**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**



BÁO CÁO ĐỒ ÁN CUỐI KÌ

LEARNING PERSONALIZED   
STORY EVALUATION

Thành viên:

21127056 – Lâm Thiều Huy

21127577 – Trịnh Hoàng An

# MỤC LỤC

[MỤC LỤC 2](#_Toc155362884)

[1. INTRODUCTION 3](#_Toc155362885)

[2. Related work 3](#_Toc155362886)

[3. PERSONALIZED STORY EVALUATION DATASET 4](#_Toc155362887)

[3.1. CONTAMINATION IN EVALUATING EXISTING STORIES 4](#_Toc155362888)

[3.2. PERSONALIZED STORY EVALUATION DATASET 6](#_Toc155362889)

[4. PERSE: PERSONALIZED STORY EVALUATION MODEL 7](#_Toc155362890)

[4.1. Mô Hình PERSE: 7](#_Toc155362891)

[4.2. Cách Thức Hoạt Động: 8](#_Toc155362892)

[4.3. Ứng Dụng của PERSE: 8](#_Toc155362893)

[5. EXPERIENCES 9](#_Toc155362894)

[5.1 Main results 9](#_Toc155362895)

[5.2 Analysis 10](#_Toc155362896)

[5.5. Example 12](#_Toc155362897)

[6. CONCLUSION AND DISCUSSION 13](#_Toc155362898)

# INTRODUCTION

* Hiệu suất của LLMs: LLMs đã cho thấy hiệu suất ấn tượng trong các nhiệm vụ ngôn ngữ tự nhiên nhưng khả năng của chúng trong việc tạo văn bản mở cuối cùng vẫn chưa được khám phá đầy đủ do thiếu các tiêu chuẩn đánh giá đáng tin cậy.
* Thách thức trong Đánh Giá Văn Bản Mở: Đánh giá tạo văn bản mở như trả lời câu hỏi dài hạn và tạo câu chuyện là thách thức do vấn đề "một-đến-nhiều" và độ phức tạp của sự liên quan và nhất quán trong dài hạn. Các chỉ số tự động truyền thống như ROUGE có mối tương quan kém với phán đoán của con người.
* Đánh Giá Cá Nhân Hóa: Tác giả khám phá đánh giá cá nhân hóa cho việc tạo câu chuyện dài hạn, nơi mà đánh giá chịu ảnh hưởng nặng nề từ sở thích cá nhân của người đánh giá.
* Mô hình Hóa Dữ Liệu và Ưu Tiên: Để giải quyết vấn đề ô nhiễm trong các bộ dữ liệu câu chuyện hiện tại do tiếp xúc với LLMs, tác giả đã tái sử dụng hai bộ dữ liệu Per-MPST và Per-DOC cho đánh giá câu chuyện cá nhân hóa. Họ cũng phát triển mô hình đánh giá câu chuyện cá nhân hóa PERSE dựa trên LLM để nắm bắt sở thích cá nhân của một người đánh giá cụ thể từ một số ít đánh giá đã ghi chú.
* Hiệu Suất của PERSE: PERSE được đào tạo để suy luận sở thích của người đánh giá dựa trên hướng dẫn điều chỉnh và vượt trội hơn GPT-4 trong cả đánh giá câu chuyện cá nhân và đánh giá so sánh.
* Quan Sát về LLMs và Cá Nhân Hóa: LLMs hiện tại thường chung chung và thiếu cá nhân hóa ngay cả khi được cung cấp sở thích cá nhân mạnh mẽ. Tuy nhiên, với việc điều chỉnh hướng dẫn trên một số lượng lớn dữ liệu, chúng có thể hiệu quả hơn trong việc nắm bắt sở thích cá nhân và trở nên mạnh mẽ hơn khi có thêm các ví dụ cá nhân hóa.

