Why Explainability Matters for Large Foundation Models in AI Systems

Robin Bashemath[1], Zhihao Ru[1],

Department of Computer Science, Chinese University of Hong Kong, Hong Kong
robin.bashemath@cuhk.edu.hk, zhihao.ru@cuhk.edu.hk

Abstract. In recent years, large foundation models, such as GPT-3, BERT, and other transformer-based architectures, have achieved state-of-the-art performance across a wide range of artificial intelligence tasks. However, their complexity and size pose significant challenges in terms of transparency, interpretability, and trust. As these models are deployed in high-stakes domains such as healthcare, finance, and law enforcement, understanding their decision-making process is crucial to ensure accountability and ethical use. This paper explores the growing need for explainable AI (XAI) in the era of large foundation models, focusing on the challenges, existing methods, and emerging trends in XAI research. We discuss state-of-the-art attribution techniques, model-agnostic approaches, and methods for visualizing and interpreting attention mechanisms and embeddings. Additionally, we highlight promising future directions, including the development of self-explainable models, multimodal explainability, and the integration of human-in-the-loop frameworks. By advancing the explainability of large models, we aim to foster greater trust in AI systems and ensure that they are both powerful and transparent, with a strong focus on fairness and ethical considerations.

Keywords: Explainable AI, large foundation models, interpretability, transparency, GPT-3, BERT, transformer models, attribution methods, SHAP, LIME, multimodal AI, attention mechanisms, self-explainable models, human-in-the-loop, fairness, accountability, ethical AI.

1 Introduction

The advent of large foundation models, such as GPT-3, BERT, and DALL·E, has led to a significant paradigm shift in the field of artificial intelligence (AI). These models, often trained on vast amounts of data and possessing billions or even trillions of parameters, have demonstrated unprecedented capabilities in a wide range of tasks, including natural language processing (NLP), computer vision, and more [1]. Their success has been driven by the ability to scale up both the size of models and the data they are trained on, resulting in the emergence of highly capable AI systems that can solve problems once thought to be uniquely human. However, the increasing complexity and opacity of these large foundation models have given rise to a number of concerns, especially in terms of transparency,

trustworthiness, and accountability [2]. In the context of AI systems, the term "explainability" refers to the ability of a model to provide understandable and interpretable justifications for its decisions, actions, or outputs [3]. The need for explainable AI (XAI) has been emphasized by researchers, practitioners, and regulators, particularly as AI is increasingly deployed in high-stakes domains such as healthcare, finance, and criminal justice [4]. In these settings, understanding why an AI system made a particular decision is crucial to ensuring that the system is not only effective but also fair, ethical, and aligned with human values. Furthermore, explainability can help uncover biases, errors, and potential risks, which are especially important when AI systems are deployed in sensitive or life-altering applications [5]. The intersection of explainability and large foundation models presents a unique set of challenges. Traditional approaches to interpretability, which work well for smaller, simpler models, struggle to scale up to the size and complexity of modern foundation models [6]. The sheer number of parameters, layers, and operations in these models makes it difficult to trace the exact reasoning behind a specific prediction or decision. Additionally, foundation models are often seen as "black-box" systems, where even the developers of the models may not fully understand how the model arrived at a particular output [7]. This lack of transparency not only hinders trust in AI but also makes it difficult to diagnose problems, improve model performance, and ensure fairness [8]. Despite these challenges, the field of explainable AI has made significant strides in recent years. Researchers have developed a variety of methods to probe and interpret the inner workings of large models. These include techniques such as saliency maps, attention mechanisms, and layer-wise relevance propagation, which aim to highlight which parts of the input data contributed most to a model's decision. Furthermore, there has been growing interest in developing post-hoc explainability methods that allow users to interpret a model's behavior without needing to understand its internal structure. While these methods have shown promise, they are not without limitations. Many of the existing techniques still face difficulties in providing meaningful and human-understandable explanations for the decisions of large foundation models. The need for explainability is further compounded by the growing use of large models in decision-making processes that directly impact individuals and society [9]. For example, AI systems used in healthcare can assist doctors in diagnosing diseases, but without clear explanations, these models could lead to potentially harmful outcomes if their recommendations are misunderstood or misinterpreted. In the financial sector, AI models that determine creditworthiness or predict market trends must be transparent to ensure fairness and prevent discrimination. Moreover, AI's increasing involvement in areas like autonomous driving and law enforcement raises concerns about accountability, where the consequences of decisions made by opaque systems can be life-altering. This paper explores the evolving landscape of explainable AI in the era of large foundation models. Specifically, we examine the challenges posed by the scale and complexity of these models, the methods being developed to address these challenges, and the broader implications for society. We will discuss the state-of-the-art approaches in XAI, focusing

