2024 11th International Conference on Advances in Computing and Communications (ICACC) | 979-8-3503-7913-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACC63692.2024.10845370

Exploring Explainable AI in Large Language Models: Enhancing Transparency and Trust

Ijas A. H
Dept. of Computer Science & Engineering, FACTS-H Lab
Indian Institute of Information Technology Kottayam, India
ijas.emoa222@iiitkottayam.ac.in

Ashly Ann Jo
Dept. of Computer Science & Engineering, FACTS-H Lab
Indian Institute of Information Technology Kottayam, India
ashlyannjo.phd2112@iiitkottayam.ac.in

Ebin Deni Raj
Dept. of Computer Science & Engineering, FACTS-H Lab
Indian Institute of Information Technology Kottayam, India
ebindeniraj@iiitkottayam.ac.in

***Abstract*—Large Language Models (LLMs) are at the forefront of technological evolution, significantly enhancing digital interactions and automating complex processes across various sectors. While these models facilitate advancements in data analytics, content generation, and strategic decision-making, their opaque nature poses challenges regarding user trust and model comprehension. Addressing the critical need for transparency, this paper focuses on enhancing LLM explainability. We propose a framework to demystify the internal mechanisms of these models, facilitating a deeper understanding that aligns with stringent ethical standards. Our approach integrates advanced explanatory tools that elucidate model decisions and foster accountability and fairness in Artificial Intelligence applications. Through rigorous analysis and the development of novel interpretative methodologies, we aim to bridge the gap between LLM capabilities and ethical AI practices, ensuring that these powerful tools are leveraged responsibly and transparently in critical infrastructures.**

***Index Terms*—Large Language Models (LLMs), Explainable AI, Transparency, Ethical AI, Model Interpretability, Accountability, Fairness, AI Applications**

I. Introduction

Large Language Models (LLMs) are increasingly central to technological advancements in numerous fields, dramatically enhancing how we interact with digital platforms by automating complex tasks with remarkable efficiency. These models leverage vast datasets to drive innovations, from improving customer service through automation to generating insightful data analyses. However, the internal mechanisms of LLMs remain largely obscure, creating significant challenges in environments that demand transparency, such as healthcare, finance, and public services. The ”black box” nature of these models not only complicates their ethical deployment but also hinders effective error diagnosis and bias mitigation, challenging the trust and reliability placed in them by users and regulatory bodies. Addressing these issues, our research focuses on unraveling the complexities of LLMs by advancing explainability techniques that illuminate their operational processes.

In this paper, we explore a variety of established and emerging methodologies designed to enhance the transparency of LLMs. We investigate gradient-based and perturbation methods alongside interpretability frameworks such as SHAP (SHapley Additive exPlanations) [1] and LIME (Local Interpretable Model-agnostic Explanations) [2], assessing their capacity to decode the reasoning behind model outputs. Through meticulous experimentation and comparative analysis, this study evaluates the effectiveness of various explainability approaches across different LLM architectures, employing rigorous metrics that quantify aspects of clarity, fairness, and transparency. Our research endeavors to contribute valuable insights to the ongoing discussions on ethical AI, aiming to equip developers, policymakers, and end-users with robust tools for better understanding and managing LLMs. By bridging the gap between cutting-edge AI capabilities and the high standards of accountability required for their application, this work fosters a foundation for future AI systems that are not only powerful but also aligned with societal values and ethical norms.

II. Literature Survey

Our literature survey focuses on two key aspects of LLM research: the exploration of LLM capabilities and the advancement of explainability within these models. Each of these areas is crucial for understanding the current landscape and directing future developments in LLM technology.

A. Exploration of LLM Capabilities

Recent advancements in LLMs have significantly extended their utility in natural language processing, evident in their increasing integration across industries for tasks such as text generation, translation, and code completion. This section critically examines the datasets, methodologies, and performance evaluations of state-of-the-art models to provide insights into their operational strengths and prevalent challenges.

