XAI.it 2024: An Overview on the Future of AI in the era of Large Language Models

Marco Polignano[1], Cataldo Musto[1], Roberto Pellungrini[2], Erasmo Purificato[3,‡], Giovanni Semeraro[1] and Mattia Setzu[2]

[1]University of Bari Aldo Moro, Italy
[2]Scuola Normale Superiore, Italy
[3]Joint Research Centre, European Commission, Ispra, Italy

Abstract

The rapid development and deployment of Large Language Models (LLMs) has the potential to transform numerous industries and aspects of our lives, from natural language processing and text generation to customer service and decision-making. However, as these models become increasingly sophisticated and pervasive, the need for AI (XAI) to ensure the transparency, interpretability, and trustworthiness of their outputs has grown more pressing. This work discusses the current state and future directions of XAI in LLMs, highlighting the challenges and opportunities in developing techniques that can handle the massive scale and complexity of modern LLMs and exploring the potential for XAI to revolutionize the way we interact with and rely on LLMs in the future. As LLMs are increasingly used to make decisions, generate content, and provide information, the lack of transparency and interpretability in their decision-making processes can have far-reaching consequences, including the potential for bias, misinformation, and harm. XAI in LLMs is essential to address these concerns, providing a means to understand the reasoning and decision-making processes behind the outputs of these models. XAI.it 2024 focused on these issues and provided a space to discuss them with the international scientific community during the annual AixiA conference focusing on new challenges and research perspectives in Artificial Intelligence.

Keywords

AI, Biases, Trustworthiness, Large Language Models, LLMs, XAI

1. Introduction

As Artificial Intelligence (AI) continues to transform the way we live and work, the importance of AI (XAI) has become increasingly evident. The widespread adoption of AI in various industries and aspects of our lives has raised a pressing need for transparency, accountability, and trust in the decision-making processes of AI systems. The lack of explainability in AI's decision-making can have far-reaching consequences, including the potential for bias, discrimination, and harm. The use of AI in high-stakes applications, such as healthcare, finance, and law enforcement, only amplifies the imperative for XAI. In recent years, high-profile examples of AI's unintended consequences have garnered significant attention, from the biased hiring algorithms that perpetuate discrimination to the autonomous vehicles that malfunction due to uninterpretable decisions. These incidents have highlighted the need for a fundamental shift in the way we approach AI development, prioritizing the accuracy and performance of

XAI.it - 5th Italian Workshop on Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024

‡The author contributed to this work while affiliated with Otto von Guericke University Magdeburg, Germany. The view expressed in this paper is purely that of the author and may not, under any circumstances, be regarded as an official position of the European Commission.

†All authors contributed equally.

marco.polignano@uniba.it (M. Polignano); cataldo.musto@uniba.it (C. Musto); roberto.pellungrini@sns.it (R. Pellungrini); erasmo.purificato@acm.org (E. Purificato); giovanni.semeraro@uniba.it (G. Semeraro); mattia.setzu@unipi.it (M. Setzu)
https://marcopoli.github.io/ (M. Polignano); https://swap.di.uniba.it/members/musto.cataldo/ (C. Musto); https://kdd.isti.cnr.it/people/pellungrini-roberto (R. Pellungrini); https://erasmopurif.com/ (E. Purificato); https://swap.di.uniba.it/members/semeraro.giovanni/ (G. Semeraro); hhttps://kdd.isti.cnr.it/people/setzu-mattia (M. Setzu)
0000-0002-3939-0136 (M. Polignano); 0000-0001-6089-928X (C. Musto); 0000-0002-1366-9833 (R. Pellungrini); 0000-0002-5506-3020 (E. Purificato); 0000-0001-6883-1853 (G. Semeraro); 0000-0001-8351-9999 (M. Setzu)

AI systems and their transparency and accountability. In decision-making, the absence of transparency and accountability can have devastating effects. For instance, an AI system that makes life-or-death decisions, such as a medical diagnosis or a self-driving car's braking decision, without providing a clear understanding of its reasoning and decision-making process can lead to catastrophic consequences. Similarly, an AI system that makes hiring or lending decisions without being able to explain its criteria can perpetuate discrimination and inequality. XAI is not only a moral imperative but also a business necessity, as organizations that fail to provide explainable AI in their decision-making processes may face reputational damage, legal liability, and even regulatory non-compliance. In this context, XAI is not just a technical challenge but a critical component of the responsible development and deployment of AI in the modern era. As we continue to push the boundaries of AI's capabilities, it is essential that we also prioritize the development of XAI techniques that can provide a clear understanding of AI's decision-making processes, thereby ensuring the trust, accountability, and transparency that are essential for the responsible use of AI in decision-making.

