Article

Exploring Explainability in Large Language Models

Fen Yin , Mu Zhong , Zhihao Ru [*]

Posted Date: 31 March 2025

doi: 10.20944/preprints202503.2318.v1

Keywords: Explainable AI; Large Language Models; Interpretability; feature attribution; attention analysis; counterfactual explanations; causal reasoning; trust; transparency; scalability; Bias Mitigation; Ethical AI; Model Transparency; trustworthy AI

Article

Exploring Explainability in Large Language Models

Fen Yin, Mu Zhong and Zhihao Ru *

Department of Computer Science, Chinese University of Hong Kong; fen.yin@cuhk.edu.hk (F.Y.); mu.zhong@cuhk.edu.hk (M.Z.)
* Correspondence: zhihao.ru@cuhk.edu.hk

Abstract: Explainable AI (XAI) has become an essential area of research, particularly in the era of Large Language Models (LLMs), which power a wide range of applications spanning natural language processing, automated decision-making, and conversational AI. These models have demonstrated remarkable capabilities in generating human-like text, answering complex queries, and assisting in diverse fields such as healthcare, finance, and legal analysis. However, despite their impressive performance, LLMs operate as black boxes with intricate, non-transparent decision-making processes. This opacity raises significant concerns regarding trust, interpretability, and accountability, particularly when these models are deployed in high-stakes domains where incorrect or biased outputs can have serious consequences. To bridge this gap, researchers have been actively developing techniques to enhance the interpretability of LLMs, enabling users to gain insights into model predictions and behavior. This paper explores various XAI methodologies, including feature attribution methods that identify the importance of input tokens, attention analysis that examines weight distributions within transformer architectures, and counterfactual explanations that highlight the minimal changes required to alter an output. Additionally, we delve into causal reasoning approaches, which attempt to establish cause-and-effect relationships within model decision-making pathways, providing a more robust understanding of model predictions. Beyond technical methodologies, this paper also discusses key challenges associated with LLM explainability. One of the foremost challenges is scalability—many XAI techniques, such as SHAP and LIME, are computationally expensive and struggle to scale effectively for billion-parameter models. Another pressing concern is the faithfulness of explanations; while methods such as attention visualization provide some level of insight, they do not necessarily align with the actual reasoning processes of the model, raising doubts about their reliability. Ethical considerations, including bias detection and mitigation, are also critical, as LLMs have been shown to inherit and propagate biases present in their training data. Ensuring that explanations are transparent, unbiased, and aligned with ethical principles remains a major research challenge. Finally, we outline potential solutions and future research directions in the field of explainable AI for LLMs. These include the development of more scalable and efficient interpretability techniques, the creation of human-centered explanation frameworks tailored to different stakeholders, and the integration of causal inference methods to provide deeper insights into model behavior. Additionally, regulatory and ethical frameworks must evolve to keep pace with advancements in AI, ensuring that models are not only interpretable but also adhere to legal and societal norms. Addressing these challenges is crucial to fostering trust and ensuring that LLMs remain transparent, fair, and aligned with human values as they continue to evolve and influence various aspects of daily life.

Keywords: Explainable AI; Large Language Models; Interpretability; feature attribution; attention analysis; counterfactual explanations; causal reasoning; trust; transparency; scalability; Bias Mitigation; Ethical AI; Model Transparency; trustworthy AI

1. Introduction

Artificial Intelligence (AI) has seen significant advancements with the emergence of large language models (LLMs) such as GPT, BERT, and PaLM [1]. These models have demonstrated remarkable capabilities in natural language understanding, generation, and reasoning [2,3]. However, their black-box nature raises concerns about transparency, accountability, and trustworthiness, particularly in high-stakes applications such as healthcare, finance, and law [4]. Explainable AI (XAI) has emerged as a critical area of research to address these concerns by providing interpretability and insights into model decision-making [5,6]. The need for explainability in AI is more pressing than ever, given the increasing reliance on LLMs in various domains [7]. While traditional XAI techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) have been effective for smaller models, their applicability to LLMs remains an open challenge [8]. The sheer scale and complexity of these models require novel approaches to interpretability that go beyond feature importance and local surrogate models [9]. In this paper, we explore the evolving landscape of explainable AI in the context of LLMs [10]. We discuss existing XAI methods, their limitations when applied to LLMs, and emerging techniques tailored for large-scale deep learning models [11]. Additionally, we highlight the trade-offs between explainability and performance, the role of human-centered AI in fostering trust, and the ethical considerations surrounding the deployment of interpretable LLMs [12]. The rest of the paper is structured as follows: Section 2 provides an overview of explainability techniques in AI, focusing on their applicability to LLMs [13]. Section 3 discusses recent advancements in explainable AI for large-scale models [14,15]. Section 4 highlights key challenges and open research questions [16]. Finally, Section 5 summarizes our findings and outlines future directions in this rapidly evolving field [17].