# Related work

* **Đánh Giá Câu Chuyện Tự Động:**
  + Các Chỉ Số Tham Chiếu và Không Tham Chiếu: Có hai loại chỉ số tự động được đề xuất: tham chiếu (đánh giá sự tương đồng giữa văn bản tham chiếu và đầu ra mô hình) và không tham chiếu (đo lường chất lượng đầu ra mô hình mà không cần tham chiếu). Tuy nhiên, cả hai loại này đều có hạn chế trong đánh giá tạo văn bản mở.
  + Sử Dụng LLMs Trong Đánh Giá: Một số nghiên cứu gần đây đã khám phá việc sử dụng LLMs trong các chỉ số đánh giá. Tuy nhiên, LLMs cũng có những hạn chế như độ thiên vị về vị trí và tính dài dòng.
* **Đánh Giá Câu Chuyện Bởi Con Người:**
  + Các Khía Cạnh Đánh Giá: Đánh giá của con người được sử dụng để đánh giá các khía cạnh khác nhau của chất lượng câu chuyện như mức độ liên quan, hấp dẫn, và sự thú vị.
  + Thách Thức: Tuy nhiên, các nghiên cứu chỉ ra rằng sự đồng thuận giữa người đánh giá thấp đối với các khía cạnh chủ quan, và đánh giá crowdsourced có thể gặp vấn đề về chất lượng và tính tái sản xuất.
* **Cá Nhân Hóa Trong Xử Lý Ngôn Ngữ Tự Nhiên:**
  + Tầm Quan Trọng của Cá Nhân Hóa: Cá nhân hóa đã được nghiên cứu kỹ lưỡng trong nhiều hệ thống gợi ý và ứng dụng tìm kiếm, và gần đây cũng được chú ý trong xử lý ngôn ngữ tự nhiên.
  + Nghiên Cứu Gần Đây: Một số nghiên cứu gần đây đã đánh giá khả năng của LLMs trong việc nắm bắt sự cá nhân hóa hoặc đề xuất cá nhân hóa.

# PERSONALIZED STORY EVALUATION DATASET

## 3.1. CONTAMINATION IN EVALUATING EXISTING STORIES

**Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, số

Mô tả được tạo tự động**

* Sử Dụng Bộ Dữ Liệu IMDB: Tác giả đã sử dụng bộ dữ liệu IMDB không thương mại, bao gồm các cốt truyện phim được đánh giá bởi hàng nghìn người xem, với điểm số từ 1 đến 10.
* Kiểm Tra Sự Ô Nhiễm: Mục tiêu là điều tra cách thức mà sự tiếp xúc trước đó với LLMs (như GPT-4) ảnh hưởng đến việc đánh giá các câu chuyện. Điều này được thực hiện bằng cách xác định liệu GPT-4 có nhận biết các cốt truyện hay không.
* Tiến Hành Thí Nghiệm:
  + Tạo 200 cặp phim, mỗi cặp bao gồm hai cốt truyện với đánh giá khác nhau.
  + Yêu cầu GPT-4 xác định tiêu đề và tiến hành so sánh cặp đôi.
  + Phân loại kết quả dựa vào việc GPT-4 có nhận biết cả hai, một, hay không nhận biết cốt truyện nào.
* Kết Quả và Quan Sát:
  + GPT-4 thường nhận biết ít nhất một cốt truyện trong mỗi cặp.
  + Khi GPT-4 chỉ nhận biết một trong hai cốt truyện, nó có xu hướng chọn cốt truyện đó, dẫn đến sự nhất quán hơn trong phán đoán và giảm thiên vị vị trí.
  + Để giảm ảnh hưởng của việc ghi nhớ, tác giả đã thực hiện thay thế tên nhân vật và địa danh bằng những cái tên tương tự, làm giảm tỷ lệ cặp biết cả hai xuống 18%. Tuy nhiên, 96% cặp vẫn có ít nhất một cốt truyện được biết.
  + Sau đó, việc tóm tắt các cốt truyện đã ẩn danh giúp giảm tỷ lệ cả hai cốt truyện được biết xuống 42.5% và tăng tỷ lệ không cốt truyện nào được biết lên 23.5%.

## 3.2. PERSONALIZED STORY EVALUATION DATASET

**A white paper with black numbers and numbers

Description automatically generated**

Table 2: Statistics of Per-MPST and Per-DOC. Length is the number of words in the instruction, which includes the instruction template, reviewer preference, and plot query. I, A, S, C, and E stand for Interestingness, Adaptability, Surprise, Character Development, and Ending. k is the number of reviews; we fix k = 1 for Per-DOC due to the length.