The unique challenges of XAI in LLMs arise from the fact that LLMs process and generate human language, which is inherently complex, nuanced, and context-dependent. LLMs' outputs are often the result of intricate interactions between the model's architecture, the input text, and the training data. This complexity makes it essential to develop XAI techniques that can provide a deep understanding of the decision-making processes of LLMs, particularly in high-stakes applications where the outputs of LLMs can have significant consequences. The application of XAI in LLMs is a new frontier in explainability, requiring the development of novel techniques that can handle the massive scale and complexity of modern LLMs. Researchers and practitioners must address the challenges of XAI in LLMs, including the need for scalable and efficient explainability methods, the development of task-oriented XAI techniques, and the integration of XAI with human-in-the-loop approaches to ensure the accuracy and trustworthiness of LLMs' outputs. Moreover, the lack of explainability in LLMs can also lead to the emergence of hallucinations and biases in their outputs. Hallucinations, in the context of LLMs, refer to the model's ability to generate text that is not supported by the input or training data, often due to the model's overfitting or the presence of adversarial examples. Biases, on the other hand, can manifest in the form of discriminatory or stereotypical language, as a result of the model's training data being imbalanced or biased. The absence of XAI in LLMs can make it difficult to detect and mitigate these issues, ultimately leading to the deployment of biased or hallucinatory models in high-stakes applications.

In the forthcoming discussion of the topic, we will engage in a nuanced examination of the issue, informed by the latest advancements in the field of AI, as presented in the papers accepted at the XAI.it 2024 workshop. By taking into account the novel approaches and methodologies showcased at the XAI.it 2024 workshop, we will delve into the topic and offer a perspective, one that not only synthesizes our current point of view but also anticipates the future directions in the development of more transparent and Interpretable AI systems in the international research context.

2. Related Work

Already in 2016, Ribeiro et al. [1] introduced LIME (Local Interpretable Model-Agnostic Explanations), an approach that allows for the explanation of individual predictions made by any machine learning classifier, regardless of its complexity. The goal of LIME is to make black-box models interpretable by providing insights into the local behavior of a model around a specific instance. In 219, Gunning et al. [2] introduced the topic of AI (XAI), defining a XAI system as *"able to explain its capabilities and understandings; explain what it has done, what it is doing now, and what will happen next; and disclose the salient information that it is acting on"*. The purpose of XAI system is to make its behavior more intelligible to humans by providing explanations. To create more understandable AI systems, there are general principles to follow, such as the ability to explain its capabilities and understandings, what it has done, what it is doing, and what will happen next, as well as revealing the salient information it is acting on. However, every explanation is context-dependent, based on the task, abilities, and

expectations of the user of the AI system. The definitions of interpretability and explainability are, thus, domain-dependent and may not be defined independently of a domain. Furthermore, users have diverse expectations based on their role within the system, for example, an intelligence analyst, a judge, or an operator, and may require different types of explanations. Since that time, a large amount of research has gone in this direction. In particular the issue of Deep Neural Network (DNNs) as **"black boxes"** arises [3]. DNNs can have many layers and numerous filters and units, making it difficult to understand the data flow and representations within the network. With millions or even billions of parameters, the complexity of DNNs increases the number of learnable variables, making it challenging to comprehend the decision-making process [4]. The design of a DNN is influenced by various factors, including activation functions, network architecture, and learning techniques, which in turn are affected by additional functions like normalization and cost functions. Due to this complexity, DNNs are often considered "black boxes", making it hard to trust and understand the decisions they make, a major problem in the field of machine learning [5]. Moreover, two crucial concepts in the field of Deep Neural Networks emerge: i.e., **interpretability** and **explainability** [6]. Interpretability enables developers to delve into the model's decision-making process, thereby boosting their confidence in understanding where the model gets its results. This concept provides an interface that gives additional information or explanations, which are essential for interpreting an AI system's underlying functioning. It allows developers, who possess the required knowledge and skills, to gain insight into the internal workings of the model, effectively "opening a door" into the black-box model.