979-8-3503-7913-6/24/$31.00

1) Training Data Diversity and Sources: The capability of LLMs to emulate human-like text generation relies significantly on the diversity and breadth of their training datasets. Models like LLAMA [3] and BLOOM [4] are built upon large-scale datasets that encompass a wide range of internet-sourced content, including CommonCrawl, GitHub, and Wikipedia. This extensive data coverage fosters a generalized language understanding but also introduces data quality and bias challenges. In contrast, ERNIE [5] integrates structured knowledge graphs with textual data, enriching its language comprehension with relational data insights. Additionally, CODE LLAMA [6] specifically targets programming languages, optimizing its capabilities for tasks such as code synthesis and bug fixing.

2) Methodological Approaches: The training methodologies of LLMs vary significantly, influencing their learning efficiency and task performance. For instance, techniques like gradient clipping and weight decay are pivotal in LLAMA's training regimen to stabilize learning and mitigate overfitting. BLOOM adopts multitask finetuning and prompt alignment to enhance versatility across different tasks. Furthermore, cutting-edge computational strategies such as distributed training and 3D parallelism are essential for managing FALCON's extensive parameters and large-scale data processing.

3) Performance Metrics and Evaluation: Evaluating LLMs involves a diverse array of metrics that assess their performance across several dimensions such as accuracy, efficiency, reliability, and their ability to handle specific tasks. These metrics are crucial for determining the practical utility and effectiveness of LLMs in various applications. Models like LLAMA and FALCON [7] are notable for their zero-shot and few-shot learning capabilities, which measure their ability to adapt to and perform on new tasks with minimal additional inputs. Metrics such as BLEU [8], ROUGE-L [9], and COMET are utilized to assess language-specific performance, providing insights into linguistic accuracy and text generation quality.

Despite their advancements, LLMs encounter several limitations. Data quality issues, including the presence of biased or inappropriate content, are significant concerns for models like LLAMA, directly impacting output quality. Specific challenges like FALCON's limited multilingual capabilities and GPT-NeoX-20B's [10] data duplication issues underscore the difficulties in developing universally robust models. Moreover, a critical hurdle across all surveyed models is the lack of explainability, which impedes error diagnosis and model behavior understanding—key for establishing trust and transparency in AI applications.

B. Explainability in Large Language Models

Several pivotal studies have laid the groundwork for methodologies and frameworks to enhance our understanding of LLMs [11].

1) Local Interpretable Model-agnostic Explanations(LIME): LIME is a seminal work in this area, offering a technique to approximate the predictions of black-box models with interpretable models, focused locally around the prediction point [12].

2) Attention Mechanisms: Initially proposed to improve model performance, attention weights have been studied for their potential to reveal which parts of the input data most influence model outputs. However, over-reliance on attention weights as definitive indicators of model rationale is cautioned against, suggesting they may not always correspond to model reasoning.

3) SHAP (SHapley Additive exPlanations): SHAP utilizes cooperative game theory to distribute "payouts" (i.e., contributions to the output) across input features, providing a granular understanding of feature contributions. SHAP quantifies the contribution of each feature across a model's decision, offering a more detailed explanation compared to methods that provide binary importance mappings [1].

4) Challenges and Gaps: In the field of explainability in LLMs, several challenges and gaps persist that hinder a deeper understanding and broader application of these technologies. Scalability and computational efficiency of explainability methods often struggle to keep pace with the increasing complexity and size of modern LLMs. Additionally, there is no established "ground truth" for explanations, making it difficult to evaluate the accuracy and quality of interpretative outputs. Potential biases within the explanations themselves, arising from underlying data or model biases, can lead to misleading or harmful interpretations. Koh and Liang (2017) introduced influence functions to trace a model's prediction back to its training data, offering another layer of transparency in black-box models [13]. These challenges highlight the need for ongoing research to develop more robust, scalable, and unbiased explainability methods, as well as standardized benchmarks and metrics for their evaluation.