2. Background and Related Work

The field of Explainable AI (XAI) has gained increasing prominence in recent years as artificial intelligence (AI) systems become more complex and pervasive [18,19]. The need for explainability is particularly crucial in domains such as healthcare, finance, and legal decision-making, where AI-driven predictions must be interpretable, trustworthy, and transparent [20]. While traditional machine learning models have benefitted from well-established explainability techniques, the emergence of Large Language Models (LLMs) has introduced new challenges and necessitated novel interpretability methods [21].

2.1. Traditional Explainability Techniques

Early efforts in XAI focused on methods that could provide local or global explanations of AI models [22,23]. Some of the widely adopted traditional explainability techniques include:

- **Feature Importance Methods**: Techniques such as SHAP (Shapley Additive Explanations) [24] and LIME (Local Interpretable Model-agnostic Explanations) [25] are designed to attribute importance to input features by approximating local decision boundaries or distributing contributions based on cooperative game theory principles [26].
- **Saliency Maps**: Methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) [**?**] and integrated gradients [27] have been effective in highlighting which input features contribute most to a model's predictions, particularly in image and text classification tasks [28].
- **Rule-based Explanations**: Decision rules and rule extraction techniques aim to simplify complex models by approximating their behavior with interpretable if-then statements [19,29].
- **Surrogate Models**: Interpretable models, such as decision trees or linear regressions, are trained to mimic the behavior of complex models in an effort to generate human-understandable explanations [30**?**].

While these techniques have proven effective for smaller-scale models, they often fail to scale appropriately to modern deep learning architectures, particularly LLMs, which contain billions of parameters and exhibit highly nonlinear decision boundaries [31].

doi:10.20944/preprints202503.2318.v1

2.2. Challenges in Explaining Large Language Models

LLMs such as GPT-4, BERT, PaLM, and LLaMA leverage deep transformer architectures and self-attention mechanisms to process and generate human-like text [32,33]. Unlike traditional machine learning models, LLMs are inherently complex due to their vast parameter space, emergent properties, and dependency on massive-scale training data [34,35]. These factors introduce significant challenges in the explainability of LLMs, including:

- **Opacity and Lack of Interpretability**: Unlike simpler models, the internal representations of LLMs are not easily understandable by humans, making it difficult to extract meaningful explanations for their predictions [36].
- **Scale and Computational Complexity**: The sheer size of modern LLMs makes traditional feature attribution methods computationally expensive and often impractical for real-time interpretability [37].
- **Contextual Dependencies**: Unlike structured machine learning models, LLMs rely on sequential token dependencies, making it difficult to attribute decisions to a specific input token or phrase [38].
- **Bias and Ethical Concerns**: LLMs are prone to biases inherited from training data, which can manifest in outputs in unpredictable ways, highlighting the need for transparency in their decision-making processes [39**?**].

2.3. Recent Advances in XAI for LLMs

To address the limitations of traditional XAI methods in the context of LLMs, researchers have proposed various novel approaches tailored for large-scale language models [40]. Some of the most promising recent advancements include:

- **Attention-based Interpretability**: Analyzing attention weights in transformer models has been a popular method to infer how LLMs process and prioritize information [41]. However, attention does not necessarily equate to explanation, as model behavior is influenced by complex interactions beyond attention scores {[42**?**] .
- **Concept-based Explanations**: Methods such as TCAV (Testing with Concept Activation Vectors) [**?**] aim to identify and attribute high-level human-understandable concepts in LLM representations, bridging the gap between black-box models and interpretable reasoning [43].
- **Counterfactual and Contrastive Explanations**: Recent studies explore counterfactual reasoning by modifying inputs and analyzing how predictions change, allowing users to understand the decision boundaries of LLMs [44,45].
- **Causal Analysis Techniques**: Causal inference methods seek to disentangle causal relationships within LLMs by identifying which internal components contribute most significantly to specific outputs [46**?**].
- **Human-in-the-Loop Explainability**: Interactive approaches that allow human users to query and refine explanations dynamically have gained traction, providing adaptive and context-specific insights [7,47,48].

