dụng mô hình vào thực tiễn.

B. POST-HOC EXPLAINABILITY (GIẢI THÍCH SAU KHI RA KẾT QUẢ)



Hình 4. *Mô hình XRec*

Bài báo “XRec: Large Language Models for Explainable Recommendation” của tác giả Qiyao Ma cùng cộng sự [16], công trình nghiên cứu giới thiệu một mô hình mới, gọi là XRec (Hình 4), nhằm cung cấp các giải thích chi tiết cho hành vi của người dùng trong hệ thống khuyến nghị. Mô hình được xây dựng bằng cách kết hợp giữa lọc cộng tác dựa trên đồ thị và mô hình LLMs để tạo ra các giải thích toàn diện và dễ hiểu.

Cơ chế hoạt động của XRec:

- Bộ mã hóa quan hệ cộng tác (Collaborative Relation Tokenizer): XRec sử dụng một bộ mã hóa để chuyển đổi các quan hệ giữa người dùng và mục tiêu thành các token, giúp LLMs hiểu và xử lý các quan hệ này một cách hiệu quả.
- Bộ điều hợp kết hợp (Mixture of Experts Adapter): Để kết nối tín hiệu cộng tác với ngữ nghĩa ngôn ngữ, XRec áp dụng một bộ điều hợp kết hợp, cho phép LLMs tạo ra các giải thích dựa trên cả quan hệ người dùng-mục tiêu và ngữ cảnh ngôn ngữ.

Công trình nghiên cứu đã mang lại những đóng góp đáng kể trong lĩnh vực hệ thống khuyến nghị có giải thích, đặc biệt với việc giới thiệu mô hình XRec. Đây là một mô hình mới, tích hợp phương pháp lọc cộng tác với các kỹ thuật diễn giải, kết hợp giữa bộ lọc cộng tác và phân tích ngữ nghĩa để cung cấp các khuyến nghị chính xác hơn. Đặc biệt, XRec đã thành công trong việc kết hợp các tín hiệu tương tác của người dùng và mục tiêu, giúp hệ thống hiểu rõ hơn về sở thích cá nhân và đưa ra các gợi ý có ý nghĩa. Một trong những điểm nổi bật của mô hình này là khả năng tạo ra các giải thích chi tiết và toàn diện, giúp người dùng hiểu rõ hơn lý do đằng sau từng khuyến nghị, từ đó tăng cường mức độ tin cậy vào hệ thống.

Tuy nhiên, nghiên cứu cũng chỉ ra một số thách thức quan trọng cần được giải quyết để nâng cao hiệu quả của mô hình. Một trong những thách thức chính là việc tích hợp các tín hiệu cộng tác và ngữ nghĩa từ dữ liệu người dùng, giúp hệ thống không chỉ hiểu rõ hơn về hành vi tương tác mà còn tránh được sự lệch lạc trong quá trình tạo khuyến nghị. Ngoài ra, để XRec có thể được áp dụng rộng rãi, cần đảm bảo tính tương thích với nhiều loại mô hình khuyến nghị khác nhau, đồng thời duy trì được hiệu suất hoạt động ổn định trong các môi trường ứng dụng thực tế. Bên cạnh đó, vấn đề đảm bảo rằng các lời giải thích do hệ thống cung cấp phải đủ rõ ràng và có ý nghĩa đối với người dùng cũng là một thách thức quan trọng cần tiếp tục nghiên cứu và cải thiện.

Bài báo: “Attention Is Not the Only Choice: Counterfactual Reasoning for Path-Based Explainable Recommendation” của tác giả Yicong Li cùng cộng sự [17], nhóm tác giả đã chỉ ra trong các mô hình khuyến nghị dựa trên đồ thị, cơ chế "attention" thường được sử dụng để xác định các đường dẫn quan trọng. Tuy nhiên, các trọng số "attention" này thường được thiết kế để tối ưu hóa độ chính xác của mô hình hơn là khả năng giải thích, dẫn đến sự không ổn định và thiếu trực quan trong các giải thích. Để khắc phục hạn chế trên, tác giả đề xuất sử dụng lập luận phản thực từ lý thuyết học nhân quả. Cụ thể, họ phát triển hai thuật toán lập luận phản thực từ góc độ biểu diễn đường dẫn và cấu trúc hình học của đường dẫn, nhằm học các trọng số giải thích thay thế cho trọng số "attention". Bài báo cũng đề xuất một bộ giải pháp đánh giá tính giải thích, bao gồm cả phương pháp định tính và định lượng. Phương pháp suy luận phản thực trong hệ thống khuyến nghị mang lại nhiều lợi ích đáng kể, đặc biệt trong việc tăng cường khả năng giải thích. Bằng cách xác định rõ vai trò của các đường dẫn trong quá trình khuyến nghị, phương pháp này giúp cung cấp lời giải thích minh bạch và dễ hiểu cho người dùng. Bên cạnh đó, nhóm tác giả cũng đề xuất một bộ công cụ đánh giá toàn diện, bao gồm cả phương pháp định tính và định lượng, giúp đo lường hiệu quả của hệ thống một cách chính xác. Tuy nhiên, phương pháp này vẫn tồn tại một số thách thức. Đầu tiên, việc áp dụng suy luận phản thực đòi hỏi tài nguyên tính toán đáng kể, đặc biệt khi xử lý các mạng lưới lớn và phức tạp. Thêm vào đó, khả năng tổng quát hóa của phương pháp cũng là một vấn đề cần quan tâm, mặc dù đã cho kết quả khả quan trên ba tập dữ liệu thực tế, nhưng khi áp dụng vào các bối cảnh khác nhau, đặc biệt với các tập dữ liệu có đặc điểm khác biệt, hiệu quả của phương pháp có thể bị ảnh hưởng.

Hiện tượng hallucination hay còn gọi là “Ảo giác” là khi AI tạo ra những nội dung có độ thuyết phục cao nhưng lại sai lệch với thực tế hoặc hoàn toàn vô nghĩa. Hiện tượng này có nguyên nhân do AI hiện tại rất giỏi về “mộng mơ” nhưng lại thiếu đi khả năng suy luận, giống như một người ngủ say thì các cơ chế tư duy, suy luận thông thường của não bộ bị tắt bớt [18]. Trong các hệ thống khuyến nghị, việc tạo ra các đánh giá sản phẩm chứa một lượng thông tin phong phú về sở thích người dùng, không chỉ giúp tăng cường tính minh bạch của hệ thống mà còn hỗ trợ người dùng đưa ra quyết định sáng suốt. Tuy nhiên, một thách thức lớn của phương pháp này chính là hiện tượng “ảo giác”, khi các mô hình sinh văn bản tạo ra thông tin không thể xác minh hoặc không được suy luận từ các đánh giá thực tế của người dùng. Điều này làm suy giảm tính đáng tin cậy của hệ thống khuyến nghị, do các đề xuất không còn dựa trên cơ sở hợp lý vững chắc. Vì vậy, việc đảm bảo sự trung thực và tính thực tế trong các bài đánh giá được tạo ra là yếu tố quan trọng để duy trì độ tin cậy và giá trị của hệ thống khuyến nghị.

Để giải quyết vấn đề này, công trình nghiên cứu trong bài báo "Counterfactual Path Augmentation for Reinforcement Reasoning in Explainable Recommendation" của tác giả HaoJie Trang cùng cộng sự [19] đã đề xuất một mô hình khuyến nghị có khả năng giải thích, sử dụng phương pháp tăng cường đường dẫn phản thực để cải thiện quá trình suy luận trong hệ thống khuyến nghị. Cụ thể, nghiên cứu tập trung vào việc học sở thích người

dùng thông qua việc phân tích các đường dẫn trong đồ thị tri thức và áp dụng học tăng cường để tạo ra các khuyến nghị cùng với giải thích tương ứng. Cụ thể, giải quyết những vấn đề sau:

- Nâng cao khả năng giải thích của hệ thống khuyến nghị: Thay vì chỉ đưa ra các đề xuất dưới dạng kết quả "hộp đen", mô hình được đề xuất cho phép hệ thống giải thích chi tiết lý do tại sao một mục nào đó được khuyến nghị (hoặc không được khuyến nghị) cho người dùng.
- Tăng cường đường dẫn phản thực (Counterfactual Path Augmentation): Bằng cách phân tích các đường dẫn (path) trong đồ thị tri thức liên quan đến người dùng và sản phẩm, mô hình sử dụng các biến thể phản thực để xác định các yếu tố quyết định trong quá trình đưa ra khuyến nghị. Điều này giúp xác định được những đường dẫn quan trọng nhất góp phần vào quyết định của mô hình.
- Học tăng cường với cơ chế thưởng kép: gồm phần thưởng liên quan đến đường dẫn (để đánh giá mức độ quan trọng của các đường dẫn dẫn đến khuyến nghị), và phần thưởng liên quan đến mục tiêu (để đảm bảo rằng sản phẩm được khuyến nghị phù hợp với sở thích và nhu cầu của người dùng)
- Đánh giá chất lượng giải thích: bài báo cũng đề xuất các chỉ số mới như tính ổn định (stability) và hiệu quả (effectiveness) để đo lường mức độ trung thực và hữu ích của các giải thích được tạo ra bởi hệ thống.