on both their potential and limitations in explaining large models. Additionally, we will highlight the ethical, legal, and societal considerations that arise when deploying AI systems in high-stakes environments [10]. Our goal is to provide a comprehensive overview of the current state of XAI research, identify key areas where further work is needed, and propose a roadmap for advancing the field in a way that promotes transparency, fairness, and accountability. In the following sections, we will first delve into the theoretical foundations of explainable AI, providing an overview of the different types of explanations and their importance [11]. We will then explore the challenges and current limitations faced by XAI in the context of large foundation models, before reviewing the most promising approaches to addressing these challenges. Finally, we will conclude by discussing the future directions of research in this area and the potential societal impact of explainable AI technologies.

2 Background and Related Work

In this section, we provide a comprehensive review of the existing literature on explainable AI (XAI) and its evolution, particularly in the context of large foundation models [12]. We begin by discussing the foundational concepts of interpretability and explainability, followed by an overview of the approaches traditionally used in smaller, more interpretable AI models. We then explore how these methods have been adapted to handle the increasing complexity of large models. Finally, we examine key works in the area of XAI for large foundation models, highlighting both the progress made and the ongoing challenges [13].

2.1 Interpretability and Explainability in AI

The terms "interpretability" and "explainability" are often used interchangeably in the AI literature, but they carry subtle distinctions. *Interpretability* generally refers to the degree to which a human can understand the cause of a decision made by a model. It is concerned with the internal workings of the model and how easily they can be understood by humans. *Explainability*, on the other hand, refers to the ability to provide an intelligible justification for the behavior or decisions of a model, often in the form of a post-hoc explanation that can be easily communicated to non-experts [14]. In the early days of AI, models such as decision trees, linear regressions, and rule-based systems were prized for their interpretability. These models are relatively simple in nature and provide clear mappings between input features and predictions. For instance, in a decision tree, each decision node is associated with a feature that splits the data based on a specific criterion, making it easy to trace the path leading to a particular decision [15]. Similarly, linear regression models explicitly quantify the contribution of each feature, allowing for direct interpretability. However, as the complexity of AI models increased, particularly with the rise of deep learning, the ability to interpret and explain these models diminished. Deep neural networks, for example, consist of multiple layers of nonlinear transformations and millions

(or even billions) of parameters, making them highly opaque and difficult to understand [16]. As a result, the need for explainability methods that could work with more complex models emerged as a major research focus.

2.2 Traditional Approaches to Explainability

Various methods have been developed over the years to address the challenge of explainability in AI. In the context of machine learning, these methods can be broadly classified into two categories: *intrinsic* and *post-hoc* explainability.

Intrinsic Explainability Intrinsic explainability refers to methods that aim to make the model itself interpretable [17]. In other words, these approaches attempt to design models that are inherently transparent and provide direct explanations for their decisions [18]. Examples of intrinsically interpretable models include decision trees, linear models, and rule-based systems, as mentioned earlier [19]. These models have the advantage of being simple and directly understandable by humans, but they often struggle with handling complex tasks and large datasets. Recent advances in interpretable models have also led to the development of more sophisticated techniques such as attention mechanisms in neural networks [20]. Attention-based models, particularly in the field of natural language processing (NLP), allow the model to "focus" on specific parts of the input, which can be visually interpreted to understand which parts of the input are driving the model's predictions. While attention mechanisms have proven useful in providing a degree of interpretability, they have limitations in that the attention distribution is not always a direct explanation of the decision-making process.