III. Methodology

The proposed methodology is designed to enhance the explainability of Large Language Models (LLMs) while maintaining robust performance across a variety of tasks. This multi-stage approach involves selecting appropriate foundational models, applying advanced interpretability techniques, and optimizing fine-tuning processes to balance transparency with performance.

A. Selection and Setup of Foundational Models with explainability

Given the computational constraints, including limited GPU resources, we focus on models with up to 7 billion parameters with 4 bit quantization [14] for feasibility and efficiency. The selected foundational models for initial experiments are:

- Llama 7b - 4 bit quantization applied
- Mistral 7b - 4 bit quantization applied
- Gemma 7b - 4 bit quantization applied
- Falcon 7b - 4 bit quantization applied
- Bloom 7b - 4 bit quantization applied
- Mpt 7b - 4 bit quantization applied

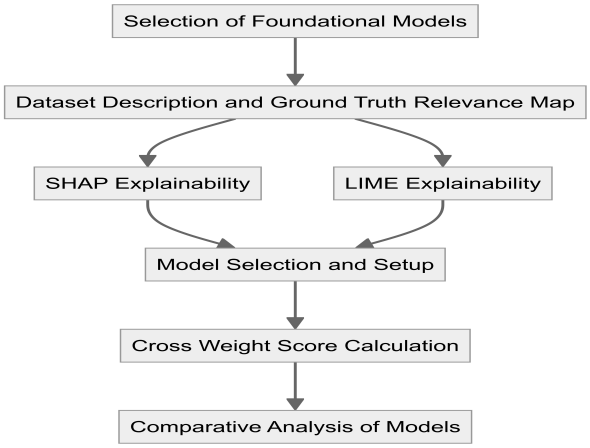

Fig. 1. Overview of Proposed Methodology

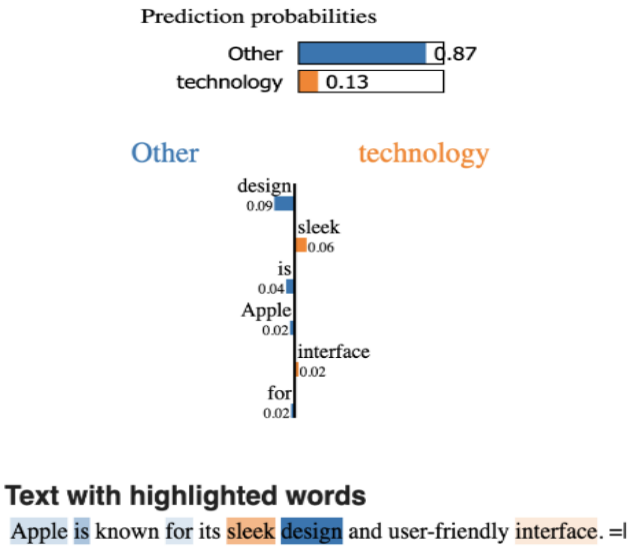

Fig. 2. LIME prediction visualization

B. LIME Explainability

Local Interpretable Model-Agnostic Explanations (LIME) is employed to demystify the decision-making processes of various foundational Large Language Models (LLMs), enhancing their transparency and explicability.

1) Model Selection and Setup: A range of LLMs with diverse architectural properties were chosen to evaluate the generalizability of LIME: **Mistral 7b, Gemma 7b, Falcon 7b, Bloom 7b, Mpt 7b**: These models were selected based on their varied capabilities to ensure a broad understanding of LIME's applicability.

These models are operationalized on infrastructure equipped with 16GB RAM and an Nvidia T4 GPU, optimized for demanding computational tasks.

2) Dataset Description and Ground Truth Relevance Map: To effectively assess the explainability mechanisms, we crafted a custom classification dataset designed to challenge the models with contextually rich examples. The dataset aims to test the model's ability to interpret nuanced semantic differences, using contextually variant uses of words like "Apple" to signify different entities (technology company vs. fruit).