Công trình nghiên cứu này đã tạo bước đột phá trong lĩnh vực hệ thống khuyến nghị, không chỉ đạt được độ chính xác cao trong dự đoán mà còn cung cấp các giải thích minh bạch và cụ thể về quá trình ra quyết định. Nhờ đó, người dùng có thể hiểu rõ những yếu tố cốt lõi dẫn đến mỗi đề xuất, từ đó tăng cường niềm tin và khuyến khích sự tương tác hiệu quả với hệ thống. Tuy nhiên việc xác định và chọn lọc các đường dẫn từ đồ thị tri thức sao cho chúng phản ánh chính xác các yếu tố quyết định trong quá trình đưa ra khuyến nghị là một thách thức lớn.

Bài báo "Why Should I Trust You? Explaining the Predictions of Any Classifier" của tác giả Marco Tulio Ribeiro cùng cộng sự [20] tập trung vào vấn đề thiếu tính minh bạch của các mô hình dự đoán phức tạp – những "hộp đen" mà người dùng và nhà phát triển thường khó hiểu được cơ sở ra quyết định của chúng. Để giải quyết vấn đề này, tác giả đã giới thiệu phương pháp LIME (Local Interpretable Model-agnostic Explanations), một cách tiếp cận độc lập với mô hình, nhằm cung cấp lời giải thích cục bộ cho từng dự đoán cụ thể.

Bằng cách tạo ra các biến thể của đầu vào thông qua nhiễu loạn (perturbation) và huấn luyện một mô hình giải thích đơn giản (thường là mô hình tuyến tính) trên các mẫu dữ liệu xung quanh điểm cần giải thích, LIME giúp xác định được mức độ đóng góp của từng đặc trưng vào dự đoán. Qua đó, nó cung cấp những lời giải thích dễ hiểu, giúp người dùng nắm bắt được lý do đằng sau mỗi quyết định của mô hình.

Phương pháp LIME dựa trên việc xấp xỉ cục bộ mô hình gốc. Giả sử có một mô hình phức tạp $f(x)$ (có thể là một mạng nơ-ron sâu, cây quyết định phức tạp, hoặc mô hình boost), cần giải thích dự đoán tại một điểm dữ liệu cụ thể x.

LIME tạo một mô hình thay thế đơn giản $g(x)$ sao cho nó xấp xỉ tốt nhất mô hình gốc $f(x)$ trong một vùng nhỏ xung quanh x.

LIME tìm một mô hình thay thế g bằng cách giải bài toán tối ưu sau:

$$\arg\max_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Trong đó:

- G là tập hợp các mô hình thay thế đơn giản (ví dụ: mô hình hồi quy tuyến tính hoặc cây quyết định).
- $L(f,g,\pi_x)$ là hàm mất mát đo độ khác biệt giữa mô hình thay thế g và mô hình gốc f trên một tập dữ liệu lân cận.
- π_x là trọng số của các điểm dữ liệu gần x, đảm bảo rằng mô hình thay thế tập trung mô phỏng chính xác vùng xung quanh x.
- $\Omega(g)$ là độ phức tạp của mô hình g (ví dụ: số lượng đặc trưng được sử dụng trong hồi quy tuyến tính), nhằm giữ cho mô hình dễ hiểu.

Thành tựu của công trình bao gồm việc mở ra một hướng tiếp cận mới cho việc giải thích các mô hình phức tạp, cho thấy rằng việc "mở hộp đen" có thể được thực hiện một cách hiệu quả thông qua các mô hình cục bộ. Các thí nghiệm trên nhiều bài toán khác nhau, từ phân loại văn bản đến nhận diện hình ảnh, đã chứng minh tính khả thi và hiệu quả của LIME trong việc tạo ra các lời giải thích có ý nghĩa đối với người dùng.

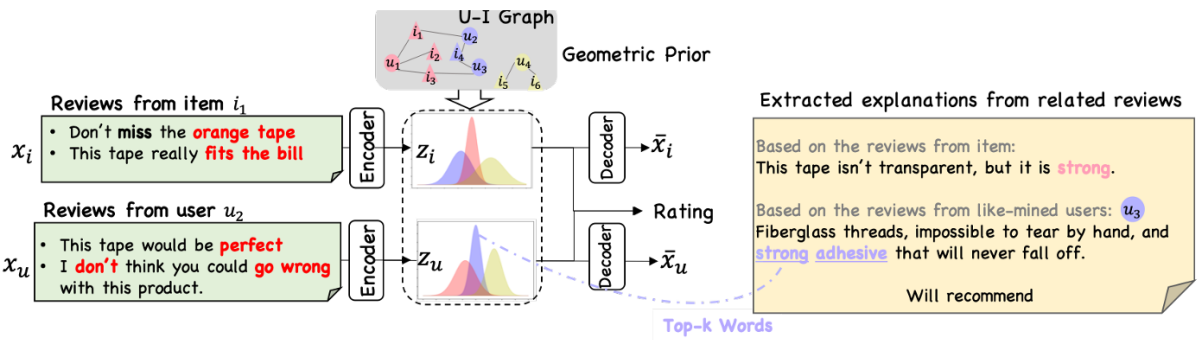
Tuy nhiên, bài báo cũng chỉ ra một số thách thức của LIME, với đặc tính tập trung vào giải thích cục bộ, không thể nắm bắt được cấu trúc toàn cục của mô hình, và kết quả của nó có thể phụ thuộc khá lớn vào cách thức tạo ra các biến thể của đầu vào. Ngoài ra, việc lựa chọn phương pháp nhiễu loạn và thiết lập các tham số phù hợp để đảm bảo tính ổn định của lời giải thích cũng là những vấn đề còn cần được cải tiến. Những thách thức này mở ra hướng nghiên cứu tiếp theo nhằm làm cho các phương pháp giải thích trở nên tin cậy và toàn diện hơn đối với các hệ thống dự đoán phức tạp.

C. USER-CENTRIC EXPLAINABILITY (GIẢI THÍCH THEO HƯỚNG NGƯỜI DÙNG)

Bài báo "Explainable Recommender with Geometric Information Bottleneck" của tác giả Hanqi Yan cùng cộng sự [21] đã giới thiệu một phương pháp mới nhằm cải thiện khả năng giải thích trong các hệ thống khuyến nghị. Các giải thích đa phần đều dựa vào đánh giá của người dùng, do đó thường bị giới hạn và không xác định được các yếu tố tiềm ẩn.

Công trình nghiên cứu đã đưa ra phương pháp tích hợp thông tin hình học học được từ tương tác giữa người dùng và sản phẩm vào một mạng biến phân (variational network) (Hình 5) để suy ra các yếu tố tiềm ẩn từ đánh giá của người dùng và sản phẩm. Điều này giúp tạo ra các giải thích vượt ra ngoài phạm vi của từng đánh giá riêng lẻ. Các thí nghiệm trên ba tập dữ liệu thương mại điện tử cho thấy mô hình này cải thiện đáng kể khả năng diễn giải của hệ thống khuyến nghị. Hệ thống hoạt động như sau:

- Học thông tin hình học từ tương tác người dùng - sản phẩm: Phương pháp đề xuất xây dựng một đồ thị tương tác giữa người dùng và sản phẩm dựa trên các đánh giá và xếp hạng. Từ đồ thị này, các cụm (clusters) của người dùng và sản phẩm được xác định, phản ánh các đặc điểm hoặc sở thích chung.
- Tích hợp thông tin hình học vào mạng biến phân: Các cụm thu được từ bước trước được sử dụng làm thông tin tiên nghiệm (prior) trong mạng biến phân. Cụ thể, mỗi người dùng hoặc sản phẩm được gán một phân phối xác suất mềm trên các cụm, tỷ lệ thuận với khoảng cách đến các tâm cụm. Thông tin tiên nghiệm này giúp điều chỉnh việc học các yếu tố tiềm ẩn từ văn bản đánh giá, đảm bảo rằng các yếu tố này phản ánh cả thông tin toàn cục từ đồ thị tương tác và thông tin cục bộ từ đánh giá.
- Suy ra yếu tố tiềm ẩn và tạo giải thích: Với một cặp người dùng - sản phẩm cụ thể, các đánh giá liên quan được đưa vào mạng biến phân để suy ra các yếu tố tiềm ẩn. Những yếu tố này sau đó được sử dụng để tạo ra các khuyến nghị và giải thích, kết hợp thông tin từ cả tương tác toàn cục và nội dung đánh giá cụ thể.



Hình 5. *Thông tin hình học từ người dùng và sản phẩm*

Công trình nghiên cứu đã mang lại những đóng góp quan trọng cho hệ thống khuyến nghị có giải thích. Một trong những điểm nổi bật của phương pháp này là khả năng tạo ra các lời giải thích không chỉ dựa trên thông tin đánh giá cá nhân mà còn phản ánh cả bối cảnh tổng thể thông qua việc tích hợp thông tin toàn cục và cục bộ từ hệ thống khuyến nghị. Nhờ việc tận dụng thông tin hình học từ suy luận phản thực trên đồ thị, phương pháp giúp cung cấp các lời giải thích sâu sắc và có ý nghĩa hơn cho người dùng. Đặc biệt, dù tập trung vào việc nâng cao tính giải thích của khuyến nghị, mô hình vẫn đảm bảo được hiệu suất dự đoán tương đương với các hệ thống khuyến nghị dựa trên mô hình truyền thống, chứng tỏ được tiềm năng ứng dụng rộng rãi trong thực tế.

Bên cạnh đó, công trình nghiên cứu cũng chỉ ra vấn đề khó khăn trong việc mở rộng cho các tập dữ liệu lớn hoặc phức tạp, và việc học các đặc trưng tiềm ẩn từ đánh giá của người dùng có thể không hoàn toàn nắm bắt được các yếu tố ngữ cảnh hoặc sở thích thay đổi theo thời gian. Ngoài ra, việc đánh giá hiệu quả của các giải thích được tạo ra vẫn cần được nghiên cứu thêm để đảm bảo rằng chúng thực sự mang lại giá trị cho người dùng.

