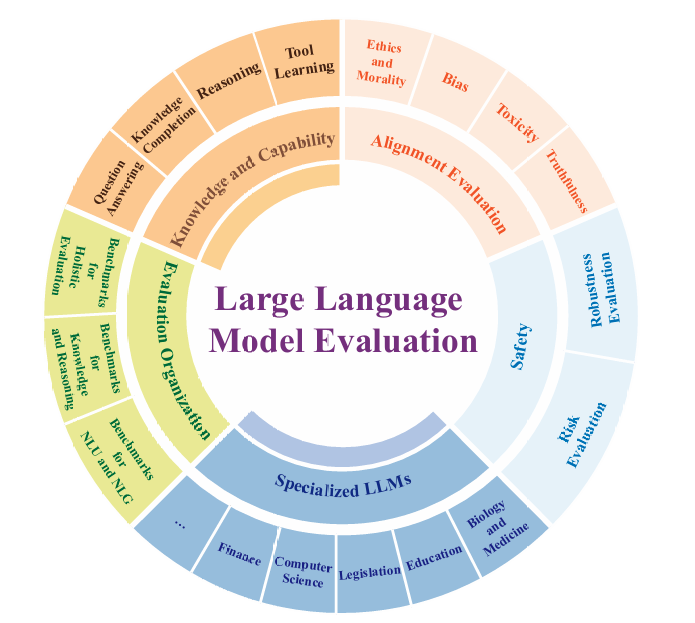
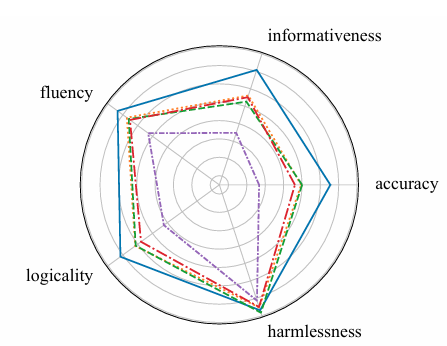
Báo cáo về tìm hiểu Seminar

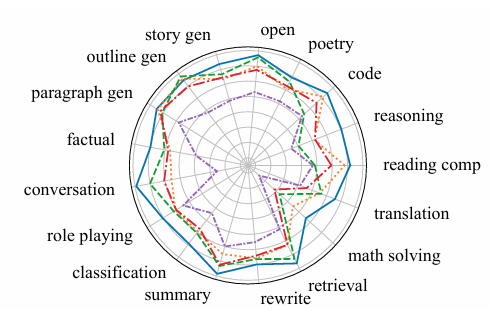
Tổng quát: Việc Đánh giá một LLM bao gồm một lượng cực lớn các task và chia thành các Domain riêng như sau, và với mỗi task phải đánh giá sẽ có kèm theo 1 dataset cũng như một benchmark riêng:



Hoặc theo các khía cạnh khái quát hơn nữa:



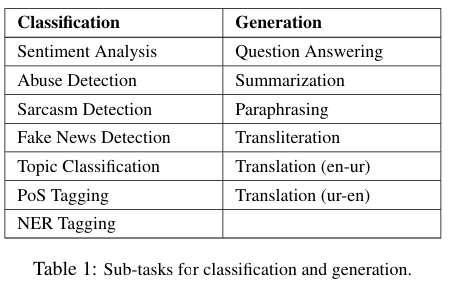
hoặc đánh giá dựa trên một loạt các task cụ thể mà hiện nay đa số mọi người quan tâm:



Tài liệu Thầy đưa:

Các phương pháp đánh giá của họ tập trung vào phần Alignment Evaluation: tức là các LLM phải đáp ứng được cán cân cân bằng trong các nhiệm vụ. Hay nói đơn giản hơn là phù hợp với đa số quan điểm của con người.