In contrast, explainability is about providing insight into the DNN's decision to the end-user, with the goal of building trust in the model's correctness and non-biased decision-making. The end-user, in this context, is not necessarily a technical expert, but a person who needs to understand and trust the AI system's outputs. A trustworthy model achieves a good balance between interpretability and accuracy, and that AI (XAI) is a means to achieve this balance by providing explanations to the end-user.

The growing attention to XAI across multiple domains has led to a surge in the development of novel methods and techniques in both industry and academia. With a diverse range of features and capabilities, from basic data exploration to complex AI model understanding, the current XAI systems present a multitude of options. In order to select the most suitable approach, it is crucial to understand the fundamental differences and characteristics of the various XAI methods. A comprehensive analysis of the most recent approaches for XAI is proposed in [7]. In particular, explainability can involve data, models, providing explainations based on features or examples. A crucial aspect of understanding a model's interpretability is its model-specific nature. This refers to the need for techniques tailored to a particular model, which involve dissecting the model's internal workings, including its intermediate processes and structures (i.e., *Model-specific techniques*). In contrast, model-agnostic techniques focus on the model's inputs, outputs, and the underlying data, seeking to uncover the relationships between these components (i.e., *Model-agnostic techniques*). Model-specific techniques are usually directly implemented into the design of the model architecture, as for an example the xDNN classifier proposed in [8]. Among Model-agnostic techniques, LIME (Local Interpretable Model-agnostic Explanations) [9] and SHAP (i.e., Shapley additive explanations) [10] are the most widely used. Unlike attempting to decipher the entire model, LIME takes a more targeted approach. By subtly altering a data sample and monitoring the effect on the prediction, LIME strives to grasp the model's behavior around that specific point. This local model interpretability is particularly relevant, as it's often what a user is interested in when examining the output of a model and how it applies to a particular case rather than the model as a whole. SHAP quantifies the contribution of each input feature to the final prediction. It is based on Shapley values, a concept from cooperative game theory, which assigns an "importance" value to each player in a game based on their contribution to the overall outcome. In the context of machine learning, each "player" is an input feature in the model, and the "game" is the prediction task. SHAP has become popular because it offers a unified approach to model interpretability that can be applied to a range of models, from linear regressions to deep neural networks. It provides insight into the "black-box" behavior of complex models like ensemble methods (e.g., Random Forest, XGBoost) and neural networks. However, it is computationally expensive to calculate the Shapley value in comparison to LIME. Unfortunately, these approaches are often not enough for a thrustable AI system. Depending

on their roles, expertise, and goals, different users require tailored explanations to understand a system's functionality and decision-making effectively. This notion of contextualized explanations is especially important in AI, Machine Learning, NLP, and LLMs where the generated outputs are often derived from complex algorithms, commonly viewed as "black boxes".

3. XAI and LLMs Research Trends

When incorporating XAI techniques into Large Language Models, a trade-off may arise between the model's interpretability and its performance. A straightforward example of this dilemma is the choice between using a transparent, yet less accurate, rule-based language generator, and a highly effective, yet opaque, large language model, capable of generating human-like text. Simple models, such as rule-based language generators, can easily reveal the decisions behind their text generations, but their capabilities are limited by their simplicity. On the other hand, complex Large Language Models (LLMs) often excel in performance, generating coherent and natural-sounding text, yet their internal workings, such as the decision-making process behind a generated sentence, are notoriously difficult to understand, rendering them effectively "black boxes" in the context of interpretability. LLMs lack true language understanding due to their subsymbolic nature. Indeed, LLMs treat all text equally, lacking the mechanisms to differentiate factual from non-factual information. Since these models encode knowledge as intricate patterns in their weights, their understanding is inaccessible and non-symbolic, rendering their "knowledge" is uninterpretable and challenging to reason with. Saba in [11], advocates for an alternative approach that incorporates symbolic, explainable, and ontologically grounded models. LLMs should be combined with symbolic systems, which could address critical linguistic challenges where stochastic models fall short. This approach could improve reasoning capabilities and allow models to handle complex linguistic phenomena, such as ambiguity and scope distinctions in language, by integrating structured, meaningful representations instead of relying solely on pattern matching. A symilar approach grounded on the concept of **Symbolic AI** is also proposed by Sullivan et al. [12] and Acharya et al. [13].