Ngoài ra, trong bài báo "An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission" của các tác giả Luis M. de Campos cùng cộng sự [22] đã đề xuất một hệ thống khuyến nghị dựa trên nội dung, giúp xác định các tạp chí khoa học phù hợp nhất để xuất bản một bài báo cụ thể. Hệ thống đề xuất tạp chí hoạt động bằng cách nhóm các bài viết theo chủ đề và tạo tiểu mục tương ứng cho từng tạp chí. Các tiểu mục này sau đó được lập chỉ mục bằng Lucene và sử dụng mô hình ngôn ngữ Jelinek–Mercer để tính toán mức độ liên quan giữa bài báo cần đề xuất và nội dung các tạp chí. Với mục tiêu là muốn có một khuyến nghị tạp chí để xuất bản, hệ thống sử dụng văn bản của bài báo đó làm truy vấn trong hệ thống truy xuất thông tin (IRS), từ đó xác định các tạp chí phù hợp dựa trên mức độ tương đồng chủ đề. Kết quả xếp hạng tiểu mục được hợp nhất thành xếp hạng tạp chí bằng thuật toán của de Campos et al. (2017) [36] giúp tổng hợp điểm số từ nhiều tiểu mục của cùng một tạp chí, đồng thời giảm ảnh hưởng của các chủ đề ít quan trọng hơn. Cuối cùng, danh sách các tạp chí phù hợp nhất sẽ được đề xuất cho người dùng (Hình 6).

Title	School-Age Outcomes of Early Intervention for Preterm Infants and Their Parents: A Randomized Trial.
Abstract	To examine the child and parental outcomes at school age of a randomized controlled trial of a home-based early preventative care program for infants born very preterm and their caregivers. At term-equivalent age, 120 infants born at a gestational age of <30 weeks were randomly allocated to intervention (n = 61) or standard care (n = 59) groups. The intervention included 9 home visits over the first year of life focusing on infant development, parental mental health, and the parent-infant relationship. At 8 years' corrected age, children's cognitive, behavioral, and motor functioning and parental mental health were assessed. Analysis was by intention to treat. One hundred children, including 13 sets of twins, attended follow-up (85% follow-up of survivors). Children in the intervention group were less likely to have mathematics difficulties (odds ratio , 0.42; 95% confidence interval [CI], 0.18 to 0.98; P = .045) than children in the standard care group, but there was no evidence of an effect on other developmental outcomes. Parents in the intervention group reported fewer symptoms of depression (mean difference, -2.7; 95% CI, -4.0 to -1.4; P < .001) and had reduced odds for mild to severe depression (odds ratio, 0.14; 95% CI, 0.03 to 0.68; P = .0152) than parents in the standard care group. An early preventive care program for very preterm infants and their parents had minimal long-term effects on child neurodevelopmental outcomes at the 8-year follow-up, whereas primary caregivers in the intervention group reported less depression.
Keywords	None
Subjects	None

Recommended Journals

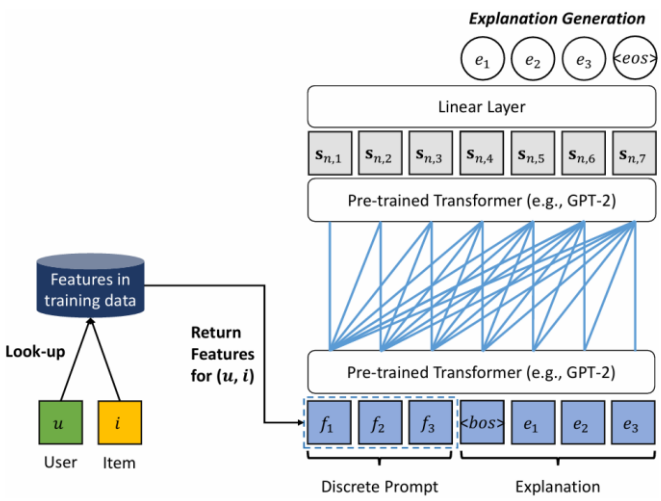
#1	Pediatrics
#2	BMC pediatrics
#3	Early human development
#4	Journal of paediatrics and child health
#5	The Cochrane database of systematic reviews
#6	Developmental medicine and child neurology
#7	JAMA
#8	BMJ Clinical research ed
#9	The Journal of pediatrics
#10	Archives of disease in childhood

Hình 6. *Danh sách các tạp chí được đề xuất*

Sau khi đã có thông tin về bài viết, hệ thống sẽ lấy danh sách các tạp chí, đưa ra đề xuất và lời giải thích. Hệ thống xem xét hai cấp độ giải thích: cấp độ toàn cầu (dựa trên xếp hạng độ tin cậy, chủ đề liên quan) và cấp độ cục bộ (bài viết liên quan đến tạp chí, chủ đề liên quan đến tạp chí, các thuật ngữ liên quan). Kết quả cho thấy rằng tăng cường khả năng giải thích có thể giúp người dùng tin tưởng vào khuyến nghị hơn.

Công trình nghiên cứu đã đóng góp đáng kể cho hệ thống khuyến nghị có giải thích bằng cách giới thiệu một cách tiếp cận mới trong việc giải thích các hệ thống khuyến nghị dựa trên nội dung, đặc biệt trong lĩnh vực đề xuất tạp chí khoa học. Phương pháp đề xuất tích hợp các yếu tố giải thích trực tiếp vào mô hình gợi ý, giúp cải thiện trải nghiệm người dùng và nâng cao độ tin cậy của hệ thống. Bên cạnh đó, nghiên cứu đã cung cấp một minh chứng cụ thể thông qua bài toán đề xuất tạp chí cho việc nộp bài, mở ra hướng đi mới cho các nghiên cứu và ứng dụng trong lĩnh vực này. Tuy nhiên, bài báo cũng chỉ ra thách thức trong việc đo lường mức độ hiệu quả của các giải thích đối với người dùng, do phản hồi của họ mang tính chủ quan và khó định lượng.

Tiếp theo, có thể kể đến công trình nghiên cứu trong bài báo "Personalized Prompt Learning for Explainable Recommendation" của tác giả Lei Li cùng cộng sự [23], bài báo tập trung vào việc cải tiến hệ thống khuyến nghị bằng cách tạo ra các lời giải thích tự nhiên, được cá nhân hóa dựa trên đặc điểm và hành vi của từng người dùng. Cụ thể, công trình này đề xuất một khung làm việc có tên là PEPLER-D (Hình 7), trong đó quá trình học prompt được sử dụng để tạo ra các ngữ cảnh cá nhân hóa, từ đó điều chỉnh lời giải thích mà hệ thống sinh ra. Prompt đề cập đến đoạn văn bản đầu vào được thiết kế đặc biệt nhằm kích hoạt các mô hình LLMs sinh ra các lời giải thích cá nhân hóa cho các đề xuất khuyến nghị. Prompt không chỉ đơn thuần là câu hỏi hay yêu cầu mà còn được bổ sung thông tin cá nhân, chẳng hạn như sở thích và lịch sử tương tác của người dùng, để định hướng cho LLM tạo ra câu trả lời phù hợp với từng cá nhân. Qua đó, prompt đóng vai trò cầu nối giữa dữ liệu cá nhân và mô hình, giúp nâng cao tính minh bạch và khả năng giải thích của hệ thống khuyến nghị.



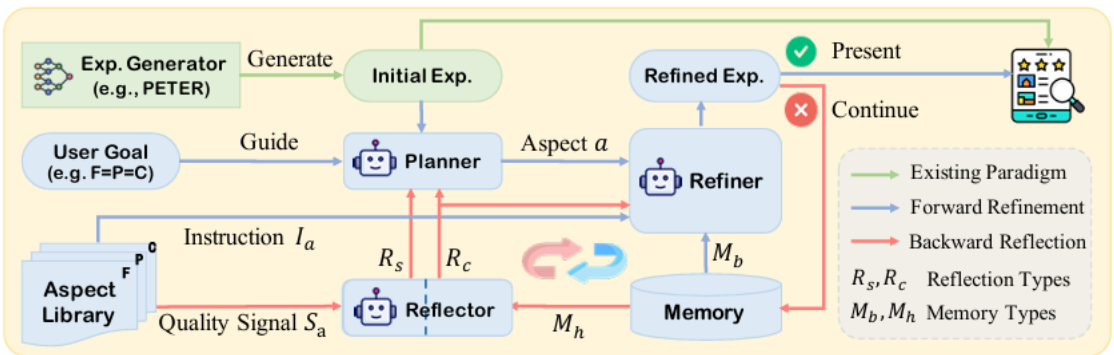
Hình 7. *Khung làm việc PEPLER-D*

Khung làm việc PEPLER-D hoạt động theo cơ chế tích hợp thông tin cá nhân của người dùng vào quá trình tạo prompt, nhằm điều chỉnh đầu ra của mô hình LLM cho phù hợp với từng cá nhân. Cụ thể:

- Mã hóa thông tin người dùng: dữ liệu cá nhân như lịch sử tương tác, sở thích và hành vi được chuyển đổi thành các vector biểu diễn thông qua một bộ mã hóa chuyên dụng. Các vector này chứa đựng những đặc trưng quan trọng giúp nắm bắt bối cảnh và ưu tiên của người dùng.
- Tạo prompt cá nhân hóa: các vector biểu diễn sau đó được tích hợp vào các prompt đầu vào dưới dạng các token đặc biệt hoặc thông tin nhúng. Quá trình này giúp "góp mặt" thông tin người dùng vào ngữ cảnh của yêu cầu, tạo nên một prompt cá nhân hóa, độc đáo cho từng cá nhân.
- Fine-tuning mô hình ngôn ngữ lớn: Prompt cá nhân hóa được đưa vào một mô hình LLM dựa trên kiến trúc Transformer. Mô hình này đã được tinh chỉnh (fine-tuned) để không chỉ dự đoán nội dung mà còn sinh ra lời giải thích minh bạch cho các đề xuất khuyến nghị.
- Sinh lời giải thích và đề xuất: dựa trên prompt cá nhân hóa, LLM tạo ra các lời giải thích tự nhiên, rõ ràng, giải thích lý do tại sao một mục cụ thể được khuyến nghị. Quá trình này kết hợp thông tin ngữ cảnh từ prompt với kiến thức tổng quát của mô hình, giúp các lời giải thích trở nên chặt chẽ và có cơ sở.
- Cơ chế học liên tục: PEPLER-D thường áp dụng cơ chế học liên tục, cho phép cập nhật prompt theo phản hồi từ hệ thống và người dùng. Điều này giúp cải thiện độ chính xác và tính nhất quán của lời giải thích theo thời gian.