Post-hoc Explainability Post-hoc explainability, in contrast, involves analyzing the behavior of a model after it has been trained to generate explanations for its predictions [21]. This approach is necessary for complex, black-box models like deep neural networks, where internal workings are not easily interpretable. Several post-hoc methods have been proposed to explain the predictions of such models. One widely used post-hoc technique is *saliency mapping*, which visualizes the parts of the input that have the greatest influence on the model's prediction [22]. Saliency maps have been particularly effective in computer vision tasks, where they highlight important regions of an image that contribute to a model's classification decision. Another popular technique is *LIME* (Local Interpretable Model-agnostic Explanations), which generates locally interpretable surrogate models around individual predictions. LIME approximates the decision boundary of a complex model in a local region of the input space, creating an interpretable explanation for a single prediction.

2.3 Explainable AI for Large Foundation Models

The emergence of large foundation models, such as GPT-3 and BERT, has raised new challenges in the realm of explainable AI [23]. These models, with their mas-

sive scale and deep architectures, are difficult to interpret and explain, especially as their decision-making processes are distributed across thousands or millions of parameters. In many cases, these models function as black boxes, where even the researchers who developed them may struggle to explain how a particular decision was made [24]. Despite these challenges, there has been significant progress in developing explainability techniques for large models [25]. One key approach is the use of *attribution methods*, which attempt to assign credit to individual input features for a given output [26]. Techniques such as Integrated Gradients, SHAP (SHapley Additive exPlanations), and Grad-CAM have gained prominence as methods for generating explanations for deep learning models, including those based on large foundation models. These methods work by quantifying the impact of each feature on the model's output and providing human-understandable explanations of the decision-making process.

Attribution Methods Attribution methods are crucial for understanding how individual components of a model contribute to its output. Integrated Gradients, for instance, measures the gradients of the model's output with respect to the input features, providing a pathway to assess how each feature influences the final prediction [27]. SHAP, based on cooperative game theory, assigns a contribution value to each feature by considering all possible combinations of features and computing their marginal contribution [28]. These methods offer more granular insights into the decision-making process, but they also have limitations in terms of scalability and interpretability when applied to large models [29].

Model-agnostic Approaches In addition to attribution methods, there has been a growing interest in model-agnostic techniques for explainability. These approaches are designed to work across different types of models, regardless of their internal architecture, and are particularly valuable when working with large foundation models. Techniques such as LIME, which creates surrogate models around individual predictions, and Counterfactual Explanations, which generate alternative scenarios that would lead to different predictions, are examples of model-agnostic explainability methods that have been applied to large models. While these techniques have proven useful, they are not without their challenges [30]. Model-agnostic methods can be computationally expensive, particularly when working with large models, and may not always produce explanations that align with the underlying behavior of the model. Furthermore, they often lack the level of precision and granularity required to provide truly actionable insights into the decision-making process [31].

2.4 Challenges and Limitations

Despite the progress made in the field of XAI, significant challenges remain, particularly when it comes to large foundation models. One of the main issues is the trade-off between model performance and interpretability [32]. Large models are often able to achieve state-of-the-art results due to their complexity and vast

parameter space, but this very complexity makes them difficult to interpret and explain [33]. As a result, there is an inherent tension between making models interpretable and maintaining their high performance. Another challenge lies in the subjective nature of explainability. Different stakeholders may require different levels of explanation depending on their expertise and the application domain [34]. For example, a medical practitioner may need a detailed, transparent explanation to understand the reasoning behind a model's diagnosis, whereas a layperson might only require a high-level justification [35]. Designing XAI methods that can cater to diverse audiences is an ongoing challenge that requires both technical and ethical considerations. Finally, there is the issue of *trust* in AI. Even with explanations, users may not always trust the decision-making process of AI systems, particularly if they cannot fully comprehend the model's reasoning. In the case of large models, trust is further complicated by the "black-box" nature of these systems. As such, building trust through explainability is a critical aspect of ensuring the successful adoption and deployment of AI technologies in high-stakes domains.

2.5 Conclusion

In summary, the field of explainable AI has made significant advancements, but the rise of large foundation models has introduced new challenges in providing meaningful and transparent explanations. While a variety of techniques exist to interpret and explain AI models, the complexity of large models requires innovative methods that balance performance with interpretability. The next section will explore the state-of-the-art approaches in XAI and discuss how they can be applied to address the unique challenges posed by large foundation models.