The trustability of an AI model go through the validity of the output provided and in LLMs this quality is not always guarantee. The presence of *hallucinations* in LLMs generated contents is one of the major concern about the use of such technology in risky environments. As elaborated previously, a Large Language Model processes information in a fundamentally different manner from human thought. Rather than being guided by logical conclusions, it generates text by predicting the likelihood of word sequences in a given context, mirroring the patterns it learned from its training data. The model does not possess a mechanism to verify the accuracy of its generated text, as the information it produces is a result of its training data, which may have been incorrect to begin with. Notably, the model was not explicitly trained to convey uncertainty or acknowledge its limitations of knowledge. To better understand this concept, we can liken the model's "thought process" to an automated, rapid, and instinctual operation, devoid of deliberate mental steps. In [14] the authors analyze common types of hallucinations in responses generated by various LLMs, including models fine-tuned for medical purposes like MedAlpaca and Robin-medical. The study identifies three main types of hallucinations: Fact Inconsistency, where responses contradict known facts; Query Inconsistency, where responses are unrelated to the question asked; and Tangentiality, where responses are somewhat related to the question but do not directly answer it. To address this, the authors propose an innovative self-reflection methodology that involves an iterative, feedback-driven process in which the model evaluates and refines its own outputs. This **Self-Reflection** approach consists of three loops: Factual Knowledge Acquiring Loop: the model initially generates background information relevant to the query, which is then evaluated for factuality; Knowledge-Consistent Answering Loop: the model generates an answer based on verified background knowledge, assessing and refining the consistency between the generated answer and the background information; Question-Entailment Answering Loop: this final loop checks whether the answer logically addresses the query itself, refining it if necessary. This iterative process continues until the response achieves high levels of factuality, consistency, and entailment. The model

is thus guided to self-correct by returning to previous loops when a threshold of accuracy or relevance is not met. A similar approach to mitigate the hallucination issue is provided by Piché [15] and Shinn [16] that introduced his concept of **Reflexion**.

Moreover, traditional automated metrics to evaluate the quality of LLM contents like BLEU, ROUGE, and METEOR are effective for structured outputs, but they often fail to capture the nuanced semantics required for free-form, open-ended responses. This shortcoming is especially critical with LLMs, which produce diverse responses that can all be valid, despite differences in wording and structure. In [17], Badshan et al. propose an LLM-driven approach where multiple LLMs are used as evaluators, or **Judges**, to provide a verdict on the quality of generated text based on context and reference answers. This approach introduces a system in which candidate model responses, reference answers, and the input prompt are evaluated by multiple LLMs to generate a verdict. These "judges" make a determination by assessing alignment with the reference answer while considering the context of the input prompt. This setup mirrors human evaluation processes by leveraging multiple judges, ensuring that diverse perspectives are captured in the verdict. The study assessed the alignment between LLM and human evaluations using metrics such as percent agreement, Fleiss's kappa, and Cohen's kappa. Findings showed that the agreement was highest with the combined judgments of multiple LLMs, which closely matched human evaluations. Individual LLMs, while reliable, showed lower consistency compared to aggregated judgments across multiple models. This finding underscores the importance of combining multiple LLMs to reduce biases and improve alignment with human evaluators.

4. XAI.it Contributions

The contributions received for the XAI.it 2024 workshop offer a comprehensive overview of applications, challenges, and emerging methodologies in the field of AI (XAI), making them particularly relevant for the scientific community.