Bài báo đã đóng góp quan trọng vào hệ thống khuyến nghị có giải thích bằng cách cá nhân hóa lời giải thích cho từng người dùng, giúp họ hiểu rõ hơn về cơ chế ra quyết định của hệ thống. Phương pháp này cũng tăng cường tính minh bạch bằng cách "mở hộp đen" của các mô hình khuyến nghị, từ đó nâng cao sự tin cậy của người dùng. Đồng thời, việc học prompt giúp hệ thống tận dụng hiệu quả thông tin cá nhân để tối ưu hóa cả khâu dự đoán và giải thích, góp phần cải thiện hiệu năng tổng thể. Tuy nhiên, bài báo cũng chỉ ra một số thách thức như khả năng mở rộng khi cá nhân hóa prompt cho hàng triệu người dùng, yêu cầu tài nguyên tính toán đáng kể. Ngoài ra, việc đảm bảo độ nhất quán và chính xác của lời giải thích, cũng như các vấn đề đạo đức và bảo mật trong sử dụng dữ liệu cá nhân, là những bài toán quan trọng cần được giải quyết.

Tiếp theo có thể kể đến bài báo "Enhancing Recommendation Explanations through User-Centric Refinement" của tác giả Jingsen Zhang cùng cộng sự [24], công trình nghiên cứu tập trung vào cải thiện các lời giải thích trong hệ thống khuyến nghị bằng cách tinh chỉnh dựa trên phản hồi và trải nghiệm của người dùng. Cụ thể, bài báo đề xuất một khung làm việc RefineX (Hình 8), lấy trọng tâm người dùng để cá nhân hóa lời giải thích, từ đó làm cho các lời giải thích trở nên có ý nghĩa, minh bạch và phù hợp với nhu cầu của từng cá nhân.



Hình 8. *Mô hình RefineX*

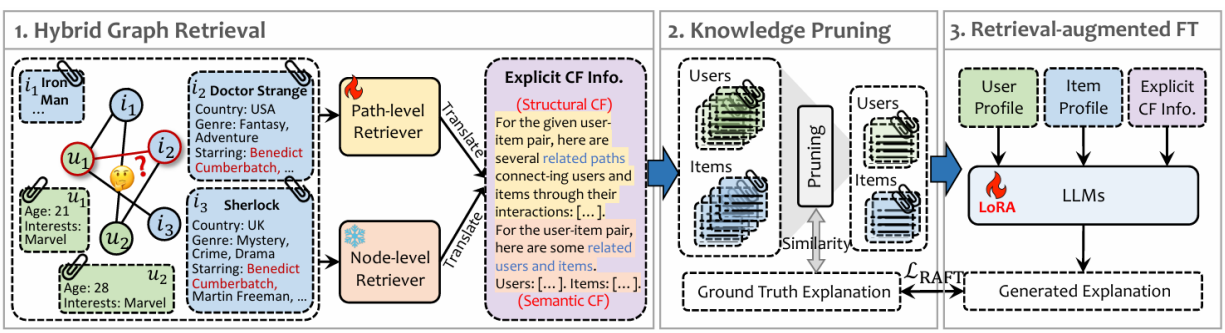
Khung làm việc RefineX được thiết kế nhằm nâng cao chất lượng lời giải thích của hệ thống khuyến nghị thông qua một quá trình tinh chỉnh liên tục dựa trên phản hồi của người dùng. Cụ thể, RefineX hoạt động qua các bước chính sau:

- Sinh lời giải thích ban đầu: hệ thống khuyến nghị tạo ra lời giải thích dựa trên dữ liệu người dùng, lịch sử tương tác và các đặc trưng của sản phẩm.
- Thu thập phản hồi người dùng: lời giải thích ban đầu được trình bày cho người dùng, sau đó hệ thống thu thập phản hồi (đánh giá, nhận xét, hành vi tương tác) để đánh giá mức độ hài lòng và hiểu biết của người dùng về lời giải thích đó.
- Phân tích và đánh giá: phản hồi thu thập được được phân tích để xác định các điểm yếu hoặc khoảng trống trong lời giải thích ban đầu. Quá trình này giúp hệ thống hiểu được những khía cạnh nào cần được cải thiện để trở nên rõ ràng, phù hợp và có giá trị hơn với người dùng.

- Tinh chỉnh lời giải thích: dựa trên phân tích phản hồi, một module tinh chỉnh (refinement module) sẽ điều chỉnh lời giải thích ban đầu. Quá trình này có thể bao gồm việc cập nhật mô hình, re-ranking thông tin hoặc fine-tuning lại các tham số để tạo ra một lời giải thích cá nhân hóa và mạch lạc hơn.
- Phản hồi vòng lặp: lời giải thích được tinh chỉnh sau đó được trình bày lại cho người dùng, qua đó hệ thống tiếp tục nhận phản hồi mới. Vòng lặp này được lặp đi lặp lại nhằm liên tục cải thiện chất lượng và độ chính xác của lời giải thích.

Bài báo đã đóng góp quan trọng trong việc tăng cường tính minh bạch và độ tin cậy của hệ thống khuyến nghị bằng cách cung cấp các lời giải thích có căn cứ, giúp người dùng hiểu rõ hơn về cơ chế đề xuất. Đồng thời, việc đưa ra các lời giải thích phù hợp và dễ hiểu không chỉ cải thiện trải nghiệm người dùng mà còn giúp họ tin tưởng hơn vào hệ thống. Ngoài ra, nghiên cứu cũng đề xuất các phương pháp đánh giá định lượng để đo lường mức độ hữu ích của lời giải thích từ góc nhìn người dùng. Tuy nhiên, vẫn còn một số thách thức như việc thu thập và xử lý phản hồi người dùng một cách hiệu quả để tích hợp vào quá trình tinh chỉnh mô hình. Bên cạnh đó, đảm bảo lời giải thích luôn nhất quán và phù hợp với từng cá nhân là một vấn đề phức tạp, đặc biệt khi dữ liệu phản hồi có thể thay đổi theo thời gian. Cuối cùng, việc xây dựng các chỉ số đánh giá khách quan nhằm đo lường chất lượng và tính hữu ích của lời giải thích vẫn là một bài toán cần tiếp tục nghiên cứu.

Bài báo "G-Refer: Graph Retrieval-Augmented Large Language Model for Explainable Recommendation" của tác giả Yuhan Li cùng cộng sự [25]. Công trình nghiên cứu đề cập đến việc cải tiến hệ thống khuyến nghị thông qua sự kết hợp giữa truy xuất thông tin từ đồ thị tri thức và các mô hình LLMs. Cụ thể, kiến trúc G-Refer (Hình 9) hoạt động bằng cách sử dụng cơ chế truy xuất để lấy các bằng chứng, bối cảnh và mối liên hệ từ đồ thị tri thức liên quan đến người dùng và sản phẩm, sau đó LLM được kích hoạt để sinh ra các lời giải thích tự nhiên, minh bạch và dễ hiểu cho các khuyến nghị.



Hình 9. *Kiến trúc G-Refer*

Kiến trúc G-Refer với ba thành phần chính, giúp hệ thống khuyến nghị không chỉ đưa ra dự đoán mà còn tạo ra lời giải thích rõ ràng, ba thành phần đó bao gồm:

- Thành phần Hybrid Graph Retrieval: Sử dụng các bộ truy xuất đa cấp độ (multi-granularity retrievers) để lấy ra các tín hiệu rõ ràng từ phương pháp lọc cộng tác (Collaborative Filtering - CF). Các tín hiệu này sau đó được chuyển đổi thành văn bản dễ hiểu (human-readable text) thông qua quá trình Graph Translation, tức là chuyển dữ liệu từ đồ thị thành dạng thông tin mà con người có thể đọc và hiểu.
- Thành phần Knowledge Pruning: Thành phần này loại bỏ những thông tin nhiễu (noise) không cần thiết từ dữ liệu đã truy xuất. Việc này không chỉ giúp hệ thống tập trung vào những thông tin hữu ích mà còn cải thiện hiệu quả huấn luyện của mô hình.
- Thành phần Retrieval-augmented Fine-tuning: Thành phần này hướng dẫn các mô hình LLMs sử dụng thông tin CF đã được truy xuất để sinh ra lời giải thích chi tiết. Quá trình fine-tuning kết hợp dữ liệu truy xuất giúp mô hình tạo ra những lời giải thích có căn cứ, mang tính thông tin và minh bạch, hỗ trợ người dùng hiểu rõ lý do đằng sau các đề xuất.