2.4. Ethical and Societal Implications of Explainability in LLMs

The growing deployment of LLMs in real-world applications necessitates careful consideration of their ethical and societal impacts [49,50]. Ensuring that LLMs are explainable is crucial for mitigating bias, preventing misinformation, and fostering user trust [51]. Some of the key ethical concerns include:

- **Fairness and Bias Mitigation**: Explainability techniques can help identify biases in LLMs, enabling interventions to reduce discriminatory outputs [25**?**].
- **Accountability and Transparency**: Regulatory frameworks increasingly demand that AI decisions be explainable, particularly in high-stakes domains such as healthcare and law [52**?**].
- **User Trust and Adoption**: Providing interpretable explanations improves user trust and facilitates broader acceptance of AI technologies in decision-making processes [53,54].

2.5. Summary and Research Gaps

Despite significant progress in XAI for LLMs, several open challenges remain [55]. Existing methods often struggle to balance fidelity and interpretability, while scalability remains a persistent issue [56]. Additionally, ensuring that explanations are meaningful and actionable for end-users requires further exploration [53]. In the next section, we delve into the latest methodologies aimed at addressing these gaps and improving the explainability of LLMs [57]. The following section will provide a deeper analysis of state-of-the-art explainability techniques and their applications in real-world scenarios [58].

3. Methodologies for Explainable AI in LLMs

Developing effective methodologies for explainable AI in large language models (LLMs) is crucial to ensuring their transparency, trustworthiness, and usability [59]. This section explores state-of-the-art techniques used to enhance interpretability in LLMs, categorized into different approaches [60].

3.1. Feature Attribution Methods

Feature attribution methods aim to identify which input tokens or features contribute most to the model's predictions [61]. Some of the most widely used techniques include:

- **Gradient-based Methods**: Techniques such as Integrated Gradients (IG) [27] and Saliency Maps analyze gradients to determine how input tokens influence model outputs [62].
- **SHAP and LIME**: While traditionally used for structured data, adaptations of SHAP (Shapley Additive Explanations) [24] and LIME (Local Interpretable Model-agnostic Explanations) [25] have been explored for language models to approximate local interpretability [63].

3.2. Attention-based Analysis

Since LLMs are based on transformer architectures, analyzing attention weights is a natural approach to understanding their decision-making process:

- **Attention Heatmaps**: Visualization of attention distributions across tokens helps infer model focus [64].
- **Limitations of Attention Weights**: Studies suggest that attention alone does not provide complete explanations, as attention is not a direct measure of causality [65**?**].

3.3. Concept-based Explanations

Concept-based methods attempt to align model representations with human-understandable concepts:

- **Testing with Concept Activation Vectors (TCAV)**: A technique that interprets model behavior by associating activations with predefined concepts [66**?** ,67].
- **Probing Classifiers**: Small classifiers are trained to extract interpretable representations from LLMs [68**?** ,69].

3.4. Counterfactual and Contrastive Explanations

Counterfactual analysis helps users understand model behavior by examining minimal input changes that alter outputs:

- **Counterfactual Text Generation**: Methods generate alternative inputs to explore model robustness and biases [44,70].
- **Contrastive Explanations**: Highlighting key differences between similar instances to justify why one output was chosen over another [71].

3.5. Causal Analysis and Model Distillation

Causal inference methods aim to determine cause-effect relationships within LLMs:

- **Intervention-based Causal Analysis**: Techniques like ablation studies and knockout experiments analyze the causal impact of different layers and neurons [72**?** ,73].
- **Model Distillation for Interpretability**: Simplifying LLMs by training smaller, interpretable models to mimic their behavior [74**?**].

3.6. Human-centered and Interactive Explainability

Explainability is most effective when tailored to human understanding [75]. Interactive and user-driven methods include:

- **Human-in-the-Loop Systems**: Enabling users to query models and refine explanations dynamically [7,76].
- **Natural Language Explanations**: Generating explanations in human-readable text to facilitate understanding and transparency [77].

This section has outlined various methodologies for explainable AI in LLMs, highlighting their strengths and limitations [78]. The next section will discuss key challenges and open research questions in the field [79].

4. Challenges and Open Research Questions

Despite significant progress in the field of explainable AI (XAI) for large language models (LLMs), numerous challenges persist [80]. This section explores key obstacles that hinder the development of effective interpretability techniques and identifies open research questions that require further investigation [81].