Creating the Ground Truth Relevance Map: This map is crucial for comparing model outputs with expected interpretations. It is constructed by initially running the dataset through a high-capacity LLM (Mixtral 8x7B in this case) to obtain preliminary prediction outputs. For each instance, expert annotators review the model's attention weights and outputs to determine the relevance of specific words or phrases to the final decision. This relevance is then quantified on a scale (e.g., 0 to 1) to establish a 'ground truth' for each input feature, reflecting its importance in the decision-making process of an ideally performing model. This relevance map serves as a benchmark for evaluating the accuracy of LIME's interpretability outputs.

3) Implementation of LIME: LIME is applied to each model to elucidate how input variations influence predictions

- Preprocessing and Relevance Mapping: Before LIME's application, initial preprocessing involves creating a word-wise relevance map using the high-capacity model. This identifies key features that are hypothesized to be influential.
- Applying LIME: Each model processes the classification task using the custom dataset, with logits used to determine prediction probabilities. LIME is then applied to these outputs to generate interpretability maps, highlighting the influence of relevant features.
- Analyzing LIME Outputs: LIME's outputs include both positively and negatively correlated features, offering detailed insights into feature influence on predictions. Comparing these with the ground truth relevance map allows for the evaluation of LIME's accuracy and the identification of any biases or errors in the model's processing.
- Mathematical Representation and Evaluation: The effectiveness of LIME in reflecting the true feature influence is quantitatively assessed using the Cross weight(w) score:

$$\text{Cross weight}(w) = \sum \text{gt(word)} \times \text{p(word)}$$

where gt(word) is the relevance score from the ground truth map, and p(word) is the relevance score derived from LIME. Unconsidered words in the ground truth are assigned a nominal relevance score to ensure consistency in the evaluation.

C. SHAP Interpretability

SHAP is a powerful interpretability tool used to explain the output of machine learning models, including Large Language Models (LLMs). SHAP leverages the concept of Shapley values from cooperative game theory to attribute the contribution of each input feature to the final prediction, providing both local and global insights into model behavior. This section details the methodology for implementing SHAP to enhance

the transparency and understanding of several foundational LLMs.

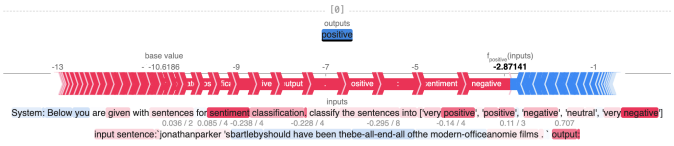

Fig. 3. SHAP Prediction visualization

1) Model Selection and Setup: To explore SHAP's applicability and effectiveness, the following LLMs have been selected:

Mistral 7b, Gemma 7b, Falcon 7b, Bloom 7b, Llama 2 7b: These models represent a range of architectures and capabilities, providing a diverse set for analysis. These models are set up on a computational infrastructure equipped with 16GB RAM and an Nvidia T4 GPU, optimizing the balance between computational efficiency and capability.

2) Dataset Description: For this experiment, we shift from the custom dataset used in previous LIME experiments to a benchmark dataset, SST-5, which is a multiclass sentiment analysis classification dataset. This dataset was chosen to ensure unbiased data inputs and to facilitate robust evaluation across different sentiment contexts. A total of 100 samples have been selected randomly across five categories, with 20 samples from each category, considering the constraints of high computational demand and resource limitations.