***A Comprehensive Strategy to Bias and Mitigation in Human Resource Decision Systems* [18].** The article contributes significantly to the topic of XAI. D'Amicantonio et al., explore the intersection of *AI, bias, and transparency* within **Human Resource (HR) systems**, emphasizing the necessity for *explainability in AI-driven decision-making processes.*

One of the primary contributions of the article is its **detailed analysis of the sources of bias in HR decision systems**, which is crucial for understanding how these biases can affect the fairness of AI applications. The authors categorize biases arising from non-representative and outdated training datasets, as well as from algorithmic limitations that fail to account for context-specific requirements. This categorization is summarized in Table 1, which outlines various sources of bias alongside corresponding mitigation techniques. For instance, the authors suggest expanding dataset sources and implementing blind recruitment practices to reduce unconscious bias. These strategies are essential for ensuring that AI systems operate on a foundation of fairness and equity, which aligns with the goals of XAI to make AI systems more interpretable and accountable. Furthermore, the article discusses the importance of knowledge sharing between AI developers and HR professionals as a means of enhancing the performance of recruitment models. This collaborative approach is vital for developing AI systems that not only perform well but also adhere to ethical standards.

In terms of experimental results, the article highlights the **effectiveness of various bias mitigation strategies through empirical evidence**. The authors advocate for independent audits and periodic assessments of AI algorithms to detect biases and ensure ongoing fairness. This approach is supported by findings that suggest transparency in audit results can foster trust among users and stakeholders, a principle that is central to the XAI framework. The emphasis on transparency and accountability in AI systems is particularly relevant in the context of HR, where decisions can significantly impact individuals' careers and lives. Moreover, the article outlines future directions for research in bias mitigation and XAI, suggesting that more complex and realistic datasets should be explored to better reflect the diversity of the population. This recommendation aligns with the workshop's focus on

advancing the field of XAI by addressing real-world challenges and improving the interpretability of AI systems. The empirical evaluation conducted by the authors reveals that **no single model can fully satisfy all fairness metrics**, highlighting the complexity of achieving both high performance and fairness in AI systems. This finding is particularly relevant for the XAI community, as it emphasizes the need for ongoing research into model architectures that can better balance these competing objectives..

Their findings and recommendations not only advance the understanding of bias in AI but also pave the way for more equitable and explainable AI systems in the future.

***An Analysis on How Pre-Trained Language Models Learn Different Aspects* [19].** The article addresses the **interpretability of Neural Language Models** (NLMs) through a *systematic exploration of NLMs layers capabilities during probing tasks*, which serves as a means to evaluate the linguistic performances of these models. The proposed approach is in line with current state of the art literature grounded on different observation angles: by analyzing self-attention weights to find relations among words, by determining whether NLMs have acquired specific world knowledge, or by investigating their linguistic capabilities. The authors emphasize the importance of understanding how these models acquire knowledge and the mechanisms behind their predictions, which is a central and timely theme in XAI in the era of LLMs.

The authors present a series of experimental results that highlight the models' performance across various probing tasks. These tasks are designed to assess different aspects of language understanding, including *grammatical correctness* and *semantic comprehension*. For instance, five distinct tasks are studied: *Causative*, *Coordinate Structures*, *Passive*, *Mix*, and *Humor*. Each task is accompanied by a dataset that has been adapted from the BLiMP benchmark, allowing for a robust evaluation of the models. Three GPT-NeoX models belonging to the Pythia benchmark suite have been used. The experimental results reveal that while *NLMs exhibit a strong grasp of basic syntactic features early in their training*, more complex semantic understanding, such as humor recognition, requires a more nuanced approach. The Mix task exhibits very low compression, indicating a limited ability of a neural language model (NLM) to determine the correctness of a sentence in general, rather than focusing on a single, specific aspect. The analysis of the learning trajectories showed that most of this general grammatical knowledge is acquired early in training, with compression generally remaining stable or, for the final layer only, slightly decreasing. This decline in performance is likely due to the specialization of the final layer on the Masked Language Modeling task for which the model is primarily trained. Moreover, it has been observed overall, the **Middle layers** achieved the *best results*, while the Bottom layers yielded the lowest performance, particularly on grammar-related tasks.

The article's contributions to XAI are multifaceted, offering valuable insights into the interpretability of NLMs through rigorous analysis of Layers performances during the training of NLMs as a means to evaluate and explain the capabilities.