4.1. Scalability and Computational Constraints

LLMs contain billions of parameters, making it computationally expensive to apply traditional XAI techniques at scale [82,83]. Methods such as SHAP and LIME, which work well for smaller models, struggle to provide real-time interpretability in LLMs due to the high computational cost [84,85].

- How can we develop efficient, scalable XAI techniques suitable for billion-parameter models [86]?
- What trade-offs exist between computational efficiency and the quality of explanations in LLMs [87]?

4.2. Faithfulness and Reliability of Explanations

Many existing XAI techniques, such as attention-based explanations, do not necessarily provide faithful representations of model reasoning [88]. The challenge lies in ensuring that explanations accurately reflect the internal decision-making processes of LLMs [89].

- How can we measure and improve the faithfulness of XAI methods for LLMs [90]?
- Are there novel approaches that can provide causal rather than correlational explanations [91]?

4.3. Interpretability vs . Performance Trade-offs

There is often a trade-off between model interpretability [92] and predictive performance [93]. While simpler models are easier to explain, they may not match the accuracy of complex LLMs [94]. Balancing interpretability without sacrificing model effectiveness remains an open research question [95][96].

- Can we develop hybrid models that optimize both interpretability and performance [97]?
- How do we evaluate the trade-offs between human-understandable explanations and model accuracy [98]?

4.4. User-Centric and Domain-Specific Explanations

Explanations need to be tailored to specific user groups and application domains [99]. What is considered interpretable to a machine learning researcher may not be useful for a healthcare professional or a legal expert [100].

doi:10.20944/preprints202503.2318.v1

- How can we design adaptive, user-centric XAI systems that adjust explanations based on domain expertise and user needs [24,101,102]?
- What role does human feedback play in refining and validating LLM explanations [103]?

4.5. Mitigating Bias and Ethical Concerns

LLMs have been shown to encode and propagate biases from their training data [104]. Explainability techniques can help detect and mitigate bias, but ensuring fairness in LLMs remains a complex challenge [105].

- What are effective strategies to use XAI for bias detection and mitigation in LLMs [106]?
- How do we ensure that explainable AI aligns with ethical and legal standards across different regions [107]?

4.6. Future Directions in Explainable AI for LLMs

Addressing these challenges requires interdisciplinary collaboration across AI researchers, domain experts, and policymakers [108,109]. Future research should focus on developing more robust, human-aligned, and efficient XAI methods that can scale to the complexity of modern LLMs [110]. The next section will summarize key findings and outline future research directions in the field of explainable AI for large language models [111].

5. Conclusion and Future Research Directions

Explainable AI (XAI) in the era of Large Language Models (LLMs) is a rapidly evolving field that seeks to address critical challenges related to transparency, trust, and accountability [112,113]. This paper has explored various methodologies for improving the interpretability of LLMs, discussed key challenges, and highlighted open research questions [114].

5.1. Summary of Key Findings

The research reviewed in this paper underscores several critical aspects of explainability in LLMs:

- Traditional XAI methods such as SHAP, LIME, and attention-based mechanisms provide insights into model behavior but are often insufficient for the complexity of LLMs [115].
- Emerging techniques, including causal analysis, concept-based explanations, and counterfactual reasoning, offer promising directions for improving interpretability [116].
- Ethical and societal implications of LLM explainability, such as bias detection, fairness, and regulatory compliance, require ongoing investigation [117,118].
- Trade-offs between interpretability and performance remain a significant hurdle, necessitating new methods that balance fidelity and usability [119].

5.2. Future Research Directions

To advance explainability in LLMs, future research should focus on the following areas:

- **Scalable and Efficient XAI Techniques**: Developing methods that can handle the complexity of billion-parameter models while remaining computationally feasible [120,121].
- **Human-Centric Explanations**: Designing adaptive and interactive XAI systems that provide explanations tailored to specific user needs and domains [122].
- **Causal and Counterfactual Approaches**: Improving causal inference techniques to provide more meaningful and actionable explanations [123].
- **Regulatory and Ethical Frameworks**: Establishing guidelines and best practices to ensure that XAI aligns with legal and ethical standards [124].
- **Integration with Human Feedback**: Enhancing XAI techniques through active learning and human-in-the-loop approaches to refine explanations dynamically [125].

