LLM for Explainable AI

Ahsan Bilal and Beiyu Lin

October 3, 2024

LLM for Explainable AI

Abstract – Explainability for large language models (LLMs) is a critical area in natural language processing because it helps users understand the models and makes it easier to analyze errors, especially as these models are widely used in various applications. The "black-box" nature of AI models raises challenges in transparency and ethics because we cannot see or understand how the model processes information to generate its output. Traditional methods, such as attention mechanisms, have enhanced explainability in AI models by improving model focus and accuracy but at the cost of increased complexity. More specifically, they use tools like gradient-based methods (e.g., Grad-CAM), making them less accessible to non-expert users. We employ in-context learning and prompt refinement techniques, focusing on the pre-trained Transformer-based large language model BART. This approach simplifies model interaction by allowing users to guide the model through natural language prompts, reducing the need for technical expertise. We validate this method via a real-life StudentLife dataset collected 48 college students over 10 weeks. Our results offer the possibility of using LLMs for XAI to achieve data mining for everyone.

I INTRODUCTION

In recent years, the demand for Explainable Artificial Intelligence (XAI) has grown significantly due to concerns about the opacity and lack of interpretability in complex machine learning models [1] and [2]. However, domain experts without experience in machine learning (ML) often find it difficult to understand and use many ML explanations, as they are typically presented in formats that are not intuitive or easily readable by humans as in [3], [4] and [5]. The motivation for XAI stems from the need to uncover what is happening inside advanced ML models, which are often treated as black boxes [6], and to make this information understandable for users. Large Language Models (LLMs), which have demonstrated their usefulness across various domains [7], presents a promising approach to advancing the field of Explainable AI (XAI).

Non-experts, while often aware of the questions they need to ask to obtain information in a consumable format, may struggle to understand the technical terms and intricacies of these AI models [8], [9]. For instance, traditional methods that require users to interact with ML models using SQL can be particularly challenging for non-technical users, making these methods less user-friendly [10].

Recent research has explored using machine learning models to interpret user queries and provide suitable explanations [11], [12]. However, these approaches still pose challenges for non-expert users, because they often require an understanding of technical terms and model-specific nuances. Moreover, the key difficulty lies not only in interpreting model behavior but also in conveying the explanations in a way that is easily consumable and useful for the user. In this paper, we improve explainability using LLMs through straightforward user prompts to extract and present information in a format that is easily consumable by users.

II LITERATURE REVIEW

The development of large language models (LLMs) has advanced rapidly, with notable examples including GPT-3.5, which enhances search relevance in Microsoft's Bing [13] and Google's Bard [14]. Notable models like Google's BERT [15] and OpenAI's GPT series have set new benchmarks in NLP tasks, for example to generate specific responses upon the input text, including text generation [16]. ChatGPT, OpenAI's latest model, effectively translates complex outputs into user-friendly language, aiding learners and advisors.

As machine learning models, such as neural networks, are increasingly used in areas where human judgment is required, it becomes crucial to explain how these models generate their outcomes [17]. This growing need for explainability and transparency in output responses has led to the development of Explainable AI (XAI), which aims to make complex models easier to understand by showing how they make decisions and offering clear, simple explanations [18]. While interpretability focuses on understanding how the entire model functions, explainability focuses on why the model gives certain results.

Researchers have employed various techniques to enhance explainability, such as feature importance methods , which identify the features that significantly impact specific responses [19]. Another example is LIME (Local Interpretable Model-agnostic Explanations), which explains a model's predictions by creating a simpler model

Table 1: Sample entries for social responses, study spaces, and class information data generated by Python script

Social Responses				
Person ID	**Location**	**Number**	**Time (Eastern)**	**Time (Unix)**
u01	43.7067925,-72.28917303	4	2013-04-16	1366092021
Study Spaces				
Person ID	**Location**	**Noise**	**Time (Eastern)**	**Place**
u30	43.7024991,-72.28938342	1	2013-05-03	Paddock library
Class Information				
Course 1	**Course 2**	**Common Week Days**	-	-
ANTH 012	COSC 089 1	[1,2,3,5]	-	-

around a specific prediction. This helps show which features are influencing that decision. More specifically, If the prompt is 'How does stress affect student GPA?', the important features would be 'stress' and 'GPA'. These techniques help in explaining how LLMs arrive at their predictions, making their outputs more understandable and actionable.