Công trình nghiên cứu đã đạt được những thành tựu quan trọng trong việc tăng cường tính minh bạch của hệ thống khuyến nghị, khi các lời giải thích được tạo ra có căn cứ rõ ràng từ dữ liệu truy xuất, giúp giảm bớt bản chất "hộp đen" của mô hình. Đồng thời, việc kết hợp thông tin đồ thị không chỉ cải thiện hiệu suất khuyến nghị mà còn mang lại những lời giải thích có giá trị thực tiễn cho người dùng. Tuy nhiên, nghiên cứu cũng chỉ ra một số thách thức còn tồn tại, bao gồm chi phí tính toán cao do việc truy xuất và xử lý thông tin từ các đồ thị tri thức quy mô lớn đòi hỏi nhiều tài nguyên. Ngoài ra, cần đảm bảo tính nhất quán giữa dữ liệu truy xuất và lời giải thích do LLM sinh ra, cũng như giải quyết bài toán mở rộng và cập nhật thông tin liên tục khi áp dụng mô hình vào thực tế.

D. NHẬN XÉT CHUNG

Trong những năm gần đây, các nghiên cứu về hệ thống khuyến nghị có giải thích đã có những bước tiến đáng kể, không chỉ tập trung vào việc nâng cao hiệu suất dự đoán mà còn chú trọng đến khả năng minh bạch và dễ hiểu của

hệ thống đối với người dùng. Các công trình được khảo sát cho thấy sự đa dạng trong cách tiếp cận, từ giải thích nội tại trong mô hình, giải thích hậu xử lý đến các phương pháp giải thích theo hướng người dùng. Mỗi hướng tiếp cận đều có những ưu điểm riêng, phù hợp với từng bài toán cụ thể và đặc điểm của dữ liệu đầu vào.

Một điểm chung dễ nhận thấy là vai trò trung tâm của dữ liệu trong việc định hình chất lượng và mức độ hữu ích của lời giải thích. Các mô hình sử dụng dữ liệu giàu ngữ cảnh hoặc kết hợp nhiều nguồn thông tin thường có khả năng tạo ra lời giải thích sâu sắc và sát thực hơn. Tuy nhiên, những hệ thống này cũng đòi hỏi xử lý dữ liệu phức tạp và chi phí tính toán cao hơn.

Nhằm hệ thống hóa các công trình nghiên cứu, bảng sau được xây dựng để so sánh các công trình tiêu biểu dựa trên phương pháp giải thích, các ưu điểm cũng như những hạn chế còn tồn tại.

Bảng 1. *So sánh các mô hình trong hệ thống khuyến nghị có giải thích*

Tên mô hình	Phương pháp giải thích	Ưu điểm	Hạn chế
KAERR	Giải thích dựa trên mô hình	Giải thích hai chiều, phù hợp bối cảnh	Khó xử lý dữ liệu tương tác khan hiếm
GaVaMoE	Giải thích dựa trên mô hình	Cá nhân hóa sâu, giảm thiếu hụt dữ liệu	Tính toán phức tạp, khó mở rộng
LLM2ER-EQR	Giải thích dựa trên mô hình	Tăng chất lượng giải thích, nhất quán	Cần tài nguyên tính toán cao
Xrec	Giải thích sau khi ra kết quả	Giải thích toàn diện, dễ hiểu	Khó đồng bộ giữa ngữ nghĩa và tín hiệu
LIME	Giải thích sau khi ra kết quả	Dễ hiểu, linh hoạt	Không bao quát toàn cục
PEPLER-D	Giải thích theo hướng người dùng	Cá nhân hóa lời giải thích, tự nhiên và rõ ràng	Tài nguyên lớn, vấn đề riêng tư
RefineX	Giải thích theo hướng người dùng	Tinh chỉnh giải thích dựa trên phản hồi người dùng	Khó thu thập và đồng bộ phản hồi
G-Refer	Giải thích theo hướng người dùng	Giải thích có căn cứ rõ ràng từ dữ liệu đồ thị	Chi phí tính toán cao, yêu cầu xử lý dữ liệu

IV. DATASET TRONG HỆ THỐNG KHUYẾN NGHỊ CÓ GIẢI THÍCH

Dataset là một tập hợp các dữ liệu. Dataset tương ứng với các nội dung có trong một bảng cơ sở dữ liệu hoặc là một ma trận của những dữ liệu thống kê, trong đó mỗi cột của bảng tính sẽ đại diện cho một biến cụ thể nhất định và mỗi hàng sẽ tương ứng với một thành viên cụ thể nhất định nào đó thuộc một tập dữ liệu được đề cập đến [26].

Tùy vào từng loại dataset khác nhau, các kỹ thuật cũng được áp dụng theo cách phù hợp để đảm bảo tính minh bạch và dễ hiểu trong hệ thống khuyến nghị. Dưới đây là các kỹ thuật phổ biến ứng với từng loại dữ liệu, có thể kể đến như sau:

- **Dataset dạng bảng** chứa thông tin có cấu trúc như dữ liệu giao dịch, thông tin khách hàng, hoặc doanh số bán hàng. Trong hệ thống khuyến nghị, loại dữ liệu này thường được sử dụng để dự đoán hành vi người dùng dựa trên các thuộc tính cụ thể. Có thể áp dụng các kỹ thuật như SHAP [4] để tính toán mức độ đóng góp của từng thuộc tính như giá sản phẩm, thương hiệu, danh mục vào quyết định khuyến nghị hoặc kỹ thuật Rule-based Explanations để diễn giải cách mô hình đưa ra khuyến nghị, ví dụ: "Nếu người dùng mua sản phẩm A, khả năng cao họ sẽ thích sản phẩm B".
- **Với dataset hình ảnh** (Image Data) thì hệ thống khuyến nghị hình ảnh thường được áp dụng trong thương mại điện tử như gợi ý sản phẩm thời trang hoặc nhận diện nội dung đa phương tiện. Trong bài báo "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" của tác giả Ramprasaath R. Selvaraju cùng cộng sự [27], các tác giả đã giới thiệu phương pháp Gradient-weighted Class Activation Mapping (Grad-CAM) nhằm tạo ra các giải thích trực quan cho các quyết định của mô hình mạng nơ-ron tích chập (CNN). Phương pháp này sử dụng thông tin gradient liên quan đến một lớp mục tiêu để tạo ra bản đồ kích hoạt, giúp xác định các vùng quan trọng trong ảnh ảnh hưởng đến dự đoán của mô hình. Trong nghiên cứu, các tác giả đã áp dụng Grad-CAM trên nhiều mô hình và nhiệm vụ khác nhau để minh họa khả năng giải thích của phương pháp này. Cụ thể, sử dụng các mô hình như VGGNet và GoogleNet, được huấn luyện trên tập dữ liệu ImageNet, để xác định các vùng ảnh hưởng đến quyết định phân loại trong bài toán phân loại ảnh. Đối với nhiệm vụ chú thích ảnh (Image Captioning), Grad-CAM

được áp dụng để xác định các vùng trong ảnh mà mô hình tập trung khi tạo ra chú thích. Ngoài ra, trong bài toán trả lời câu hỏi trực quan (Visual Question Answering - VQA), Grad-CAM giúp xác định các vùng hình ảnh liên quan đến câu hỏi và câu trả lời, làm rõ cách mô hình đưa ra phản hồi dựa trên thông tin trực quan.

- **Với dataset chuỗi thời gian** (Time Series Data) thường được sử dụng để phân tích hành vi người dùng theo thời gian, dự báo nhu cầu sản phẩm hoặc xu hướng tiêu dùng. Trong bài báo "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR" của tác giả Sandra Wachter cùng cộng sự [28], các tác giả tập trung vào việc cung cấp giải thích cho các quyết định tự động mà không cần tiết lộ nội dung bên trong của mô hình (hộp đen). Phương pháp giải thích phản thực tế được đề xuất nhằm giúp cá nhân hiểu rõ hơn về các quyết định tự động và cách họ có thể thay đổi kết quả bằng cách điều chỉnh các yếu tố đầu vào. Thay vì giải thích trực tiếp cách hoạt động của mô hình, phương pháp này cung cấp cho người dùng thông tin về những thay đổi cần thiết trong đầu vào để đạt được một kết quả khác. Điều này giúp người dùng hiểu rõ hơn về mối quan hệ giữa các đặc trưng đầu vào và kết quả đầu ra, đồng thời cung cấp các gợi ý cụ thể và có thể hành động được để thay đổi kết quả trong tương lai.

Trong mô hình XRec [16], các tác giả đã áp dụng khung XRec trên ba tập dữ liệu công khai gồm Amazon-books, Google-reviews và Yelp. Các tập dữ liệu này chứa thông tin về tương tác giữa người dùng và sản phẩm/dịch vụ, được sử dụng để huấn luyện và đánh giá khả năng của mô hình trong việc tạo ra các giải thích dễ hiểu cho các khuyến nghị. Để chuẩn bị dữ liệu cho mô hình, các tác giả đã tạo hồ sơ người dùng và sản phẩm, cũng như các giải thích, bằng cách sử dụng các công cụ xử lý ngôn ngữ tự nhiên. Quá trình này bao gồm việc sử dụng các tập lệnh Python để tạo hồ sơ và giải thích từ đầu, đảm bảo rằng mô hình có thể hiểu và tạo ra các giải thích phù hợp dựa trên thông tin từ các tập dữ liệu này.