5.3. Final Thoughts

As LLMs continue to advance and integrate into various domains, ensuring their explainability remains a critical challenge. The interdisciplinary nature of XAI research requires collaboration between AI developers, domain experts, and policymakers to create robust, transparent, and human-aligned AI systems. By addressing the outlined challenges and pursuing innovative research directions, the field of explainable AI can make significant strides toward enhancing trust and usability in LLMs.

References

1. Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; Miao, Z. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In Proceedings of the Proceedings of the CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2024; CHI '24, pp. 1–21. https://doi.org/10.1145/3613904.3641960.
2. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation* **2024**, *16*, 45–74.
3. Arras, L.; Horn, F.; Montavon, G.; Müller, K.R.; Samek, W. "What is relevant in a text document?": An interpretable machine learning approach. *PloS one* **2017**, *12*, e0181142.
4. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* **2019**.
5. Holliday, D.; Wilson, S.; Stumpf, S. User trust in intelligent systems: A journey over time. In Proceedings of the Proceedings of the 21st International Conference on Intelligent User Interfaces, 2016, pp. 164–168.
6. Maliha, G.; Gerke, S.; Cohen, I.G.; Parikh, R.B. Artificial Intelligence and Liability in Medicine. *The Milbank Quarterly* **2021**, *99*, 629–647.
7. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
8. Zhang, G.; Kashima, H. Learning state importance for preference-based reinforcement learning. *Machine Learning* **2023**, pp. 1–17.
9. Nwakanma, C.I.; Ahakonye, L.A.C.; Njoku, J.N.; Odirichukwu, J.C.; Okolie, S.A.; Uzondu, C.; Ndubuisi Nweke, C.C.; Kim, D.S. Explainable Artificial Intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Applied Sciences* **2023**, *13*, 1252.
10. Rahman, M.; Polunsky, S.; Jones, S. Transportation policies for connected and automated mobility in smart cities. In *Smart Cities Policies and Financing*; Elsevier, 2022; pp. 97–116.
11. Aubin Le Quéré, M.; Schroeder, H.; Randazzo, C.; Gao, J.; Epstein, Z.; Perrault, S.T.; Mimno, D.; Barkhuus, L.; Li, H. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, 2024; pp. 1–7. https://doi.org/10.1145/3613905.3636301.
12. Chen, V.; Liao, Q.V.; Wortman Vaughan, J.; Bansal, G. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction* **2023**, *7*, 1–32.
13. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18.
14. Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *Journal of Chemical Information and Modeling* **2022**, *62*, 447–462.
15. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) program. *AI Magazine* **2019**, *40*, 44–58.
16. Nourani, M.; Kabir, S.; Mohseni, S.; Ragan, E.D. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In Proceedings of the Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2019, Vol. 7, pp. 97–105.
17. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **2022**, *35*, 24824–24837.
18. Adhikari, A.; Tax, D.M.J.; Satta, R.; Faeth, M. LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models. In Proceedings of the 2019 IEEE International Conference on Fuzzy

Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019; pp. 1–7. https://doi.org/10.1109/FUZZ-IEEE.2019.8858846.

19. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* **2018**, *51*, 1–42.
20. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **2019**, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.
21. Jie, Y.W.; Satapathy, R.; Mong, G.S.; Cambria, E.; et al. How Interpretable are Reasoning Explanations from Prompting Large Language Models? *arXiv preprint arXiv:2402.11863* **2024**.
22. Burton, S.; Habli, I.; Lawton, T.; McDermid, J.; Morgan, P.; Porter, Z. Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence* **2020**, *279*, 103201.
23. Hamamoto, R. Application of artificial intelligence for medical research, 2021.
24. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017**, *30*.
25. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
26. Bian, Z.; Xia, S.; Xia, C.; Shao, M. Weakly supervised vitiligo segmentation in skin image through saliency propagation. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 931–934.
27. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the International Conference on Machine Learning. PMLR, 2017, pp. 3319–3328.
28. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A review of trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**.
29. Marcinkevičs, R.; Vogt, J.E. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805* **2020**.
30. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Towards safe, explainable, and regulated autonomous driving. *arXiv preprint arXiv:2111.10518* **2021**.
31. Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences* **2022**, *12*, 9423.
32. Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; Cottrell, G. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971* **2017**.
33. Zafar, M.R.; Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction* **2021**, *3*, 525–541.
34. Saraswat, D.; Bhattacharya, P.; Verma, A.; Prasad, V.K.; Tanwar, S.; Sharma, G.; Bokoro, P.N.; Sharma, R. Explainable AI for healthcare 5.0: opportunities and challenges. *IEEE Access* **2022**.
35. Li, L.; Xu, M.; Liu, H.; Li, Y.; Wang, X.; Jiang, L.; Wang, Z.; Fan, X.; Wang, N. A large-scale database and a CNN model for attention-based glaucoma detection. *IEEE transactions on Medical Imaging* **2019**, *39*, 413–424.
36. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**.
37. Mankodiya, H.; Jadav, D.; Gupta, R.; Tanwar, S.; Hong, W.C.; Sharma, R. Od-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences* **2022**, *12*, 5310.
38. Bano, M.; Zowghi, D.; Whittle, J. Exploring Qualitative Research Using LLMs **2023**.
39. van der Waa, J.; Nieuwburg, E.; Cremers, A.; Neerincx, M. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* **2021**, *291*, 103404. https://doi.org/10.1016/j.artint.2020.103404.
40. Wang, B.; Zhou, J.; Li, Y.; Chen, F. Impact of Fidelity and Robustness of Machine Learning Explanations on User Trust. In Proceedings of the Australasian Joint Conference on Artificial Intelligence. Springer, 2023, pp. 209–220.
41. Regulation, P. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)* **2016**, *679*, 2016.
42. Chern, S.; Chern, E.; Neubig, G.; Liu, P. Can Large Language Models be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate, 2024. arXiv:2401.16788 [cs].

43. Krause, J.; Perer, A.; Ng, K. Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 5686–5697.
44. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* **2017**, *31*, 841.
45. Anton, N.; Doroftei, B.; Curteanu, S.; Catãlin, L.; Ilie, O.D.; Târcoveanu, F.; Bogdănici, C.M. Comprehensive review on the use of artificial intelligence in ophthalmology and future research directions. *Diagnostics* **2022**, *13*, 100.
46. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* **2023**, *99*, 101805.
47. Job, S.; Tao, X.; Li, L.; Xie, H.; Cai, T.; Yong, J.; Li, Q. Optimal treatment strategies for critical patients with deep reinforcement learning. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1–22.
48. Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K.E.; Dickerson, J.P.; Shah, C. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* **2020**.
49. Hanawa, K.; Yokoi, S.; Hara, S.; Inui, K. Evaluation of Similarity-based Explanations, 2021. arXiv:2006.04528 [cs, stat].
50. Kha, Q.H.; Le, V.H.; Hung, T.N.K.; Nguyen, N.T.K.; Le, N.Q.K. Development and Validation of an Explainable Machine Learning-Based Prediction Model for Drug–Food Interactions from Chemical Structures. *Sensors* **2023**, *23*, 3962.
51. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners, 2023. arXiv:2205.11916 [cs].
52. Farahat, A.; Reichert, C.; Sweeney-Reed, C.M.; Hinrichs, H. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of Neural Engineering* **2019**, *16*, 066010.
53. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **2020**, *58*, 82–115.
54. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer, 2014, pp. 818–833.
55. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* **2022**, *77*, 29–52.
56. Castelnovo, A.; Depalmas, R.; Mercorio, F.; Mombelli, N.; Potertì, D.; Serino, A.; Seveso, A.; Sorrentino, S.; Viola, L. Augmenting XAI with LLMs: A Case Study in Banking Marketing Recommendation. In Proceedings of the Explainable Artificial Intelligence; Longo, L.; Lapuschkin, S.; Seifert, C., Eds., Cham, 2024; pp. 211–229. https://doi.org/10.1007/978-3-031-63787-2_11.
57. Oviedo, F.; Ferres, J.L.; Buonassisi, T.; Butler, K.T. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research* **2022**, *3*, 597–607.
58. Lertvittayakumjorn, P.; Toni, F. Human-grounded evaluations of explanation methods for text classification. *arXiv preprint arXiv:1908.11355* **2019**.
59. Kolla, M.; Salunkhe, S.; Chandrasekharan, E.; Saha, K. LLM-Mod: Can Large Language Models Assist Content Moderation? In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, 2024; pp. 1–8. https://doi.org/10.1145/3613905.3650828.
60. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-precision model-agnostic explanations. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
61. Albahri, A.; Duhaim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.; Alamoodi, A.; Bai, J.; Salhi, A.; et al. A systematic review of trustworthy and Explainable Artificial Intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* **2023**.
62. Rudin, C.; Radin, J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* **2019**, *1*, 1–9.
63. El Naqa, I.; Murphy, M.J. *What is machine learning?*; Springer, 2015.
64. Huang, Z.; Yao, X.; Liu, Y.; Dumitru, C.O.; Datcu, M.; Han, J. Physically explainable CNN for SAR image classification. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *190*, 25–37.