Ngoài ra, các bài đánh giá thường dựa trên 1 loại ngôn ngữ, hoặc một khía cạnh nhất định của Hình tròn Đánh giá LLM.



Và họ chia ra thành 2 nhóm model: đã được tinh chỉnh và được huấn luyện tổng quát?

Khái quát được ccasc bước để đánh giá một LLM:

Lựa chọn Khía cạnh muốn đánh giá -> Lựa chọn kiểu models -> lựa chọn/tạo dataset và benchmark

Lựa chọn cách tiếng hành đánh giá: zero shot, 3 shot, COT etc (liên quan đến việc khai thác tối đa kết quả thông qua promting)

**Cấu trúc bài Seminar:**

**Outline : Tiêu đề cần có.**

**I.Khái niệm LLM**

* **Giới thiệu về LLM (Language Models):**
  + **Khái niệm cơ bản về mô hình ngôn ngữ.**
  + **Sự phát triển từ các mô hình ngôn ngữ truyền thống đến các mô hình ngôn ngữ lớn (LLM) như GPT, BERT, T5, v.v.**
  + **Vai trò của LLM trong xử lý ngôn ngữ tự nhiên (NLP).**
* **Ưu và nhược điểm của LLM:**
  + **Khả năng xử lý văn bản đa dạng, tạo văn bản tự nhiên.**
  + **Hạn chế về tài nguyên tính toán và những thách thức trong việc kiểm soát nội dung.**

**II.Đánh giá LLM**

* **Cái nhìn tổng quan về việc đánh giá LLM ( Tập trung chủ yếu kiến thức từ trong Nghiên cứu** [**masterthesis/documents/references/Evaluating LLM\_ComprehensiveSurvey\_2310.19736v3.pdf at main · quangminh1512/masterthesis (github.com)**](https://github.com/quangminh1512/masterthesis/blob/main/documents/references/Evaluating%20LLM_ComprehensiveSurvey_2310.19736v3.pdf)**)**
* **Ứng dụng và tiến bộ trong đánh giá LLM:**
  + **Các ứng dụng phổ biến của LLM trong thực tế (dịch máy, trợ lý ảo, phân loại văn bản, tóm tắt, v.v.).**
  + **Tiến bộ gần đây trong việc đánh giá chất lượng và hiệu quả của LLM.**

**III.Đánh giá LLM đối với 1 ngôn ngữ riêng biệt (urdu) - Một loại ngôn ngữ hiếm**

**Tác giả đã thực hiệ các bước sau để đánh giá:**

**1. Dữ liệu kiểm tra được kiểm soát:**

**Dữ liệu cân bằng: Tạo ra bộ kiểm tra cho hai nhiệm vụ Phân loại và Generation với số lượng ngang nhau**

**2. Số liệu đánh giá:**

* **Nhiệm vụ phân loại: Họ đã sử dụng Điểm Macro-F1 để đo hiệu suất tổng thể của các mô hình trên nhiều nhiệm vụ phân loại khác nhau.**
* **Generation: Họ sử dụng các số liệu khác nhau dựa trên nhiệm vụ cụ thể:**
* **Trả lời câu hỏi: Điểm SQUAD F1**
* **Tóm tắt: Điểm ROUGE-L**
* **Diễn giải, Chuyển ngữ và Biên dịch: Điểm SacreBLEU**

**3. Prompt Engineering (Cách thức thực hiện):**

* **Mô hình tổng quát (GPT-4, Llama): Họ sử dụng zero-shot, few-shot (3-shot and 6-shot), và chain-of-thought (CoT) prompting để hiểu cách các mô hình này thích ứng với nhiệm vụ với các mức độ ngữ cảnh khác nhau.**
* **Mô hình Fine-tuned (Llama-FT, XLM-R, mT5): Các mô hình này được tinh chỉnh cụ thể cho từng nhiệm vụ, do đó chúng chỉ được đánh giá mà không cần có kỹ thuật promting.**