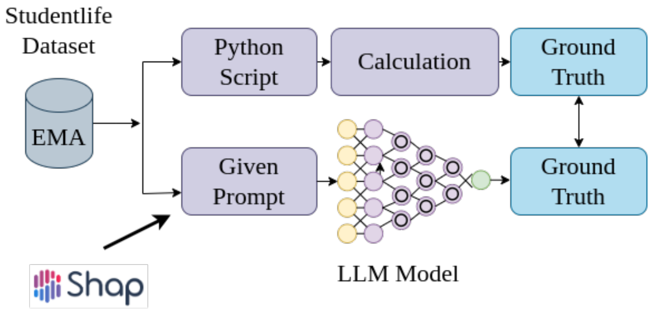

Figure 1: Comparison between the ground truth Python script written by us (Figure I) and the result obtained using the refined specific prompt (Table III). The prompt refinement process focuses on task-specific attention, calculated using the SHAP [20] tool, with the attention mask for the three categories shown in Figure 2. Due to the large dataset size and complexity of ChatGPT's architecture, the BART model [21] was employed alongside SHAP to enhance prompt explainability.

III METHODOLOGY

To enhance AI explainability using LLM, we apply the“locate-then-edi” approach [22] (originally designed for modifying weights in classification tasks), to simplify explainability with LLM using straightforward user prompts. Unlike classification tasks, our focus is on improving the explainability of the model. Prompt engineering, in this context, involves designing and refining the input or question given to a language model to guide its response effectively. A “prompt” refers to the input or question given to a language model that guides its response. For example, asking “What is machine learning?” directs the model to provide an explanation of that topic. Our method identifies critical words within the input prompt which are called prompt features(the “locate” step) using shapley additive explanations (SHAP) [20] and then refines the prompt to focus on these features (the “edit” step). We choose SHAP because it measures feature importance by calculating each feature's contribution to a model's prediction using Shapley values from cooperative game theory. Unlike other methods, SHAP shows how each feature contributes to the model's prediction. For example, in a model predicting health outcomes, both age and blood pressure might be important together. SHAP not only shows that these features matter but also explains how they work together to affect the prediction. By emphasizing relevant aspects in the input, we aim to enhance the accuracy and clarity of the explanations provided by the LLM. We evaluated this approach by comparing the explanatory table of data generated in the LLM response III with a ground truth table created using a python script. This showed how a single prompt could produce clear, user-friendly explanations, making the system more intuitive for non-experts and presenting information in a format that was easily consumable. In doing so, we improved the explainability of the AI model using LLM.

IV EXPERIMENTS

In this preliminary study, we use SHAP with large language models (LLMs) to map prompt features. We then refine the prompts to generate explanations in the form of visualization tables for the Experience Sampling Method (EMA) section of the dataset. We focused on the social, stress, and study space categories in the EMA section of the studentlife dataset [23].

We have done experiments with three key tasks. First, it extracts entries with locations containing the word “library,” enabling users to easily identify available study spaces across different libraries on campus. This feature