doi:10.20944/preprints202503.2318.v1

65. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning; Precup, D.; Teh, Y.W., Eds. PMLR, 06–11 Aug 2017, Vol. 70, *Proceedings of Machine Learning Research*, pp. 3319–3328.
66. Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* **1936**, *7*, 179–188. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
67. Munir, M. Thesis approved by the Department of Computer Science of the TU Kaiserslautern for the award of the Doctoral Degree doctor of engineering. PhD thesis, Kyushu University, Japan, 2021.
68. Hu, T.; Zhou, X.H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv preprint arXiv:2404.09135* **2024**.
69. Puiutta, E.; Veith, E.M. Explainable reinforcement learning: A survey. In Proceedings of the International Cross-domain Conference for Machine Learning and Knowledge Extraction. Springer, 2020, pp. 77–95.
70. Ma, S.; Chen, Q.; Wang, X.; Zheng, C.; Peng, Z.; Yin, M.; Ma, X. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making, 2024. arXiv:2403.16812 [cs].
71. Weller, A. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*; Springer, 2019; pp. 23–40.
72. Yilma, B.A.; Kim, C.M.; Cupchik, G.C.; Leiva, L.A. Artful Path to Healing: Using Machine Learning for Visual Art Recommendation to Prevent and Reduce Post-Intensive Care Syndrome (PICS). In Proceedings of the Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–19.
73. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561* **2021**.
74. Ismail, A.A.; Gunady, M.; Corrada Bravo, H.; Feizi, S. Benchmarking deep learning interpretability in time series predictions. *Advances in Neural Information Processing Systems* **2020**, *33*, 6441–6452.
75. Sadeghi Tabas, S. Explainable Physics-informed Deep Learning for Rainfall-runoff Modeling and Uncertainty Assessment across the Continental United States **2023**.
76. Plumb, G.; Wang, S.; Chen, Y.; Rudin, C. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In Proceedings of the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp. 1677–1686.
77. Crocker, J.; Kumar, K.; Cox, B. Using explainability to design physics-aware CNNs for solving subsurface inverse problems. *Computers and Geotechnics* **2023**, *159*, 105452.
78. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593.
79. Weber, L.; Lapuschkin, S.; Binder, A.; Samek, W. Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion* **2023**, *92*, 154–176.
80. Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; Höhne, M.M.C. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research* **2023**, *24*, 1–11.
81. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215.
82. Schlegel, U.; Keim, D.A. Time series model attribution visualizations as explanations. In Proceedings of the 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX). IEEE, 2021, pp. 27–31.
83. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
84. Alharin, A.; Doan, T.N.; Sartipi, M. Reinforcement learning interpretation methods: A survey. *IEEE Access* **2020**, *8*, 171058–171077.
85. Zhu, Y.; Zhou, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Soft proposal networks for weakly supervised object localization. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1841–1850.
86. Cooper, J.; Arandjelović, O.; Harrison, D.J. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognition* **2022**, *129*, 108743.
87. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **2017**, *60*, 84–90.