3 Challenges in Explainability for Large Foundation Models

While significant strides have been made in developing methods for explainable AI (XAI), large foundation models, such as GPT-3, BERT, and similar architectures, present unique and unprecedented challenges in the realm of interpretability and explainability. These models, characterized by their massive parameter space, deep architectures, and the vastness of the data they are trained on[36, 37], complicate traditional methods of generating meaningful and transparent explanations. In this section, we explore the primary challenges associated with explainability for large foundation models, including the complexity of the models themselves, the trade-offs between performance and interpretability, the difficulty in understanding model decisions at scale, and issues related to the transparency of the training process.

3.1 Model Complexity and Scale

The defining characteristic of large foundation models is their size [38]. These models often have billions or even trillions of parameters, which makes them in-

herently difficult to interpret. The sheer scale of the models adds multiple layers of complexity: large models generally consist of many more layers, have a greater number of operations, and involve highly complex interactions between parameters that are not easily understood [39]. In comparison to simpler models such as decision trees or linear regression, which have transparent decision-making processes, large foundation models operate through highly abstract learned representations, often making it difficult to trace how specific inputs contribute to outputs. For instance, models like GPT-3 are based on transformer architectures, which rely on multi-head attention mechanisms to process input sequences [40]. While attention mechanisms offer some insight into how a model allocates focus to different parts of the input, the complexity of these mechanisms in large models makes it difficult to pinpoint which aspects of the input led to a specific output. Additionally, since large models often involve multiple layers of abstraction, understanding how information propagates through these layers and how it influences final predictions becomes a highly non-trivial task. Another factor contributing to complexity is the non-linear interactions that occur between various components of a model [41]. Unlike simpler models where feature importance can be directly measured, large models are highly non-linear, and the contribution of each parameter or feature may not be immediately apparent [42]. This complicates the task of explaining the specific decisions made by the model, as the decision-making process is distributed across numerous parameters and layers, each of which may have a nuanced impact on the outcome.

3.2 Trade-off between Performance and Interpretability

One of the core challenges of large foundation models is the trade-off between model performance and interpretability. In general, larger models with more parameters tend to deliver better performance on a wide range of tasks. These models are capable of capturing complex patterns and relationships in data, which allows them to excel in areas such as natural language understanding, image recognition, and even generative tasks [43]. However, this increased performance comes at a cost [44]. As models grow in size and complexity, they become more difficult to interpret. The increase in parameters and layers introduces a level of opacity that diminishes the ability to understand the decision-making process. This trade-off has significant implications for domains where interpretability is critical, such as healthcare, criminal justice, and finance. In these contexts, the ability to trust and understand the reasoning behind AI decisions is paramount. Unfortunately, achieving higher performance often means sacrificing some degree of interpretability, creating a difficult balancing act between maximizing performance and ensuring transparency. For example, a model that can predict patient outcomes with high accuracy may be less interpretable if it requires millions of parameters to achieve that level of performance [45]. On the other hand, a more interpretable model, such as a decision tree, may offer less predictive power, thereby making it harder to achieve state-of-the-art results on complex tasks [46]. This dilemma is central to the ongoing discourse in XAI and underscores

the need for new techniques that can provide meaningful explanations without compromising model effectiveness.

3.3 Understanding Model Decisions at Scale

As the size of foundation models increases, so too does the difficulty in understanding their decisions at scale [47]. Large models are often trained on vast amounts of data, which can make it difficult to understand how specific data points or features influence the model's outputs. The sheer volume of data can overwhelm traditional interpretability methods, which often focus on explaining individual predictions or small subsets of data [48]. With large models, explanations must account for the influence of a large number of features and how they interact, which can result in explanations that are difficult to grasp or overly complex. Moreover, large foundation models can make use of highly sophisticated learned representations, such as embeddings, which are not directly interpretable [49]. For instance, in the case of NLP models, input text is often transformed into dense vectors (embeddings) that encapsulate semantic meaning in a way that is not immediately understandable by humans. While these embeddings are highly effective for model performance, their abstract nature makes it challenging to interpret the model's reasoning process in terms that a human could easily follow. Additionally, scaling the interpretability of large models often requires examining an extensive set of features, parameters, and decision-making paths [50]. With billions of parameters, the computational resources required to generate and process explanations at scale become prohibitively expensive [51]. This challenge makes it increasingly difficult to provide explanations that are both accurate and computationally feasible for large foundation models.