3) Implementation of SHAP: SHAP is applied to each LLM under consideration as follows:

- Task Setup: Each model is tasked with a classification challenge using the SST-5 dataset.
- Using Mixtral 8x7b for Ground Truth Generation: To create a reliable benchmark for SHAP's explanations, the Mixtral 8x7b model, a higher-capacity LLM, processes the same SST-5 dataset. This model's outputs are analyzed to determine the impact of each word on sentiment classification, creating a preliminary relevance map.
- Teacher Forcing Technique: This technique is employed to use the model's logits for generating predictions, ensuring that the input to SHAP is the output directly from the model, which captures the probability distribution over possible outputs.
- Applying SHAP: For each prediction, SHAP values are computed to determine the contribution of each input feature (word) to the prediction. This involves:
  - Using the Input-Target Pair: The input sentence and its corresponding target class are used to focus SHAP's explanation on the specific decision-making process for that classification.
- Weight Calculation:

$$\text{Cross weight(w)} = \sum \text{gt(word)} \times \text{p(word)}$$

Here, gt(word) is the relevance score from the ground truth map generated using Mixtral 8x7b, and p(word) is the relevance predicted by SHAP. This score helps quantify the alignment of SHAP explanations with the expert-validated ground truth, indicating the accuracy and reliability of the explanations provided by SHAP.

Implementing SHAP in the evaluation of LLMs allows for a detailed examination of how specific input features influence model predictions, thus enhancing model transparency and building trust. Through this methodology, we aim to illuminate the inner workings of sophisticated LLMs, providing valuable insights into their operational mechanics and decision-making processes. This approach not only aids in model validation and improvement but also fosters more informed and responsible use of AI technology in real-world applications.

D. Specific Utility of Cross weight

The Cross weight(w) calculation compares the ground truth relevance scores (gt(word))—expert-defined values indicating the importance of each word in making a decision—with the predicted relevance scores (p(word)) generated by the model's explainability method. By multiplying these scores and summing across all features, the Cross weight(w) effectively measures how well the model's assessments of feature importance align with expert judgements.

1) Cross weight relevance:

- Correlation with Expert Judgement: A higher Cross weight(w) indicates a stronger correlation between the model's interpretation of feature importance and the expert-defined benchmarks. This correlation is essential in contexts where understanding the reasoning behind a model's decision is as important as the decision itself, such as in medical diagnostics or financial forecasting.
- Identifying Misalignment: If certain features consistently result in negative contributions to the Cross weight(w), it suggests that the model either overestimates or underestimates their importance compared to expert expectations. This can identify specific areas where the model's understanding is flawed or biased, guiding further refinement and training.
- Model Debugging and Improvement: By analyzing the components contributing to the Cross weight(w), developers can pinpoint specific features where the model's explainability needs improvement. This can be particularly valuable in iterative model development, where successive versions of a model are refined based on detailed feedback on their performance.
- Model Comparison: When evaluating multiple models or explainability approaches, the Cross weight(w) provides a standardized metric to compare their effectiveness in mirroring expert understanding. Models with higher Cross weight(w) scores are deemed more reliable in terms of explainability, assuming the ground truth is accurately defined.

IV. Experiments and Results

The evaluation of the modified Large Language Models (LLMs) using LIME interpretability techniques produced in-

sightful results, particularly when assessed through the metric of Cross Average (C) scores across different categories. The models under consideration included Mixtral, Gemma, Falcon, Bloom, and mpt7b. These models were evaluated across four categories: Entertainment, Health, Sports, and Technology. The results provide a nuanced view of each model's ability to handle different contexts and highlight areas where explainability can be further enhanced.

A. LIME Explanability Analysis

In the LIME analysis, the comparative performance of various language model models across different categories reveals distinct patterns in their ability to align predictions with expert-annotated ground truth relevance. This section provides a detailed examination of how models such as Mixtral, Gemma, Falcon, Bloom, and mpt7b perform across Entertainment, Health, Sports, and Technology categories, highlighting their strengths and weaknesses in interpretability and feature relevance.

1) Comparative Analysis of Models: The Cross Average scores, which measure the alignment of model predictions with the ground truth relevance established through expert annotations, revealed varied performance across models and categories:

Mixtral: This model generally underperformed, showing negative scores in Entertainment (-0.01717), Health (-0.03993), and Technology (-0.00825), with a slightly positive score in Sports (0.00421). The negative values indicate a significant misalignment between the model's predictions and the expert-defined relevance, suggesting areas where Mixtral may misinterpret or overweight certain features irrelevant to the correct categorization. Gemma, Falcon, Bloom, and mpt7b: These models demonstrated more favorable outcomes, with all posting positive scores across the categories. Notably, Falcon and mpt7b showed a marked improvement in Technology, with scores of 0.01647 and 0.01644, respectively, indicating a strong alignment with the expected relevance values. The consistent positive scores across these models suggest a better handling of feature relevance, contributing to more accurate and explainable outcomes.