Trong kiến trúc G-Refer [25], các tác giả đã áp dụng khung G-Refer trên ba tập dữ liệu công khai gồm Amazon-books, Yelp, và Google-reviews. Các tập dữ liệu này được sử dụng để đánh giá hiệu suất của G-Refer trong việc cung cấp các khuyến nghị có thể giải thích được. Cụ thể, các tác giả đã áp dụng cơ chế truy xuất đồ thị kết hợp để trích xuất các tín hiệu lọc cộng tác rõ ràng từ cả góc độ cấu trúc và ngữ nghĩa. Thông tin CF được truy xuất sau đó được chuyển đổi thành văn bản dễ hiểu để hỗ trợ các mô hình LLMs trong việc tạo ra các giải thích cho khuyến nghị.

Nhằm hệ thống hóa và làm rõ vai trò của dữ liệu trong các mô hình khuyến nghị có giải thích, bảng sau được xây dựng để so sánh các công trình tiêu biểu dựa trên nguồn dữ liệu sử dụng, đặc điểm nổi bật của từng loại dữ liệu, cũng như ảnh hưởng của chúng đến khả năng giải thích mà mô hình mang lại.

Bảng 2. *So sánh dataset trong các mô hình khuyến nghị có giải thích*

Tên mô hình	Nguồn dữ liệu sử dụng	Đặc điểm dữ liệu	Tác động đến giải thích
KAERR	Dữ liệu từ hệ thống hẹn hò và tuyển dụng	Tương tác song phương, khan hiếm	Cần phản ánh ý định của cả hai bên
GaVaMoE	3 tập dữ liệu thực tế	Thiếu hụt tương tác người dùng	Dùng GMM + VAE để mô hình hóa sở thích phức tạp
LLM2ER-EQR	Dữ liệu có chất lượng thấp	Không đồng nhất, nhiều nhiễu	Fine-tuning giúp sinh giải thích ổn định
Xrec	Dữ liệu hành vi + văn bản + đồ thị	Kết hợp tín hiệu cộng tác và ngữ nghĩa	Giải thích toàn diện nhưng phức tạp
LIME	Dữ liệu bất kỳ: ảnh, văn bản, bảng…	Tự do, mô hình bất khả tri (model-agnostic)	Giải thích cục bộ, dễ tiếp cận
PEPLER-D	Lịch sử hành vi cá nhân	Dữ liệu cá nhân hóa cao	Tạo prompt sinh giải thích riêng từng người
RefineX	Phản hồi người dùng về lời giải thích	Dữ liệu định tính + hành vi	Giải thích được cá nhân hóa theo phản hồi
G-Refer	Đồ thị tri thức + truy xuất + văn bản mô tả	Truy xuất từ graph lớn, cần xử lý nhiễu	Giải thích giàu ngữ nghĩa, rõ ràng

V. HƯỚNG PHÁT TRIỂN

A. THÁCH THỨC

Hệ thống khuyến nghị có giải thích đã và đang thu hút sự quan tâm rộng rãi trong cộng đồng nghiên cứu, nhờ vào khả năng nâng cao tính minh bạch và độ tin cậy của các khuyến nghị được đưa ra. Mặc dù đã có nhiều tiến bộ đáng kể trong việc phát triển các phương pháp giải thích, nhưng hệ thống vẫn phải đối mặt với nhiều thách thức quan trọng, có thể kể đến như sau:

- **Hạn chế trong khả năng mô hình hóa sở thích phức tạp**: Các mô hình khuyến nghị dựa trên LLMs hiện nay gặp khó khăn trong việc nắm bắt sở thích phức tạp giữa người dùng và sản phẩm. Thêm vào đó, mô hình cần xử lý tình huống dữ liệu tương tác bị thiếu hụt, đồng thời cá nhân hóa các giải thích để phù hợp với từng người dùng. Việc này có thể thấy rõ trong mô hình GaVaMoE [14].
- **Hiệu năng và tài nguyên tính toán khi tinh chỉnh LLMs trong hệ thống khuyến nghị**: Các mô hình LLM có kích thước lớn như GPT-3 [13] tuy mang lại bước tiến lớn cho xử lý ngôn ngữ tự nhiên nhưng vẫn gặp các thách thức lớn về hiệu suất, đạo đức và tính giải thích. Việc tối ưu hóa các mô hình này để đảm bảo vừa có độ chính xác cao vừa có thể cung cấp lời giải thích hợp lý vẫn là một thách thức chưa được giải quyết triệt để. Bên cạnh đó, việc tinh chỉnh các mô hình LLMs cũng đòi hỏi tài nguyên tính toán rất lớn, làm tăng chi phí triển khai và hạn chế khả năng mở rộng của hệ thống khuyến nghị có giải thích.
- **Tính nhất quán và ổn định của lời giải thích trong hệ thống khuyến nghị**: Một trong những hạn chế lớn của các mô hình dựa trên suy luận phản thực là việc lựa chọn và lọc các đường dẫn từ đồ thị tri thức để đảm bảo rằng chúng phản ánh chính xác mối quan hệ giữa người dùng và mục tiêu. Ngoài ra, vấn đề nhất quán trong các lời giải thích vẫn chưa được giải quyết hoàn toàn. Các hệ thống cần đảm bảo rằng các giải thích đưa ra phản ánh đúng logic hoạt động của mô hình và không bị thay đổi đáng kể khi dữ liệu hoặc bối cảnh thay đổi. Có thể thấy ở mô hình của tác giả Haojie Trang [19] cùng cộng sự.
- **Thử thách về cá nhân hóa lời giải thích**: Mặc dù phương pháp cá nhân hóa giúp người dùng dễ hiểu hơn về khuyến nghị của hệ thống, nhưng việc tạo ra lời giải thích phù hợp cho từng cá nhân trên quy mô lớn đòi hỏi lượng tài nguyên tính toán đáng kể. Một thách thức khác là làm thế nào để đảm bảo lời giải thích phản ánh đúng sở thích của người dùng mà vẫn đảm bảo tính ổn định và chính xác của hệ thống. Cụ thể có thể thấy thử thách đó ở khung làm việc PEPLER-D [23].
- **Hiện tượng "ảo giác" trong mô hình sinh lời giải thích**: Một trong những vấn đề quan trọng của các hệ thống khuyến nghị có giải thích dựa trên mô hình sinh ngôn ngữ là hiện tượng "ảo giác". Các mô hình có thể tạo ra các giải thích có vẻ hợp lý nhưng thực tế lại không có cơ sở dữ liệu hỗ trợ, gây mất tin cậy cho hệ thống khuyến nghị.
- **Vấn đề đánh giá chất lượng lời giải thích vẫn còn nhiều hạn chế**, hầu hết các nghiên cứu chỉ dừng lại ở đánh giá định tính hoặc khảo sát người dùng, chưa có các thước đo định lượng chuẩn hóa để so sánh khách quan giữa các phương pháp giải thích.
- **Việc áp dụng các mô hình giải thích vào thực tiễn gặp rào cản về chi phí tính toán và khả năng mở rộng**. Ví dụ, mô hình GaVaMoE [14] mặc dù cung cấp giải thích cá nhân hóa nhưng có độ phức tạp tính toán cao, khó triển khai trên quy mô lớn. Tương tự, LLM2ER-EQR [15] đòi hỏi tài nguyên tính toán đáng kể khi fine-tuning mô hình ngôn ngữ lớn, gây khó khăn khi triển khai thực tế.
- **Phần lớn các nghiên cứu hiện nay tập trung vào các lĩnh vực phổ biến như thương mại điện tử hoặc phim**, trong khi các lĩnh vực nhạy cảm như y tế, tài chính, giáo dục vẫn chưa được khai thác đầy đủ, dù đây là những lĩnh vực yêu cầu tính minh bạch cao trong quyết định.

B. HƯỚNG PHÁT TRIỂN

1. CẢI THIỆN KHẢ NĂNG MỞ RỘNG VÀ TỐI ƯU HÓA TÍNH TOÁN

Một trong những thách thức chính của hệ thống khuyến nghị có giải thích là chi phí tính toán cao, đặc biệt khi sử dụng mô hình LLMs hoặc các phương pháp học sâu kết hợp với đồ thị. Điều này đặt ra nhu cầu nghiên cứu các kỹ thuật giảm chi phí tính toán mà vẫn duy trì chất lượng giải thích, chẳng hạn như:

- Tối ưu hóa bộ nhớ và tài nguyên tính toán: Nghiên cứu các thuật toán giảm chiều dữ liệu và tối ưu hóa tham số mô hình giúp giảm dung lượng bộ nhớ và tốc độ suy luận của mô hình mà không làm suy giảm chất lượng dự đoán.
- Phân tán và song song hóa: Sử dụng các phương pháp học phân tán (Distributed Learning) hoặc Federated Learning [29] để giảm tải tài nguyên tính toán bằng cách chia nhỏ dữ liệu và xử lý trên nhiều thiết bị khác nhau.
- Chiến lược khởi tạo prompt thông minh: Trong các phương pháp dựa trên mô hình LLMs, cần nghiên cứu cách thiết kế prompt tối ưu để tăng hiệu quả giải thích mà không làm gia tăng chi phí tính toán.