***Ethical AI Systems and Shared Accountability: The Role of Economic Incentives in Fairness and Explainability* [20].** The article contributes significantly to the topic of XAI by addressing the critical intersection of **ethical alignment and transparency in AI systems**. The authors present a robust framework that integrates economic modeling with ethical considerations, thereby enhancing our understanding of how to govern AI systems effectively. This is particularly relevant as AI technologies become increasingly autonomous and integrated into various societal functions.

One of the key contributions of the article is its exploration of the principal-agent problem as it pertains to AI alignment. The authors draw parallels between traditional economic theories and the challenges faced in AI development, particularly the *misalignment of incentives* between *developers* (agents) and *users* (principals) (**Principal-Agent problem**). This misalignment can lead to ethical dilemmas, where the objectives of developers may not fully align with the ethical expectations of users. The article posits that by structuring contracts that clearly delineate responsibilities and incentives, it is possible to foster a more ethical approach to AI development.

The authors highlight how different levels of risk aversion among developers influence their willing-

ness to adhere to ethical guidelines. The results indicate that contracts that incorporate performance-based incentives lead to a higher degree of ethical compliance, as developers are motivated to align their outputs with the ethical standards set forth by users. Moreover, they argue that when developers are held accountable through well-structured contracts, it not only enhances compliance with ethical standards but also builds trust among users. This is particularly crucial in high-stakes applications such as autonomous vehicles and healthcare, where the consequences of misaligned AI behavior can be severe. The authors suggest that future research should explore more complex incentive structures that account for these dynamic factors, thereby enriching the dialogue on ethical AI governance and emphasizing the need for collaborative efforts among developers, users, and regulators to ensure that AI technologies are developed and deployed responsibly.

***ExplainBattery: Enhancing Battery Capacity Estimation with an Efficient LSTM Model and Explainability Features* [21].** The article presents significant contributions in the context of Battery Management Systems (BMS) for lithium-ion battery capacity prediction. The authors emphasize the importance of transparency and **interpretability in machine learning models**, especially in *critical applications* such as Prognostic and Health Management (PHM). This focus on explainability is crucial, as it fosters **trust** in the model's predictions and decision-making processes, which is a central theme in XAI discussions.

The salient contribution of the article is the development of a novel **LSTM-based model** that not only *enhances prediction accuracy* but also *reduces the complexity of the neural architecture*. The experimental results demonstrate that this model outperforms existing state-of-the-art models, achieving substantial improvements across various evaluation metrics. For instance, it shows a reduction of 46.45% in Mean Squared Error (MSE), 21.21% in Root Mean Squared Error (RMSE), 13.59% in Mean Absolute Error (MAE), and 35.86% in Mean Absolute Percentage Error (MAPE). These metrics are critical in assessing the performance of predictive models, and the reported improvements highlight the effectiveness of the proposed approach. Moreover, the authors provide a detailed analysis of the model's efficiency, noting a *75.67% decrease in trainable parameters compared to previous models*. This reduction in complexity is particularly relevant in the context of battery management systems, where computational resources may be limited. The trade-off between model complexity and performance is a recurring theme in XAI, as simpler models are often preferred for their interpretability and ease of deployment.

The article also introduces the **ExplainBattery Web Application**, which serves as a practical tool for users to interact with the LSTM model and explore the underlying data. This application is designed to facilitate the visualization of battery capacity predictions and the investigation of the model's decision-making process. By incorporating explainability techniques such as SHAP (SHapley Additive exPlanations) and Saliency Maps, the application allows users to gain insights into the influence of various features on the model's predictions. This aligns with the workshop's focus on developing methods that enhance the interpretability of AI systems.

By advancing the state of the art in battery capacity prediction through an efficient and accurate LSTM model, and by providing a user-friendly application that emphasizes explainability, the authors address critical challenges in the deployment of AI systems in safety-sensitive domains. Their work not only enhances the understanding of model behavior but also promotes the development of trustworthy AI solutions and their real-world implementations, making this research particularly relevant to ongoing discussions in the field.

***Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps* [22].** The authors address a critical challenge in the realm of **AI art classification**: *the opacity of decision-making processes* in deep learning models. By integrating advanced techniques such as Grad-CAM with Large Language Models (LLMs), the research aims to enhance the interpretability of AI systems, making them more accessible and understandable to non-expert users.