* Bộ Dữ Liệu Per-MPST:
  + Số Lượng Người Đánh Giá: Số lượng người đánh giá giảm dần từ 1412 (k=1) đến 1336 (k=5).
  + Số Lượng Ví Dụ Đào Tạo: Tăng từ 13254 ví dụ cho k=1 lên đến 12041 ví dụ cho k=5.
  + Độ Dài Trung Bình của Hướng Dẫn: Độ dài trung bình của hướng dẫn tăng từ 868.9 từ (k=1) lên đến 2123.3 từ (k=5).
* Bộ Dữ Liệu Per-DOC (k = 1):
  + Chỉ số I, A, S, C, và E: Các chỉ số này đại diện cho các khía cạnh Interestingness (Thú vị), Adaptability (Khả năng thích ứng), Surprise (Bất ngờ), Character Development (Phát triển nhân vật), và Ending (Kết thúc).
  + Số Lượng Người Đánh Giá và Ví Dụ Kiểm Tra: Có 92 người đánh giá cho mỗi khía cạnh, với số lượng ví dụ kiểm tra và độ dài trung bình của hướng dẫn tương ứng được cung cấp.
* Mục Đích của Bảng:
  + Bảng này cung cấp một cái nhìn tổng quan về cấu trúc và quy mô của hai bộ dữ liệu.
  + Nó cho thấy sự đa dạng trong số lượng người đánh giá và số lượng ví dụ đào tạo, cũng như độ dài của hướng dẫn, phản ánh độ phức tạp và sự cụ thể của các trường hợp đánh giá cá nhân hóa.

# PERSE: PERSONALIZED STORY EVALUATION MODEL

## 4.1. Mô Hình PERSE:

* PERSE là một mô hình dựa trên LLMs (Large Language Models) được thiết kế để tạo ra đánh giá cá nhân hóa cho một người đánh giá cụ thể.
* Mô hình này dựa trên khả năng suy luận và tuân theo hướng dẫn của LLMs để suy ra sở thích ngầm từ các đánh giá trước đó của người đánh giá.
* Dưới đây là một số thông tin cụ thể về mô hình PERSE:
* Xác Định Sở Thích Cá Nhân
* Tạo Đánh Giá Cá Nhân Hóa
* Học từ Phản Hồi:
* Đánh Giá So Sánh Tinh Tế:

## 4.2. Cách Thức Hoạt Động:

* Định Nghĩa Vấn Đề: Cho một câu chuyện x và người đánh giá u, mô hình M dự đoán một đánh giá cá nhân hóa y của câu chuyện theo sở thích của người đánh giá này.
* Dự Đoán Đánh Giá: Đánh giá có thể là một điểm số số cho một cốt truyện riêng lẻ hoặc một đánh giá so sánh giữa hai cốt truyện. Nó cũng có thể bao gồm một giải thích văn bản cho quyết định.
* Xác Định Sở Thích: Sở thích được định nghĩa là lịch sử các đánh giá của người đánh giá này về các câu chuyện khác, bao gồm các cặp cốt truyện-đánh giá và mô tả văn bản của đánh giá.
* So Sánh và Phân Tích : So sánh đánh giá mới từ PERSE với đánh giá thực từ người đánh giá (nếu có) để xác định mức độ chính xác của mô hình trong việc hiểu sở thích cá nhân.

## 4.3. Ứng Dụng của PERSE:

* Mô hình PERSE được sử dụng để tạo đánh giá cho cốt truyện mới hoặc chọn ra cốt truyện ưa thích trong một cặp cốt truyện, dựa trên sở thích đã được suy luận từ các đánh giá trước đó của người đánh giá.
* Trong thực tế, PERSE có thể hỗ trợ trong việc cá nhân hóa các hệ thống gợi ý câu chuyện, cung cấp phản hồi tự động cho các nhà văn, hoặc được sử dụng trong các ứng dụng AI khác nhau để hiểu và tương tác tốt hơn với sở thích cá nhân của người dùng.