2. CẢI THIỆN TÍNH NHẤT QUÁN VÀ CHẤT LƯỢNG CỦA LỜI GIẢI THÍCH

Mặc dù nhiều phương pháp đã được đề xuất để cải thiện tính giải thích của hệ thống khuyến nghị, nhưng một trong những thách thức lớn là đảm bảo sự nhất quán và chính xác của các giải thích. Một số hướng nghiên cứu có thể giải quyết vấn đề này gồm:

- Tích hợp các mô hình kiểm tra tính hợp lý của giải thích: Sử dụng mô hình kiểm định tính nhất quán để đánh giá xem lời giải thích có phản ánh chính xác quyết định của hệ thống không. Các kỹ thuật như SHAP

(Shapley Additive Explanations) có thể giúp xác định mức độ đóng góp của từng đặc trưng vào quyết định khuyến nghị.
- Giảm hiện tượng "ảo giác" (Hallucination): Các mô hình LLMs có thể tạo ra những giải thích không có căn cứ trong dữ liệu gốc. Do đó, cần phát triển các phương pháp như Knowledge Pruning để loại bỏ thông tin nhiễu và đảm bảo lời giải thích có cơ sở logic chặt chẽ.
- Đánh giá chất lượng lời giải thích dựa trên phản hồi của người dùng: Xây dựng các bộ chỉ số tự động đo lường độ tin cậy và tính hữu ích của lời giải thích thông qua phản hồi từ người dùng thực tế.

3. CÁ NHÂN HÓA LỜI GIẢI THÍCH ĐỂ NÂNG CAO TRẢI NGHIỆM NGƯỜI DÙNG

Lời giải thích có thể trở nên hữu ích hơn khi được tùy chỉnh theo nhu cầu và đặc điểm cá nhân của người dùng. Tuy nhiên, việc cá nhân hóa lời giải thích ở quy mô lớn vẫn còn nhiều thách thức. Một số hướng nghiên cứu tiềm năng bao gồm:

- Học Prompt cá nhân hóa: Áp dụng kỹ thuật Personalized Prompt Learning, như mô hình PEPLER hoặc XRec, để hướng dẫn mô hình LLM tạo ra lời giải thích phù hợp với từng người dùng.
- Tích hợp phản hồi của người dùng vào quá trình huấn luyện: Sử dụng Reinforcement Learning từ phản hồi của người dùng (RLHF - Reinforcement Learning from Human Feedback) [30] giúp điều chỉnh lời giải thích theo sở thích cá nhân.
- Đánh giá hiệu quả của lời giải thích theo từng nhóm người dùng: Cần nghiên cứu về các đặc điểm người dùng (chẳng hạn như mức độ hiểu biết về hệ thống khuyến nghị) để cung cấp lời giải thích phù hợp với từng đối tượng.

4. KẾT HỢP NHIỀU PHƯƠNG PHÁP GIẢI THÍCH KHÁC NHAU ĐỂ TẠO RA LỜI GIẢI THÍCH ĐA DẠNG

Các nghiên cứu hiện tại thường tập trung vào một phương pháp giải thích cụ thể, nhưng thực tế cho thấy mỗi người dùng có thể cần nhiều loại giải thích khác nhau. Một số hướng phát triển có thể bao gồm:

- Kết hợp các phương pháp giải thích: Phát triển các mô hình lai, tích hợp nhiều phương pháp như giải thích dựa trên đặc trưng (Feature Importance Analysis), mô hình thay thế (Surrogate Models), Rule-Based Explainability, và Attention-Based Explanation để tăng tính toàn diện của lời giải thích.
- Sử dụng phương pháp học sâu (Deep Learning) kết hợp với kỹ thuật SHAP: Nhằm xác định chính xác hơn tầm quan trọng của từng đặc trưng và tạo ra lời giải thích trực quan hơn cho người dùng.
- Khai thác đồ thị tri thức và lý thuyết trò chơi: Như trong mô hình G-Refer, việc kết hợp Hybrid Graph Retrieval và Counterfactual Reasoning có thể giúp giải thích không chỉ từ đặc trưng của sản phẩm mà còn từ mối quan hệ giữa các sản phẩm, người dùng và các yếu tố ngữ nghĩa khác.

5. ĐẢM BẢO BẢO MẬT VÀ TÍNH ĐẠO ĐỨC TRONG GIẢI THÍCH KHUYẾN NGHỊ

Việc cung cấp lời giải thích dựa trên dữ liệu cá nhân đặt ra nhiều thách thức về quyền riêng tư và bảo mật thông tin, đặc biệt trong bối cảnh dữ liệu người dùng có thể bị lạm dụng hoặc tiết lộ ngoài ý muốn. Một số hướng phát triển quan trọng bao gồm:

- Áp dụng các kỹ thuật bảo vệ dữ liệu như Differential Privacy [31] và Federated Learning [29] để bảo vệ thông tin cá nhân mà không làm giảm hiệu suất của mô hình.
- Xây dựng cơ chế kiểm soát quyền riêng tư cho phép người dùng tùy chỉnh mức độ chi tiết của lời giải thích và thông tin cá nhân được sử dụng trong hệ thống khuyến nghị.
- Giải quyết vấn đề thiên vị và sự công bằng trong khuyến nghị để đảm bảo hệ thống không thiên vị theo giới tính, chủng tộc hoặc các đặc điểm cá nhân khác, giúp tạo ra lời giải thích khách quan và công bằng.

VI. KẾT LUẬN

Bài viết đã khảo sát toàn diện các phương pháp giải thích trong hệ thống khuyến nghị có giải thích, bao gồm giải thích dựa trên mô hình, giải thích sau khi ra kết quả và giải thích theo hướng người dùng. Qua phân tích các công trình tiêu biểu, có thể thấy mặc dù lĩnh vực này đã đạt được nhiều tiến bộ đáng kể, nhưng vẫn còn tồn tại nhiều thách thức cần giải quyết như chi phí tính toán cao, hạn chế trong khả năng cá nhân hóa lời giải thích, cũng như các vấn đề liên quan đến bảo mật. Trong thời gian tới, các nghiên cứu có thể tập trung vào việc tối ưu hóa chi phí và khả năng mở rộng mô hình, đảm bảo tính nhất quán và chất lượng của lời giải thích, phát triển các phương pháp cá nhân hóa phù hợp với từng người dùng, kết hợp đa dạng các phương pháp giải thích, và xây dựng cơ chế bảo mật thông tin hiệu quả. Những hướng đi này không chỉ giúp nâng cao chất lượng và độ tin cậy của hệ thống khuyến nghị có giải thích mà còn góp phần mở rộng phạm vi ứng dụng của chúng trong các lĩnh vực quan trọng như y tế, tài chính và giáo dục.

VII. TÀI LIỆU THAM KHẢO

[1] *Hong-Quan Do (2024).* Hệ thống khuyến nghị trong bối cảnh Dữ liệu lớn, Trường Đại học FPT, *https://csr.greenwich.edu.vn/paper-sharing-he-thong-khuyen-nghi-trong-boi-canh-du-lieu-lon/25/4/2024.* Ngày truy cập 20/02/2025.

[2] Yongfeng Zhang and Xu Chen (2020), *Explainable Recommendation: A Survey and New Perspectives,* Foundations and Trends® in Information Retrieval: Vol. 14, No. 1, pp 1–101. DOI: 10.1561/1500000066.

[3] Boxuan Ma, Tianyuan Yang, and Baofeng Ren (2024), *A Survey on Explainable Course Recommendation Systems,* DOI: 10.1007/978-3-031-60012-8_17.

[4] Scott M. Lundberg, Su-In Lee (2017), *A Unified Approach to Interpreting Model Predictions,* Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp.4765-4774.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin (2017), *Attention Is All You Need,* Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, pp. 5998-6008.

[6] Mohamed Hussein Abdi, George Onyango Okeyo & Ronald Waweru Mwangi (2018), *Matrix Factorization Techniques for Context-Aware Collaborative Filtering Recommender Systems: A Survey,* Computer and Information Science; Vol. 11, No. 2, pp.1-22.

[7] Nickil Maveli, Demystifying Post-hoc Explainability for ML models, *https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability.* Access date: 14/03/2025.

[8] Nava Tintarev, Judith Masthoff (2007), *A Survey of Explanations in Recommender Systems,* Proceedings of the 7th International Conference on Intelligent User Interfaces, Honolulu, HI, USA, pp.47-54.

[9] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumnin Chen, Zhaochun Ren (2023), *Towards Explainable Conversational Recommender Systems,* SIGIR '23, July 23–27, 2023, Taipei, Taiwan.

[10] Mohamed Amine Chatti, Mohamed Guesmi, Abderrazak Muslim (2023), *Visualization for Recommendation Explainability: A Survey and New Perspectives,* ACM Transactions on Interactive Intelligent Systems, Volume 14, Issue 3, Article No.19, pp.1-40, https://doi.org/10.1145/3672276.

[11] Kathing Wardatzky, Oana Inel, Luca Rossetto, Abraham Bernstein (2025) , *Whom do Explanations Serve? A Systematic Literature Survey of User Characteristics in Explainable Recommender Systems Evaluation,* arXiv:2412.14193v2 [cs. HC], 3 Feb 2025.

[12] Kai-Huang Lai, Zhe-Rui Yang, Pei-Yuan Lai, Chang-Dong Wang, Mohsen Guizani, Min Chen (2024), *Knowledge-Aware Explainable Reciprocal Recommendation,* TheThirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24), pp.8636-8644.