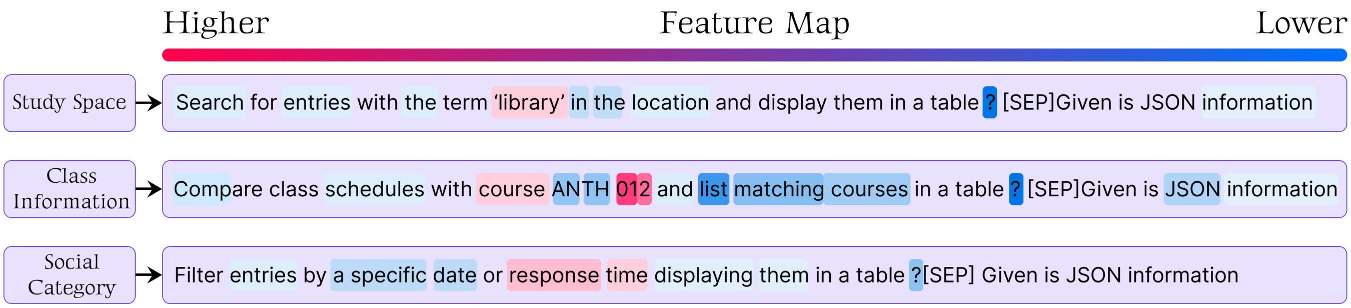
Higher
Feature Map
Lower
Study Space
Search for entries with the term 'library' in the location and display them in a table ? [SEP]Given is JSON information
Class Information
Compare class schedules with course ANTH 012 and list matching courses in a table ? [SEP]Given is JSON information
Social Category
Filter entries by a specific date or response time displaying them in a table ?[SEP] Given is JSON information

Figure 2: Feature map of prompts for specific outputs

Social				
ID	**Location**	**Number**	**Time (Unix)**	**Time (Eastern)**
U01	43.70,-72.289	4	1366092021	2013-04-16
U02	43.70,-72.289	2	1366092013	2013-04-16
U04	43.70,-72.289	2	1366092013	2013-04-16
Study Space				
ID	**Location**	**Noise**	**Place**	**Productivity**
u30	43.70,-72.289	1	Paddock library	4
u45	43.71,-72.30	1	Dona library	2
u56	Unknown	2	Library	2
Class Information				
Course 1	**Course 2**	**Common Week Days**		
ANTH 012	COSC 089	[1,2,3,5]		
ANTH 012	EARS 003	[1,2,3,5]		
ANTH 012	ENGL 047	[1,2,3,5]		
ANTH 012	JAPN 033	[1,2,3,5]		
ANTH 012	MATH 022	[1,2,3,5]		
ANTH 012	MATH 023	[1,2,3,5]		
ANTH 012	NAS 035	[1,2,3,5]		

Table 2: Data table for social, study space, and class information generated by LLM

helps students plan their study sessions based on location and availability. Second, it matches schedules for courses like "ANTH 012" and other selected courses, allowing students to consolidate their class timings in one place, helping them avoid scheduling conflicts. In the final task, the system filters entries by response time, ensuring that users receive the most timely and relevant information, particularly useful for quick responses in time-sensitive situations such as social events or announcements.

V RESULTS

The initial understanding of our work is illustrated in Figure 1. We compared the ground truth, represented by our Python script I, with the results from a refined specific prompt III. This refinement, guided by the SHAP tool, focused importance on key aspects of the task. Highlighting terms like "library", "course," and "response time" demonstrated basic explainability using LLM. By providing a more transparent view of how specific terms or features influence the model's output. Figure 2 shows the prompt's feature map across three categories. These results, including the table generated using LLMs to convey data, suggest that leveraging explainability with LLMs can effectively convey information in a way that is easily consumable, even for non-expert users. In summary, we have explored the potential of LLMs to enhance model explainability and generate clear, explainable outputs, laying the groundwork for future advancements in this field.

VI FUTURE DIRECTIONS

Future improvements could involve LLM training based on user feedback (i.e., LLMs continuously fine-tuning based on user interactions and feedback in real-time), using in-context learning, and the locate-and-edit technique in parallel to enhance explainability in ML model.

However, the current study has certain limitations, such as reliance on predefined prompts and the limited interpretability of complex model outputs. Additionally, the accuracy of LLM-generated explanations can vary depending on the quality of input data and the effectiveness of feature importance techniques. To overcome these limitations, we can improve by continuously refining prompts to better-fit user needs. This refinement will help convey information in a clearer and more relevant way for users in the context of XAI.