**4. Đánh giá của con người (Chủ quan):**

* **Với các nhiệm vụ cụ thể: Họ tập trung vào đánh giá của con người đối với các nhiệm vụ chủ quan nhất: tóm tắt, diễn giải, chuyển ngữ và cả hai hướng dịch tiếng Urdu-Anh.**
* **Xếp hạng: Người chú thích là con người (người bản ngữ nói tiếng Urdu) xếp hạng đầu ra từ các mô hình khác nhau (GPT-4, Llama, Llama-FT, mT5) dựa trên các tiêu chí cụ thể. Tiêu chí xếp hạng: Mỗi người chú thích được cung cấp một bộ tiêu chí (được nêu chi tiết trong Phụ lục G của bài báo) để đánh giá đầu ra. Các tiêu chí này có thể bao gồm các khía cạnh như:**
  + **Độ chính xác: Đầu ra phản ánh ý nghĩa của văn bản gốc tốt như thế nào?**
  + **Độ trôi chảy và tự nhiên: Đầu ra có đúng ngữ pháp và dễ đọc bằng tiếng Urdu không?**
  + **Tính liên quan: Đầu ra có duy trì thông tin và phong cách thiết yếu của văn bản gốc không?**

**Hệ thống xếp hạng: Người chú thích xếp hạng đầu ra của các mô hình khác nhau cho mỗi nhiệm vụ theo thang điểm từ 1 đến 4, trong đó 1 biểu thị đầu ra tốt nhất và 4 biểu thị đầu ra kém nhất. Họ được phép xếp hạng nhiều mô hình cùng một thứ hạng nếu chúng được đánh giá là có chất lượng như nhau.**

* **Độ tin cậy giữa những người đánh giá: Sử dụng Krippendorff's alpha để đo tính nhất quán của thứ hạng chú thích là con người. Điều này giúp hiểu được mức độ đồng thuận giữa các đánh giá chủ quan của con người.**

**5. Đánh giá dựa trên LLM:**

**GPT-4 và Llama: Các nhà nghiên cứu cũng yêu cầu GPT-4 và Llama xếp hạng đầu ra của các mô hình khác, sử dụng cùng tiêu chí như khi người xếp hạng là con người.**

**So sánh: Họ so sánh thứ hạng do LLM tạo ra với thứ hạng của con người để xem LLM có thể phù hợp với đánh giá của con người tốt như thế nào.**

**4.Đánh giá LLM tiếng việt với 1 miền giáo dục. Chỉ cần đặc trưng và hữu ích nhất 2 đến 3 task.**

**Các nhiệm vụ Macro như (Chủ yếu để kiểm tra kiến thức nội hàm của LLM):**

**Hỏi đáp, diễn giải trên các vùng tri thức trong Domain giáo dục:**

* **Hệ thống Giáo dục**
* **Phương pháp Giảng dạy**
* **Chương trình Giáo dục**
* **Đánh giá và Kiểm tra trong hệ thống Giáo dục**
* **Giáo dục Đặc biệt**
* **Giáo dục các bậc từ mầm non tới Đại học**
* **Giáo dục Quốc tế**
* **Chính sách và Quản lý Giáo dục**
* **Giáo dục và Công nghệ**

**Đánh giá với các vi phạm về sắc tộc, văn hóa, ngôn ngữ.**

**Các nhiệm vụ micro như (chủ yếu để kiểm tra khả năng thích nghi với các cấp độ context phức tạp hơn):**

**Tóm tắt, phân tích, giải thích với các promting đầu vào của người dùng, bao gồm cả việc cung cấp tri thức ngoài để hỏi. Có thể tập trung chủ yếu và các phân vùng kiến thức cụ thể hơn như ngữ văn, toán, lý etc**

**Với toàn bộ các nhiệm vụ kể trên, đều được đánh giá thông qua Con người và LLM như trong phần III.4 và III.5 đã mô tả.**