[13] TomB. Brown, Benjamin Mann, Prafulla Dhariwal, Nick Ryder, Arvind Neelakantan, Melanie Subbiah, Pranav Shyam, Amanda Askell, Rewon Child, Sandhini Agarwal, Aditya Ramesh, Christopher Hesse, Benjamin Chess, Mark Chen, Ariel Herbert-Voss, Daniel M. Ziegler, Eric Sigler, Jack Clark, Gretchen Krueger, Jeffrey Wu, Mateusz Litwin, Girish Sastry, TomHenighan, Clemens Winter, Scott Gray, Christopher Berner, SamMcCandlish, Alec Radford, Ilya Sutskever, Dario Amodei (2020), *Language Models are Few-Shot Learners,* Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2020), pp. 1877-1901.

[14] Fei Tang, Yongliang Shen, Hang Zhang, Zeqi Tan, Wenqi Zhang, Guiyang Hou, Kaitao Song, Weiming Lu, Yueting Zhuang (2024), *GaVaMoE: Gaussian-Variational Gated Mixture of Experts for Explainable Recommendation,* arXiv:2410.11841v1 [cs.IR], 15 Oct 2024.

[15] Mengyuan Yang, Mengying Zhu, Yan Wang, Linxun Chen, Yilei Zhao, Xiuyuan Wang, Bing Han, Xiaolin Zheng, Jianwei Yin (2024), *Fine-Tuning Large Language Model Based Explainable Recommendation with Explainable Quality Reward,* The Thirty-Eight AAAI Conference on Artificial Intelligence (AAAI-24), pp.9250-9259.

[16] Qiyao Ma, Xubin Ren, Chao Huang (2024), *XRec: Large Language Models for Explainable Recommendation,* arXiv:2406.02377v2 [cs.IR], 22 Sep 2024.

[17] *Yicong Li, Xiangguo Sun, Hongxu Chen, Sixiao Zhang, Yu Yang, Guandong Xu (2024), Attention Is Not the Only Choice: Counterfactual Reasoning for Path-Based Explainable Recommendation,* IEEE Transactions on knowledge and data engineering, 11 Jan 2024.

[18] Đăng Khoa (2024), Chuyên gia lý giải tình trạng "ảo giác" của AI, *https://viettimes.vn/chuyen-gia-ly-giai-tinh-trang-ao-giac-cua-ai-post175852.html.* Ngày truy cập: 28/02/2025.

[19] Haojie Trang, Wei Emma Zhang, Weitong Chen, Jian Yang, Quan Z. Sheng (2024), *Counterfactual Path Augmentation for Reinforcement Reasoning in Explainable Recommendation,* ACMTrans. Intell. Syst. Technol., Vol. 16, No. 1, Article 9. Publication date: December 2024.

[20] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016), *Why Should I Trust You? Explaining the Predictions of Any Classifier,* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 1135-1144.

[21] Hanqi Yan, Lin Gui, Menghan Wang, Kun Zhang, Yulan He (2024), *Explainable recommender with geometric information bottleneck.* IEEE Transactions on Knowledge and Data Engineering, 36(7), 3036–3046. https://doi.org/10.1109/TKDE.2024.3350447.

[22] Luis M. de Campos, Juan M. Fernández-Luna và Juan F. Huete (2024), *An explainable content-based approach for recommender systems: a case study in journal recommendation for paper submission,* User Modeling and User-Adapted Interaction (2024) 34:1431–1465, https://doi.org/10.1007/s11257-024-09400-6.

[23] Lei Li, Yongfeng Zhang, Li Chen (2023), *Personalized Prompt Learning for Explainable Recommendation, Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, pp. 1178-1188.

[24] Jingsen Zhang, Zihang Tian, Xueyang Feng, Xu Chen (2025), *Enhancing Recommendation Explanations through User-Centric Refinement,* arXiv:2502.11721v1 [cs.IR], 17 Feb 2025.

[25] Yuhan Li, Xinni Zhang, Linhao Luo, Heng Chang, Yuxiang Ren, Irwin King, Jia Li (2025), *G-Refer: Graph Retrieval-Augmented Large Language Model for Explainable Recommendation,* arXiv:2502.12586v1 [cs.IR], 18 Feb 2025.

[26] Giang Nguyễn (2024), Giải mã khái niệm dataset. Các loại dataset nào được sử dụng trong học máy, *https://fptshop.com.vn/tin-tuc/danh-gia/dataset-174624.* Ngày truy cập: 09/07/2025.

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra (2019), *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626.

[28] Sandra Wachter, Brent Mittelstadt, Chris Russell (2018), *Counterfactual Explanations Without Opening The Black Box: Automated Decisions And The Gdpr,* Harvard Journal of Law & Technology, Vol. 31, No. 2, pp. 841-887.

[29] Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, Lizhen Cui (2022), *A Survey on Federated Recommendation Systems, IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 4, pp. 1478-1498.

[30] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe (2022), *Training language models to follow instructions with human feedback,* Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, Louisiana, USA, pp. 27735-27750.

[31] Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, Wanrong Zhang (2024), *Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment,* https://hdsr.mitpress.mit.edu/pub/sl9we8gh/release/3, Access date: 16/03/2025.

A SURVEY ON EXPLAINABLE RECOMMENDATION FROM METHODS AND APPLICATIONS TO CHALLENGES

Tran Nguyen Minh Hieu, Tran Van Lang, Nguyen Quoc Huy

ABSTRACT— Explainable Recommendation Systems (ERS) not only provide relevant recommendations but also offer clear explanations to enhance transparency and user trust. This paper surveys the major explanation approaches in ERS, including model-based, post-hoc, and user-centric methods, and analyzes representative studies applying SHAP, LIME, PEPLER-D, GaVaMoE, and G-Refer. The results highlight several critical challenges, such as limited capability to model complex user preferences, high computational costs when using LLMs, hallucination phenomena in explanations, lack of standardized datasets and quantitative evaluation metrics, as well as potential risks to user data privacy. To address these issues, potential future directionts are proposed, including optimizing computational cost and scalability, ensuring explanation consistency and quality, personalizing explanations for individual users, integrating multiple explanation methods for comprehensiveness, and developing privacy-preserving and ethical mechanisms for ERS. This study provides a systematic overview that informs future research to improve the quality and practical applicability of ERS across various domains.

Keyword— Explainable Recommendation, Model-Based Explainability, Post-Hoc Explainability, User-centric explainability

Trần Nguyễn Minh Hiếu là Thạc sĩ Hệ thống thông tin. Cô tốt nghiệp Cử nhân ngành Công nghệ thông tin tại trường Đại học Khoa học Tự nhiên Tp.HCM (thuộc ĐHQG-HCM). Hiện cô đang là chuyên viên của Phòng Đào tạo đồng thời là giảng viên của khoa Công nghệ thông tin trường Đại học Sài Gòn. Lĩnh vực mà cô

Trần Văn Lăng là Phó Giáo sư ngành Công nghệ thông tin. Ông tốt nghiệp Cử nhân tại Khoa Toán Trường Đại học Tổng hợp TP.HCM năm 1982, nay là trường Đại học Khoa học Tự nhiên (thuộc ĐHQG-HCM). Ông cũng đã nhận học vị Tiến sĩ Toán - Lý tại trường này vào năm 1995. Trong sự nghiệp của mình, ông đã đảm nhiệm nhiều chức vụ

nghiên cứu quan tâm đó là Hệ thống khuyến nghị có giải thích.

Nguyễn Quốc Huy hoàn thành tiến sĩ ngành Khoa học máy tính vào năm 2014 tại viện JAIST, Nhật Bản. Tốt nghiệp Cử nhân, Thạc sỹ tại trường Đại học Khoa học Tự nhiên Tp.HCM (thuộc ĐHQG-HCM). Hiện đang công tác tại Khoa Công nghệ thông tin - Đại học Sài Gòn, đảm nhiệm chức vụ Trưởng Bộ môn Kỹ thuật phần mềm. Hiện ông có trên 30 công trình đăng trên các tạp chí khoa học trong và ngoài nước theo các hướng nghiên cứu Máy học, Khai thác dữ liệu. Bên cạnh đó, ông cũng đã hướng dẫn thành công nhiều học viên tốt nghiệp đại học và cao học chuyên ngành Kỹ thuật phần mềm và Khoa học máy tính.

như Phân Viện trưởng Phân Viện Công nghệ thông tin TP.HCM, Phó Viện trưởng Viện Cơ học ứng dụng và Tin học thuộc Viện Hàn lâm Khoa học và Công nghệ Việt Nam (VAST); Trưởng khoa Công nghệ thông tin, Trường Đại học Lạc Hồng và Trường Đại học Nguyễn Tất Thành. Hiện nay là Chủ tịch Hội đồng Khoa học và Đào tạo của Trường Đại học Ngoại ngữ - Tin học TP.HCM. Các lĩnh vực nghiên cứu quan tâm của ông bao gồm Sinh tin học, Hóa tin học, Tính toán song song và phân tán, Tính toán khoa học và Trí tuệ tính toán. Ông thành thạo nhiều ngôn ngữ lập trình như FORTRAN, C/C ++, Java, Python, Perl. Ngoài ra, ông cũng có nhiều đóng góp cho việc đào tạo nguồn nhân lực CNTT tại Việt Nam, đặc biệt là khu vực phía Nam; tham gia các hoạt động cộng đồng đặc biệt là nhiều năm ông là thành viên của Ban chỉ đạo với tư cách là Chủ tịch Ban chương trình của Hội nghị Khoa học quốc gia về nghiên cứu cơ bản và ứng dụng (FAIR). Thông tin chi tiết hơn về ông có thể được tìm thấy tại https://fair.conf.vn/~lang.