# 5. EXPERIENCES

## 5.1 Main results

A table with numbers and letters

Description automatically generated

Table 3: Pearson, Spearman and Kendall correlations of each method with human ratings on Per-MPST. We use three reviews (k = 3) to represent user preferences. All results have a p-value less than 0.05. PERSE-7b is comparable to GPT-4 and PERSE-13b significantly out performs GPT-4.

* Đánh Giá Cá Nhân Hóa:
* Mô hình PERSEind-13b (được tinh chỉnh trên LLaMA-13b-chat) có hiệu suất đáng kể, vượt trội so với tất cả các phương pháp cơ sở trong việc tương quan với đánh giá của con người đối với người đánh giá chưa biết.
* PERSEind-7b (tinh chỉnh trên LLaMA-7b-chat) có hiệu suất tương đương với GPT-4.
* PERSEind-13b đạt hệ số tương quan Pearson cao (0.345), cho thấy khả năng của mô hình trong việc nắm bắt sở thích của người đánh giá từ các đánh giá được cung cấp.
* Đánh Giá So Sánh Tinh Tế:
* Mô hình PERSEcomp được sử dụng để đánh giá so sánh tinh tế trên Per-DOC.
* Kết Quả Các Phương Pháp Khác:
* Một baseline đơn giản sử dụng điểm số trung bình từ các đánh giá trước đó làm dự đoán.
* Các mô hình LLM khác như LLaMA-2-chat (từ 7b đến 70b) và GPT-4 cũng được đánh giá với cùng các hướng dẫn và cài đặt tạo văn bản.
* Kết Quả Chính:
* Hệ số tương quan Pearson, Spearman và Kendall giữa dự đoán của PERSE và điểm số của con người cho thấy PERSEind-13b có hiệu suất vượt trội, với kết quả đáng tin cậy (p-value nhỏ hơn 0.05).
* Baseline đơn giản và LLaMA-2-7b có hiệu suất thấp hơn so với PERSE.

## 5.2 Analysis

A graph of different colored bars

Description automatically generated

* **Khả Năng Hiểu Sở Thích Cá Nhân:**
* Mô hình PERSE cho thấy khả năng hiểu sở thích của người đánh giá và sử dụng chúng để tạo điểm số cá nhân hóa.
* Điều này trái ngược với khả năng của các LLMs đã qua tiền đào tạo, vốn gặp khó khăn trong việc hiểu và sử dụng sở thích cá nhân của người đánh giá.
* **Đánh Giá So Sánh Tinh Tế:**
* PERSEcomp đạt hiệu suất tốt nhất trên tất cả các khía cạnh, bao gồm Interestingness, Adaptability, Surprise, Character Development, và Ending.
* So với phiên bản PERSEcomp-13b, phiên bản PERSEcomp-7b có hiệu suất tương đương trong khía cạnh Surprise nhưng thấp hơn trong các khía cạnh khác.
* **So Sánh với Các Baseline Khác:**
* Các mô hình LLaMA-2 chỉ đạt hiệu suất tương đương với baseline đơn giản, với độ chính xác khoảng 50% trên hầu hết các khía cạnh.
* GPT-4 thể hiện tốt hơn các baseline LLM khác trong việc nắm bắt khía cạnh Surprise, nhưng không có ưu thế trong các khía cạnh khác.
* **Kết Quả Từ Bảng 3:**
* Hiệu suất của các mô hình và baseline trên từng khía cạnh cụ thể.
* Mô hình PERSE cho thấy sự vượt trội so với các baseline khác trên tất cả các khía cạnh đánh giá.