References

[1] ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, v. 58, p. 82–115, 2020.

[2] BHATT, U. et al. Explainable machine learning in deployment. In: ASSOCIATION FOR COMPUTING

MACHINERY. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York, NY, USA, 2020. p. 648–657.

[3] JIANG, H.; SENGE, E. On two xai cultures: A case study of non-technical explanations in deployed ai systems. In: *Human Centered AI (HCAI) workshop at NeurIPS 2021*. [s.n.], 2021. Disponível em: ⟨http://arxiv.org/abs/2112.01016⟩.

[4] NYRE-YU, M. et al. Considerations for deploying xai tools in the wild: Lessons learned from xai deployment in a cybersecurity operations setting. In: US DOE. *ACM SIG Knowledge Discovery and Data Mining Workshop on Responsible AI*. Singapore, Singapore, 2021.

[5] YANG, W. et al. Survey on explainable ai: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, v. 3, n. 3, p. 161–188, 2023.

[6] ZAHAVY, T.; BEN-ZRIHEM, N.; MANNOR, S. Graying the black box: Understanding dqns. In: PMLR. *International conference on machine learning*. [S.l.], 2016. p. 1899–1908.

[7] CHANG, Y. et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[8] ZHAO, H. et al. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023. Disponível em: ⟨https://arxiv.org/abs/2309.01029⟩.

[9] SHEN, L. et al. Towards natural language interfaces for data visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[10] WU, X. et al. Usable xai: 10 strategies towards exploiting explainability in the llm era. *Preprint*, March 2024.

[11] SHEN, H. et al. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. In: *Computer Supported Cooperative Work and Social Computing*. [S.l.: s.n.], 2023. p. 384–387.

[12] SLACK, D. et al. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, v. 5, n. 8, p. 873–883, 2023.

[13] MEHDI, Y. *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web, 2023.* 2023. Disponível em: ⟨https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.⟩

[14] THOPPILAN, R. et al. Lamda: Language models for dialog applications. *arXiv preprint*, arXiv:2201.08239, 2022. Disponível em: ⟨https://arxiv.org/abs/2201.08239⟩.

[15] DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805, 2018. Disponível em: ⟨https://arxiv.org/abs/1810.04805⟩.

[16] BROWN, T. B. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020. v. 33, p. 1877–1901. Disponível em: ⟨https://arxiv.org/abs/2005.14165⟩.

[17] SUSNJAK, T. A prescriptive learning analytics framework: Beyond predictive modelling and onto explainable ai with prescriptive analytics and chatgpt. *arXiv preprint arXiv:2208.14582*, 2023. Revision of the original paper to include ChatGPT integration. Disponível em: ⟨https://doi.org/10.48550/arXiv.2208.14582⟩.

[18] MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. Interpretable machine learning – a brief history, state-of-the-art and challenges. *arXiv preprint arXiv:2010.09337*, 2020. Disponível em: ⟨https://doi.org/10.48550/arXiv.2010.09337⟩.

[19] ALFEO, A. L. et al. From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine*, v. 236, p. 107550, 2023. ISSN 0169-2607. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0169260723002158⟩.

[20] LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017. To appear in NIPS 2017. Disponível em: ⟨https://arxiv.org/abs/1705.07874⟩.

[21] LEWIS, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. ArXiv:1910.13461 [cs.CL]. Disponível em: ⟨https://doi.org/10.48550/arXiv.1910.13461⟩.

[22] FEIGENBAUM, I. et al. Editing arbitrary propositions in llms without subject labels. *arXiv preprint arXiv:2401.07526*, 2024. ArXiv:2401.07526v1

[cs.CL]. Disponível em: ⟨https://arxiv.org/abs/2401.07526⟩.

[23] WANG, R. et al. Studentlife: Using smartphones to assess mental health and academic performance of college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, v. 1, n. 1, p. 1–26, 2017. Disponível em: ⟨https://www.cs.dartmouth.edu/~xia/papers/mobilehealth17.pdf⟩.