doi:10.20944/preprints202503.2318.v1

88. Tjoa, E.; Guan, C. A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2020**, *32*, 4793–4813.
89. Madhav, A.S.; Tyagi, A.K. Explainable Artificial Intelligence (XAI): connecting artificial decision-making and human trust in autonomous vehicles. In Proceedings of the Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021. Springer, 2022, pp. 123–136.
90. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects, 2019. arXiv:1812.04608 [cs].
91. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Towards better analysis of deep convolutional neural networks. *International Conference on Learning Representations (ICLR)* **2015**.
92. Chowdhary, K.; Chowdhary, K. Natural language processing. *Fundamentals of Artificial Intelligence* **2020**, pp. 603–649.
93. Wang, Z.; Yan, W.; Oates, T. Time series classification from scratch with deep neural networks: A strong baseline. In Proceedings of the 2017 International joint Conference on Neural Networks (IJCNN). IEEE, 2017, pp. 1578–1585.
94. Dai, J.; Upadhyay, S.; Aivodji, U.; Bach, S.H.; Lakkaraju, H. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In Proceedings of the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 203–214.
95. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
96. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
97. Huber, T.; Weitz, K.; André, E.; Amir, O. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* **2021**, *301*, 103571.
98. Anguita-Ruiz, A.; Segura-Delgado, A.; Alcalá, R.; Aguilera, C.M.; Alcalá-Fdez, J. eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Computational Biology* **2020**, *16*, e1007792.
99. Ye, Y.; Zhang, X.; Sun, J. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. *Transportation Research Part C: Emerging Technologies* **2019**, *107*, 155–170.
100. Chakraborty, S.; Tomsett, R.; Raghavendra, R.; Harborne, D.; Alzantot, M.; Cerutti, F.; Srivastava, M.; Preece, A.; Julier, S.; Rao, R.M.; et al. Interpretability of deep learning models: A survey of results. In Proceedings of the 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI). IEEE, 2017, pp. 1–6.
101. Heuillet, A.; Couthouis, F.; Díaz-Rodríguez, N. Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Computational Intelligence Magazine* **2022**, *17*, 59–71.
102. Ward, A.; Sarraju, A.; Chung, S.; Li, J.; Harrington, R.; Heidenreich, P.; Palaniappan, L.; Scheinker, D.; Rodriguez, F. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digital Medicine* **2020**, *3*, 125.
103. Li, J.; King, S.; Jennions, I. Intelligent Fault Diagnosis of an Aircraft Fuel System Using Machine Learning—A Literature Review. *Machines* **2023**, *11*, 481.
104. Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; et al. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403* **2024**.
105. Jain, S.; Wallace, B.C. Attention is not explanation. *arXiv preprint arXiv:1902.10186* **2019**.
106. Ribeiro, M.T.; Singh, S.; Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* **2016**.
107. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys* **2023**, *55*, 1–42. https://doi.org/10.1145/3583558.
108. Fuhrman, J.D.; Gorre, N.; Hu, Q.; Li, H.; El Naqa, I.; Giger, M.L. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics* **2022**, *49*, 1–14.

109. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors* **2023**, *23*, 634.
110. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y.A.; Gomaa, M.M.; Hassanien, A.E. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review* **2023**, *56*, 5975–6037.
111. El-Sappagh, S.; Alonso, J.M.; Islam, S.R.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports* **2021**, *11*, 2660.
112. Loh, H.W.; Ooi, C.P.; Seoni, S.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of Explainable Artificial Intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine* **2022**, p. 107161.
113. Zhou, X.; Tang, J.; Lyu, H.; Liu, X.; Zhang, Z.; Qin, L.; Au, F.; Sarkar, A.; Bai, Z. Creating an authoring tool for K-12 teachers to design ML-supported scientific inquiry learning. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–7.
114. Kim, T.S.; Lee, Y.; Shin, J.; Kim, Y.H.; Kim, J. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In Proceedings of the Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–21. arXiv:2309.13633 [cs], https://doi.org/10.1145/3613904.3642216.
115. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable Artificial Intelligence: a comprehensive review. *Artificial Intelligence Review* **2022**, pp. 1–66.
116. Feng, J.; Lansford, J.L.; Katsoulakis, M.A.; Vlachos, D.G. Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. *Science advances* **2020**, *6*, eabc3204.
117. Shumway, R.H.; Stoffer, D.S.; Stoffer, D.S. *Time series analysis and its applications*; Vol. 3, Springer, 2000.
118. Lipton, Z.C.; Kale, D.C.; Wetzel, R.; et al. Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare* **2016**, *56*, 253–270.
119. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* **2014**, *15*, 3221–3245.
120. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
121. Mankodiya, H.; Obaidat, M.S.; Gupta, R.; Tanwar, S. XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles. In Proceedings of the 2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). IEEE, 2021, pp. 1–5.
122. Fan, F.L.; Xiong, J.; Li, M.; Wang, G. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences* **2021**, *5*, 741–760.
123. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* **2021**, *296*, 103473.
124. Dam, H.K.; Tran, T.; Ghose, A. Explainable software analytics. In Proceedings of the Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, 2018, pp. 53–56.
125. Awotunde, J.B.; Adeniyi, E.A.; Ajamu, G.J.; Balogun, G.B.; Taofeek-Ibrahim, F.A. Explainable Artificial Intelligence in Genomic Sequence for Healthcare Systems Prediction. In *Connected e-Health: Integrated IoT and Cloud Computing*; Springer, 2022; pp. 417–437.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.