A graph of different colored bars

Description automatically generated

* **Hiệu Suất trên Các Khía Cạnh Đánh Giá:**
* Bảng 5 cho thấy hiệu suất của PERSEcomp-7b và PERSEcomp-13b so với các baseline khác trên các khía cạnh Interestingness, Adaptability, Surprise, Character Development, và Ending.
* **Kết Quả của Các Mô Hình:**
* PERSEcomp-7b và PERSEcomp-13b:
* Cả hai phiên bản của PERSEcomp đều thể hiện hiệu suất cao trên tất cả các khía cạnh, với độ chính xác từ 0.560 đến 0.621.
* Các Baseline Khác:
* Các mô hình baseline khác, bao gồm LLaMA-2 và GPT-4, có hiệu suất thấp hơn so với PERSEcomp, với độ chính xác từ 0.466 đến 0.596 trên các khía cạnh khác nhau.
* **Nhận Xét về Hiệu Suất:**
* PERSEcomp-13b, nói riêng, thể hiện hiệu suất vượt trội so với các mô hình khác, đặc biệt trên các khía cạnh như Interestingness và Surprise.
* Điều này chứng tỏ khả năng của mô hình PERSE trong việc hiểu và phản ánh sở thích cá nhân của người đánh giá một cách chính xác.

## 5.5. Example

A screenshot of a computer

Description automatically generated

* **Hiệu Suất trên Per-MPST với Số Lượng Đánh Giá Khác Nhau (k):**
* Các thí nghiệm cho thấy mô hình PERSE có hiệu suất tương quan cao hơn với tăng số lượng đánh giá.
* Bảng 6 thể hiện hiệu suất của các mô hình với số lượng đánh giá khác nhau từ người đánh giá (k).
* **Kết Quả Các Mô Hình:**
* PERSE, đặc biệt là phiên bản PERSE-13b, có khả năng nắm bắt sở thích của người đánh giá tốt hơn khi tăng số lượng đánh giá.
* Các mô hình baseline yếu hơn, như LLaMA-2 đã qua tiền đào tạo, không thể hiện lợi ích từ việc tăng số lượng đánh giá.
* Hiệu suất của baseline đơn giản cũng giảm sau 4 đánh giá, cho thấy việc tăng số lượng đánh giá có thể làm tăng độ phức tạp của bối cảnh và giới thiệu nhiễu.
* **Ổn Định của PERSE với Sự Thay Đổi Trật Tự Các Ví Dụ:**
* PERSE, đặc biệt là phiên bản 13b, ổn định hơn so với các baseline khác trên trung bình, với sự tăng cường ổn định khi tăng số lượng đánh giá.
* Điều này cho thấy khả năng của PERSE trong việc nắm bắt sở thích ngầm của người đánh giá từ các đánh giá này.

# 6. CONCLUSION AND DISCUSSION

* **Tổng Kết Nghiên Cứu:**
* Nghiên cứu tập trung vào đánh giá cá nhân hóa của tạo văn bản mở cuối cùng.
* Phương pháp nghiên cứu bao gồm việc tạo ra các bộ dữ liệu cá nhân hóa không thiên vị từ các nguồn dữ liệu hiện có, giúp giảm thiểu vấn đề về việc mô hình đánh giá LLMs tiếp xúc với các bộ dữ liệu đã tồn tại.
* Hai bộ dữ liệu đánh giá câu chuyện, Per-MPST và Per-DOC, được tái sử dụng để đánh giá cá nhân hóa toàn diện trong các tình huống khác nhau.
* Mô hình PERSE được đề xuất cho đánh giá câu chuyện cá nhân hóa và đã cho thấy hiệu suất cao, vượt trội so với GPT-4 trong cả đánh giá chung và so sánh tinh tế giữa các cặp.
* **Hạn Chế của Nghiên Cứu:**
* PERSE sử dụng đánh giá từ người đánh giá để nắm bắt sở thích của họ, với giả định rằng sở thích này không thay đổi theo thời gian, điều này có thể không phản ánh chính xác các tình huống thực tế.
* Hạn chế về độ dài ngữ cảnh của các mô hình ngôn ngữ lớn hiện tại có thể ảnh hưởng đến việc hiểu biết toàn diện về sở thích của người đánh giá.
* **Hướng Phát Triển Tiếp Theo:**
* Nghiên cứu khuyến khích phát triển tạo câu chuyện cá nhân hóa, với mục tiêu sử dụng PERSE để cải thiện sự phù hợp giữa mô hình tạo câu chuyện và yêu cầu đặc biệt của người đánh giá.
* Mục tiêu tiếp theo là sử dụng PERSE để cải thiện sự phù hợp giữa mô hình tạo câu chuyện và yêu cầu riêng biệt của người đánh giá