The average across each categories calculated as follows:

$$\text{Overall Average Cross Weight} = \frac{1}{N}\sum_{i=1}^{N}\text{Cross weight}(c_i)$$

where N is the no of data-point considered within the particular category.

TABLE I
LIME CROSS WEIGHT AVERAGE RESULTS (ACROSS CATEGORIES)

Category	**MISTRAL**	**GEMMA**	**FALCON**	**BLOOM**	**MPT**
Entertainment	-0.01717	0.01128	0.01153	0.01170	0.01199
Health	-0.03993	0.01437	0.01411	0.01573	0.01487
Sports	0.00421	0.01026	0.01225	0.01012	0.00935
Technology	-0.00825	0.01496	0.01647	0.01393	0.01644
Grand Total	**-0.01609**	**0.01285**	**0.01367**	**0.01300**	**0.01336**

2) Model Performance by Category: Entertainment and Health: All models except Mixtral performed well in these categories, with positive scores that indicate a good grasp of relevant features contributing to the sentiment analysis tasks. Gemma and Bloom showed particularly strong performance in Health, with scores of 0.01437 and 0.01573, respectively. Sports: This category saw generally positive scores from all models, indicating a uniform ability to correctly weigh features relevant to sports-related contexts. Falcon's score of 0.01225 was the highest, suggesting slight advantages in feature handling for sports content. Technology: Notably, this category showcased the highest scores for Falcon and mpt7b, reinforcing their capability in more technical or specific content, likely due to better contextual understanding or feature processing.

B. SHAP Interpretability Analysis

The SHAP interpretability analysis was conducted across a spectrum of foundational Large Language Models (LLMs), including Bloom, Llama, Falcon, Gemma, and Mistral. These models were evaluated using a standardized dataset, focusing on their ability to accurately interpret and predict sentiment across five categories: negative, neutral, positive, very negative, and very positive. The results, expressed as average SHAP values (Cross Average scores), provide insight into each model's effectiveness and the transparency of their decision-making processes.

1) Analysis of SHAP Values Across Categories: Negative Sentiment: Gemma and Mistral exhibited strong performance with average scores of 0.40880 and 0.37999, respectively, suggesting a robust ability to attribute correct feature importance in negative contexts. However, Falcon under-performed significantly with a negative score of -0.19666, indicating potential misalignment in interpreting features associated with negative sentiments. Neutral Sentiment: Mixed results were observed in the neutral category, with Mistral achieving the highest positive score (0.09786) among the models, suggesting effective handling of features for neutral sentiment. In contrast, Falcon and Llama displayed negative scores (-0.04157 and -0.01012 respectively), pointing to difficulties in accurately capturing the subtleties of neutral language cues. Positive Sentiment: In positive sentiments, Mistral again showed a strong understanding with a score of 0.19809, closely followed by Bloom with 0.18047. Falcon's performance was notably poorer, with a negative score of -0.15530, reflecting challenges in recognizing or weighing positive sentiment features correctly. Very Negative and Very Positive Sentiments: In these more extreme sentiment categories, Gemma and Bloom displayed particularly high scores, 0.67156 and 0.68208 respectively for very negative and very positive sentiments, indicating their capability to handle distinct and strong sentiment features effectively. Llama and Mistral also performed well, suggesting a generally good model sensitivity to more intense emotional expressions. The average calculation across the different categories is the same as previous.