3.4 Opacity of the Training Process

Another critical challenge in explaining large foundation models is the opacity of the training process itself [52]. Many large models, particularly those used in deep learning, rely on complex training regimes involving unsupervised or semi-supervised learning, massive datasets, and distributed computing architectures. The resulting models are not just difficult to interpret but are also opaque in terms of how they learn and evolve over time [53]. For example, large language models like GPT-3 are trained on enormous corpora of text data and learn patterns and relationships from the data in a highly unsupervised manner. While this training process allows the model to capture a wide range of linguistic structures and knowledge, it also means that the model may learn spurious or unintended patterns that are not immediately observable or explainable. Moreover, the distributed nature of modern training processes, where training is often performed across multiple GPUs or even data centers, adds another layer of complexity to understanding how a model is trained. The opacity of the training process further complicates efforts to explain the model's behavior [54]. Traditional methods of interpreting models often focus on analyzing the parameters and weights of a trained model [55]. However, for large foundation models, these

methods become less effective as the sheer number of parameters and the complexity of their interactions make it impossible to interpret the model in any straightforward way. Understanding how the model was trained and why it behaves the way it does often requires insights into the dataset, the optimization process, and the model architecture—factors that are not always accessible or transparent [56].

3.5 Ethical and Legal Considerations

The challenges associated with explaining large foundation models also intersect with important ethical and legal considerations. As AI systems are deployed in real-world applications, the inability to explain their decisions can lead to unintended consequences, such as bias, discrimination, or unfair treatment of individuals. In domains like healthcare, law enforcement, and finance, where decisions can significantly impact people's lives, the lack of transparency raises serious ethical concerns [57]. For example, consider an AI system that is used to determine eligibility for a loan [58]. If the system makes a decision based on complex and opaque criteria, it could potentially discriminate against certain groups without any clear justification. Similarly, in the case of autonomous vehicles, if an AI system makes a decision that results in harm, it is crucial to understand the reasoning behind that decision to assign responsibility and ensure accountability [59]. Furthermore, the opacity of AI models raises legal concerns, especially when it comes to regulatory compliance and accountability. In many jurisdictions, there are legal requirements for AI systems to provide explanations for their decisions, particularly in high-stakes domains such as finance or healthcare. The inability to offer clear and understandable explanations for model outputs can create challenges for meeting these legal requirements and ensuring that AI systems are used in a responsible and ethical manner [60].

3.6 Conclusion

The challenges associated with explainability in large foundation models are numerous and multifaceted. The complexity and scale of these models, combined with the trade-offs between performance and interpretability, pose significant barriers to providing meaningful and transparent explanations [61]. Additionally, the opacity of the training process and the difficulty in understanding model decisions at scale further complicate the task of explainability [62]. However, addressing these challenges is crucial for ensuring that large foundation models are deployed responsibly and ethically, especially in high-stakes applications [63]. The following section will discuss current efforts to mitigate these challenges and explore potential future directions for research in XAI for large foundation models [64].

4 State-of-the-Art Approaches in Explainable AI for Large Foundation Models

As large foundation models, such as GPT-3, BERT, and other transformer-based architectures, have become increasingly prevalent, significant efforts have been made to develop explainability techniques that can scale to the complexity of these models. In this section, we explore the state-of-the-art approaches in explainable AI (XAI) specifically designed for large foundation models. We will discuss various methodologies for providing post-hoc interpretability, model-agnostic approaches, and emerging techniques that aim to integrate explainability into the design and training of large models. These techniques can help us understand the decision-making process behind these black-box systems while addressing the challenges identified in the previous section [65].

4.1 Attribution Methods for Large Models

Attribution methods aim to assign importance scores to input features to explain how each feature influences the model's output [66]. These methods can be critical for understanding the behavior of large foundation models, as they provide a way to break down the contribution of individual components within complex architectures.

Integrated Gradients Integrated Gradients (IG) is one of the most widely used attribution methods, particularly for deep neural networks [67]. IG provides a way to attribute the model's output to the input features by computing the integral of the gradients of the output with respect to the input along the path from a baseline input (often a vector of zeros or another reference input) to the actual input. This method is well-suited for large foundation models, as it allows for gradient-based analysis of individual features and provides a clear explanation of which features are most responsible for a model's prediction. Integrated Gradients can be particularly effective for large NLP models such as BERT and GPT-3, where the contribution of each word or token in a sequence is assessed with respect to the model's output [68]. Despite its advantages, IG can be computationally expensive, especially for large models with vast parameter spaces, and it may struggle with capturing interactions between input features that are highly non-linear [69].