One of the key contributions of the article is the proposed framework that combines **visual explanations from Grad-CAM with textual descriptions generated by LLMs**. This dual approach not only

provides visual insights into which areas of an artwork influenced the model's classification but also offers coherent and relevant textual explanations that articulate the reasoning behind these decisions. The integration of these two modalities is particularly important in the context of art classification, where the subtleties of artistic style and composition can be complex and nuanced.

The experimental results presented in the article are particularly noteworthy. The authors conducted a comprehensive evaluation using a dataset of 100 images, evenly split between AI-generated and original artworks. This dataset was carefully curated from larger repositories, ensuring a diverse representation of artistic styles and genres. The experiments were designed to assess both qu*antitative and qualitative aspects of the models' performance*. In the quantitative analysis, the authors employed two primary metrics: image-to-text similarity and text-to-label similarity. The image-to-text similarity was measured using the CLIP model, which computes cosine similarity between the Grad-CAM overlay and the generated textual description. A higher score indicates a better alignment between the visual content and the generated text. The text-to-label similarity was assessed using the S-BERT model, which evaluates the consistency between the generated text and the classification label. *These metrics provide a robust framework for evaluating how well the LLMs can generate explanations* that are not only relevant but also reflective of the visual content.

The qualitative analysis involved a manual examination of the generated explanations, focusing on their coherence, relevance, and insightfulness. The authors compared the descriptions with the **Grad-CAM heatmaps** to determine whether the explanations **provided meaningful insights into the model's decision-making process**. This qualitative assessment is crucial, as it allows for a deeper understanding of how well the LLMs can articulate the reasoning behind the model's focus on specific areas of the artwork. The results of the experiments revealed that the selected LLMs—LLaVa-NeXt, InstructBLIP, and KOSMOS-2 demonstrated encouraging effectiveness in generating coherent and insightful explanations. Overall, the article contributes to the ongoing discourse in XAI by demonstrating how the integration of visual and textual explanations can enhance the interpretability of AI models.

Probabilistic Abstract Interpretation on Neural Networks via Grids Approximation **[23]** The authors, Zhuofan Zhang and Herbert Wiklicky, focus on e**nhancing the interpretability of neural networks** *through the lens of probabilistic abstract interpretation*, a method that allows for a deeper understanding of how neural networks process inputs and make predictions.

One of the key contributions of the paper is the introduction of a novel framework that utilizes grid approximation to **analyze the density distribution of input spaces in neural networks**. By employing a probabilistic framework, the authors aim to extract meaningful insights about the behavior of neural networks, which is crucial for building trust and transparency in AI systems.

In their experiments, the authors demonstrate the application of probabilistic abstract interpretation to a digit classification task using the MNIST dataset. They provide a comprehensive evaluation of the model's performance under various conditions, including adversarial attacks. The results indicate that the *probabilistic approach not only enhances the robustness of the model against such attacks but also provides a clearer picture of the probability density flow of inputs towards predictions.* This aspect is particularly noteworthy, as it highlights the dual purpose of the method: it extracts features of the network while simultaneously illustrating how different inputs influence the model's decisions. The authors acknowledge that while their approach shows promise, there are still challenges to address, particularly in scaling the method to more complex neural network architectures and exploring additional abstract domains beyond grid approximations.

The insights gained from this work are poised to influence ongoing discussions at the XAI.it 2024 workshop, where the importance of interpretability in AI continues to gain traction.

5. Conclusion

In this overview, we comprehensively examined the evolving landscape of eXplainable AI (XAI) in the context of Large Language Models (LLMs). We highlights the transformative potential of LLMs across

various sectors, emphasizing their applications in natural language processing, decision-making, and customer service. At the same time, we focus on the critical need for transparency and interpretability in AI systems, particularly as these technologies become increasingly autonomous and integrated into societal functions.

The analysis we presented reveals that while LLMs exhibit remarkable capabilities, they also pose significant challenges, including issues related to bias, misinformation, and ethical alignment. The exploration of learning trajectories within neural models illustrates the complexities of achieving both performance and fairness, indicating that no single model can adequately satisfy all fairness metrics. Furthermore, the importance of structuring economic incentives and accountability mechanisms to foster ethical AI development, thereby enhancing user trust, has been discussed.