Overall, Gemma demonstrated the highest average SHAP value (0.36543), indicating its superior capability across differ-

TABLE II
SHAP CROSS WEIGHT AVERAGE RESULTS (ACROSS CATEGORIES)

Label Text	BLOOM	LLAMA	FALCON	GEMMA	MISTRAL
Negative	0.19088	0.18167	-0.19666	0.40880	0.37999
Neutral	0.08320	-0.01012	-0.04157	0.00464	0.09786
Positive	0.18047	0.12185	-0.15530	0.06005	0.19809
Very Negative	0.15434	0.22579	0.11284	0.67156	0.36215
Very Positive	0.40303	0.26238	0.22300	0.68208	0.30700
Grand Total	**0.20238**	**0.15631**	**-0.01154**	**0.36543**	**0.26902**

ent sentiments in attributing correct importance to the relevant features. Mistral also performed well, with an overall average of 0.26902, reflecting its effectiveness in diverse sentiment analyses. Conversely, Falcon exhibited an overall negative average (-0.01154), highlighting potential deficiencies in its explainability framework or feature handling mechanisms.

C. Limitations and Considerations

The application of LIME, while revealing in terms of localized explanations, also surfaced inherent limitations in the methodology:

Model Sensitivity: The variability in scores, particularly the negative values observed with Mixtral for LIME results, may reflect a sensitivity to how features are represented and weighted. For instance, the term "Apple" can be challenging as it may refer to both a fruit and a corporation, leading to disparate interpretations depending on the surrounding context used by the model.

Ambiguity in Feature Relevance: In cases where LIME's randomized sampling fails to capture the complete context in which a word is used (e.g., "Apple" as part of a broader discussion on technology versus health), the resulting explanations may not accurately reflect the reasons behind a model's prediction, leading to potential misinterpretations or erroneous relevance attributions.

V. CONCLUSION AND FUTURE WORK

The analysis using LIME and SHAP interpretability methods revealed key insights into the explainability of LLMs like Mistral, Gemma, Falcon, Bloom, and Llama. The study highlighted significant differences in how each model processes and interprets input features, emphasizing the importance of tailored explainability for understanding AI decision-making. This enhances transparency and builds trust in AI systems by making their decisions more interpretable.

Future research should focus on developing custom explainability metrics for specific LLM architectures to better reflect their decision-making processes. Creating engineered datasets to address model weaknesses and implementing fine-tuning protocols based on these datasets can improve models' handling of complex inputs. Iterative evaluations will help validate improvements in explainability.

In conclusion, while this research lays a foundation for understanding LLM explainability, ongoing development is needed. Future efforts will aim to create more nuanced tools and methods, refine datasets, and enhance fine-tuning processes to ensure LLMs are both effective and understandable, fostering greater trust and wider adoption of these AI systems.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozire, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[5] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[6] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin *et al.*, "Code llama: Open foundation models for code," *arXiv preprint arXiv:2308.12950*, 2023.

[7] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic *et al.*, "The falcon series of open language models," *arXiv preprint arXiv:2311.16867*, 2023.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[9] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[10] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang *et al.*, "Gpt-neox-20b: An open-source autoregressive language model," *arXiv preprint arXiv:2204.06745*, 2022.

[11] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[12] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of lime," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 1287–1296.

[13] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[14] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[15] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[16] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *International Conference on Machine Learning*. PMLR, 2023, pp. 2397–2430.

[17] D. Zan, B. Chen, F. Zhang, D. Lu, B. Wu, B. Guan, Y. Wang, and J.-G. Lou, "Large language models meet nl2code: A survey," *arXiv preprint arXiv:2212.09420*, 2022.

[18] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[19] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[20] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu *et al.*, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, p. 100017, 2023.

[21] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[22] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?" *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, "Chatgpt and large language models in academia: opportunities and challenges," *BioData Mining*, vol. 16, no. 1, p. 20, 2023.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[25] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[28] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivire, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[30] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[31] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[32] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," 2016. [Online]. Available: https://arxiv.org/abs/1606.04155

[33] M. Christoph, *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020.