SHAP (SHapley Additive exPlanations) SHAP is another popular attribution method based on cooperative game theory, which provides a unified measure of feature importance by calculating the average contribution of each feature across all possible combinations of features. In essence, SHAP uses Shapley values from game theory to assign an importance value to each feature in a prediction [70]. SHAP has been widely adopted in the XAI community due to its theoretical foundations and consistency across different model types [71]. SHAP has been successfully applied to large foundation models, especially in NLP

tasks, where the importance of each token or word can be computed based on the model's output. The method provides an interpretable, additive explanation for the predictions made by a model. However, SHAP can be computationally expensive, particularly when applied to large datasets or models with millions of parameters, due to the need to compute all possible feature subsets.

Grad-CAM (Gradient-weighted Class Activation Mapping) Grad-CAM is a visualization technique that provides insights into the parts of an input that contribute most to the model's decision. It is often used for convolutional neural networks (CNNs) in computer vision tasks, but it has also been adapted for transformer-based architectures [72]. Grad-CAM computes the gradient of the predicted class with respect to the feature maps in the final convolutional layer, highlighting the regions of the input that most strongly influence the model's decision [73]. While Grad-CAM has proven effective in explaining image-based tasks, its adaptation to large foundation models, particularly those used in NLP and multimodal tasks, has been a subject of active research [74]. Recent work has explored how to extend Grad-CAM to transformers by using attention weights and final hidden states to identify the tokens or sections of an input sequence that are most relevant to the model's decision. This approach has shown promise in providing visual explanations for text-based tasks, but challenges remain in fully capturing the complex reasoning process that occurs in large models.

4.2 Model-Agnostic Explainability Techniques

Model-agnostic methods are designed to work independently of the underlying architecture of the model, allowing for the interpretation of any model, including large foundation models [75]. These techniques focus on generating explanations that do not require access to the internals of the model, but instead rely on approximating the decision-making process using simpler, more interpretable surrogate models or other techniques.

LIME (Local Interpretable Model-agnostic Explanations) LIME is a well-known model-agnostic technique that aims to explain individual predictions by approximating the local decision boundary of a complex model with a simpler, interpretable model [76]. LIME works by perturbing the input data to create a set of similar but slightly modified samples and then training a surrogate model on this local region. The surrogate model is typically interpretable (e.g., a linear model or decision tree), and it provides insights into how the original complex model arrived at the given prediction. LIME can be applied to large foundation models by generating local approximations of the model's behavior around a specific input [77]. This makes it possible to generate explanations for complex predictions without needing to fully understand the underlying model [78]. However, LIME's reliance on creating perturbed versions of the input data can be computationally expensive, especially when working with large datasets or high-dimensional input spaces such as text or images. Additionally, the quality

of the explanations generated by LIME can depend on how well the surrogate model approximates the complex model's behavior [79].

Counterfactual Explanations Counterfactual explanations provide a form of interpretability by showing what changes would need to be made to the input in order to achieve a different outcome [80]. In other words, they explain what the model would have predicted if certain features of the input had been different [81]. For example, in a binary classification task, a counterfactual explanation would show which input features would need to be modified in order to change the model's decision from "not approved" to "approved."

Counterfactual explanations are particularly useful because they are intuitive and easy to understand, as they provide concrete examples of how changes to the input can affect the output [82]. These explanations are inherently human-friendly, as they provide actionable insights into how specific aspects of the data influence the decision. However, generating counterfactual explanations for large models can be challenging, as it often requires navigating complex feature spaces and generating meaningful perturbations that result in realistic, interpretable explanations.

4.3 Interpretable Attention Mechanisms

Attention mechanisms, particularly in transformer models, have gained significant attention in the XAI community due to their ability to provide insight into how a model processes input data [83]. In large foundation models such as BERT and GPT-3, attention mechanisms allow the model to "focus" on specific tokens or parts of the input when making predictions. This has led to the development of several techniques aimed at interpreting the attention weights and understanding how they influence the model's decision [84].