A common trend found in ongoing research in XAI moves forward model architectures and structures that can better balance competing objectives of performance, fairness, and explainability. The discussion contributes to the broader discourse on XAI, emphasizing the necessity for collaborative efforts among developers, users, and regulators to ensure responsible AI deployment.

The overarching theme emphasizes that as AI models grow in scale and sophistication, so too must our capacity to interpret, understand, and ethically align their decision-making processes. The approaches discussed within the XAI.it 2024 workshop contributions reflect the urgent need for scalable, nuanced, and practical XAI methods capable of enhancing transparency across various AI applications. By focusing on interpretability, accountability, and trustworthiness, this paper ultimately reinforces the necessity of XAI as an integral element in the responsible deployment and future evolution of AI systems. We hope this overview serves as a foundational reference for future investigations into the intersection of LLMs and XAI, paving the way for advancements prioritizing ethical considerations and user trust in AI technologies.

Acknowledgments

This research is partially funded by PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] M. T. Ribeiro, S. Singh, 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.

[2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science robotics 4 (2019) eaay7120.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[4] Y. Liang, S. Li, C. Yan, M. Li, C. Jiang, Explaining the black-box model: A survey of local interpretation methods for deep neural networks, Neurocomputing 419 (2021) 168–182.

[5] V. Buhrmester, D. Münch, M. Arens, Analysis of explainers of black box deep neural networks for computer vision: A survey, Machine Learning and Knowledge Extraction 3 (2021) 966–989.

[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 80–89.

[7] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[8] P. Angelov, E. Soares, Towards explainable deep neural networks (xDNN), Neural Networks 130 (2020) 185–194.

[9] S. Mishra, B. L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis, in: ISMIR, volume 53, 2017, pp. 537–543.
[10] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).
[11] W. S. Saba, Stochastic LLMs do not understand language: towards symbolic, explainable and ontologically based LLMs, in: International Conference on Conceptual Modeling, Springer, 2023, pp. 3–19.
[12] R. Sullivan, N. Elsayed, Can Large Language Models Act as Symbolic Reasoners?, arXiv preprint arXiv:2410.21490 (2024).
[13] K. Acharya, A. Velasquez, H. H. Song, A survey on symbolic knowledge distillation of large language models, IEEE Transactions on Artificial Intelligence (2024).
[14] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating LLM hallucination via self reflection, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 1827–1843.
[15] A. Piché, A. Milios, D. Bahdanau, C. Pal, LLMs can learn self-restraint through iterative self-reflection, arXiv preprint arXiv:2405.13022 (2024).
[16] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, Advances in Neural Information Processing Systems 36 (2024).
[17] S. Badshah, H. Sajjad, Reference-Guided Verdict: LLMs-as-Judges in Automatic Evaluation of Free-Form Text, arXiv preprint arXiv:2408.09235 (2024).
[18] S. D'Amicantonio, M. K. Kulangara, H. D. Mehta, S. Pal, M. Levantesi, M. Polignano, E. Purificato, E. W. De Luca, A Comprehensive Strategy to Bias and Mitigation in Human Resource Decision Systems, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.
[19] E. Gjinika, N. Arici, L. Putelli, A. E. Gerevini, I. Serina, An Analysis on How Pre-Trained Language Models Learn Different Aspects, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.
[20] D.-H. Yoo, C. Giannetti, Ethical AI Systems and Shared Accountability: The Role of Economic Incentives in Fairness and Explainability, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.
[21] L. Dal Ronco, E. Purificato, ExplainBattery: Enhancing Battery Capacity Estimation with an Efficient LSTM Model and Explainability Features, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.
[22] G. Castellano, M. G. Miccoli, R. Scaringi, G. Vessio, G. Zaza, Using LLMs to explain AI-generated art classification via Grad-CAM heatmaps, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.
[23] Z. Zhang, H. Wiklicky, Probabilistic Abstract Interpretation on Neural Networks via Grids Approximatio, in: Proceedings of 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence, Bolzano, Italy, November 25-28, 2024, CEUR. org, 2024.