Attention Visualization Attention visualization techniques aim to provide a visual interpretation of the attention weights in transformer-based models. By visualizing how the model attends to different tokens in a sequence, these methods can offer insight into which parts of the input are deemed important for the model's decision [85]. Attention maps can be plotted to show the flow of information through the model and identify key interactions between tokens. Although attention visualization has proven useful in many cases, it is not without its limitations. The attention mechanism in transformer models does not always correspond directly to human-like reasoning, meaning that high attention scores do not always indicate that a model is using the "right" information. Moreover, the complexity of large models means that interpreting attention scores across multiple layers and heads can be challenging [86].

Interpretable Embeddings Another area of focus in explainability for large foundation models is the interpretation of embeddings. In models such as BERT

and GPT-3, the input tokens are transformed into dense vector representations (embeddings) that capture semantic meaning [87]. These embeddings are highly effective for task-specific performance, but they can be difficult to interpret. Recent research has focused on developing methods to visualize and understand embeddings, such as by projecting them into lower-dimensional spaces (e.g., using t-SNE or PCA) or by identifying clusters of similar embeddings. Interpretable embeddings allow for a more intuitive understanding of how models represent and process input data. These techniques provide a way to analyze which aspects of the data are emphasized in the model's learned representations and how these representations influence the final prediction. However, visualizing and interpreting embeddings at scale remains a challenge, particularly in models that operate over large corpora or multimodal data.

4.4 Integrating Explainability into Model Design

As the demand for explainable AI continues to grow, there is increasing interest in integrating explainability directly into the design and training of large foundation models. Instead of relying solely on post-hoc methods, researchers are exploring how to build explainable components into the models themselves during training. This includes designing models that produce more interpretable representations, as well as developing training objectives that promote transparency. For example, recent work has focused on developing *self-explaining models*, where the model generates its own explanations alongside its predictions. These models are designed with explainability in mind, and they aim to produce explanations that are both accurate and faithful to the underlying decision-making process. By integrating explainability into the model design, it is hoped that large foundation models can achieve both high performance and transparency without relying on external interpretability methods.

4.5 Conclusion

State-of-the-art approaches to explainable AI for large foundation models have made significant progress in providing more transparent, understandable, and interpretable models. Attribution methods, model-agnostic techniques, and interpretable attention mechanisms have all contributed to our ability to understand how large models make decisions [88]. However, challenges remain in balancing model performance with interpretability, especially as the scale of models continues to grow [89]. The integration of explainability into the design of models is a promising direction for future research, offering the potential for more transparent AI systems that are both effective and interpretable by design [90].

5 Future Directions in Explainable AI for Large Foundation Models

While substantial progress has been made in the development of explainable AI (XAI) techniques for large foundation models, many challenges remain [91]. As

these models continue to evolve, both in size and complexity, there is a growing need for innovative solutions that can provide transparent, interpretable, and accountable AI systems. This section explores potential future directions in the field of XAI for large foundation models, highlighting emerging research areas and approaches that could address the current limitations and advance the state-of-the-art in explainability.

5.1 Towards Self-Explainable Models

One promising direction for the future of XAI in large foundation models is the development of self-explainable models [92]. These models would not only provide accurate predictions but would also generate their own explanations alongside those predictions. This contrasts with current methods that typically rely on post-hoc explanations, which are generated after the model has made a decision. Self-explainable models could use a variety of techniques to generate explanations directly [93]. For example, a model might produce an interpretable rationale or justification for its output, potentially in natural language or through visualizations, depending on the task. Recent research in natural language processing (NLP) has explored using language models to generate textual explanations, where the model can output a series of logical steps or reasoning that led to its decision [94]. Similarly, self-explanatory models in computer vision could explain their predictions by highlighting key regions in the input image and providing verbal descriptions of what was important for the model's decision [95]. Such self-explaining models could represent a breakthrough in XAI, as they would offer a more seamless and integrated form of interpretability, making the reasoning process inherently tied to the model's predictions [96]. However, the development of such models poses significant challenges, as it requires the models to not only be accurate but also capable of generating meaningful, coherent, and human-understandable explanations.

5.2 Explainability in Multimodal Models

As large foundation models increasingly handle multimodal data—such as text, images, and audio—developing techniques that can explain predictions across different modalities is a critical research area [97]. Multimodal models, which process multiple types of input simultaneously, present unique challenges in terms of explainability. The interactions between different modalities can be highly complex, and understanding how the model integrates information from different sources is crucial for providing accurate explanations [98]. For example, in a multimodal model that combines text and images, the model may make a prediction based on both the content of an image and the surrounding text [99]. Understanding how these two modalities influence the final decision requires new methods for explaining not just individual modalities but also their interaction. Current approaches to explainability often treat each modality separately, but the future of XAI in multimodal models will likely involve integrated techniques

that can jointly explain decisions across different types of input [100]. Furthermore, these models will need to handle the alignment between modalities effectively [101]. For instance, how does the model align visual objects with specific words or concepts in a textual description? Addressing these questions and developing methods that can provide explanations for multimodal decision-making will be a key area of focus in the coming years [72].

5.3 Improving Post-Hoc Explanation Methods for Large Models

Although post-hoc explanation methods, such as Integrated Gradients, SHAP, and LIME, have been extensively used for explaining large foundation models, these methods are still subject to several limitations, including computational expense and difficulty in capturing complex interactions between features [102]. Future research could focus on improving the efficiency, scalability, and accuracy of these methods, making them more suitable for large models with billions of parameters [103]. One potential avenue for improvement is the development of more efficient algorithms for calculating attributions, especially for highly complex models [104]. Researchers could explore ways to reduce the number of computations required to generate explanations or apply approximation techniques that preserve the accuracy of the explanation while improving efficiency. For example, novel methods for approximating Shapley values or integrating gradient-based techniques with other interpretability strategies could be explored. Another important direction is the refinement of post-hoc methods to capture higher-order interactions between input features [105]. Current methods typically focus on the individual importance of each feature, but many decision-making processes in large models involve complex interactions between multiple features. Developing post-hoc techniques that can effectively capture and explain these interactions will be crucial for providing a deeper understanding of model behavior.

5.4 Interpretable Neural Architecture Design

As the architecture of large foundation models continues to evolve, there is increasing interest in designing models with built-in interpretability. One approach is to integrate explainability into the architecture itself, rather than relying on external post-hoc methods. This could involve designing neural networks with transparent or inherently interpretable components, such as attention mechanisms or modular architectures that allow for easier understanding of the decision-making process[106, 107]. For instance, one possible direction is the development of "modular" neural architectures, where different components of the model specialize in different aspects of the decision-making process. By separating the model's decision process into distinct, understandable modules, it may be possible to generate more interpretable outputs [108]. Similarly, attention mechanisms, which have already been used in transformer models to highlight important tokens or regions in input data, could be extended to other forms of neural architectures to improve interpretability [109]. Another approach is to design neural networks that explicitly incorporate constraints or regularization terms

to promote interpretability [110]. For example, regularization strategies could be used to encourage the model to produce sparse representations or to make its reasoning process more transparent [111]. By incorporating interpretability into the model design itself, it may be possible to achieve a more natural balance between performance and transparency [112].

5.5 Human-in-the-Loop Explainability

An emerging trend in XAI is the inclusion of human-in-the-loop (HITL) methodologies, where the explanations provided by AI models are used in conjunction with human expertise to improve decision-making. In many real-world applications, AI systems are used in collaboration with human experts, who need to trust the model's predictions and explanations to make informed decisions. As such, it is crucial to design explanation systems that can work alongside human decision-makers, enhancing their understanding of the model's behavior and supporting their ability to act on its outputs [113]. In HITL frameworks, the AI system could provide explanations that are tailored to the user's level of expertise and the context of the decision-making task [114]. For example, a medical AI system might generate simple, intuitive explanations for a doctor, while providing more detailed technical explanations for researchers. The human expert would then be able to review the explanation, verify its validity, and provide feedback, which could be used to refine the model and its explanations over time. HITL approaches are particularly important in high-stakes domains such as healthcare, finance, and criminal justice, where the implications of AI-driven decisions can have profound consequences. By incorporating human judgment into the explanation process, it is possible to create a more robust and transparent AI system that can be trusted in critical applications.

5.6 Ethical Considerations and Accountability

As large foundation models become more integrated into decision-making systems, ensuring their accountability and fairness is becoming an increasingly important research area [115]. The opacity of these models raises significant ethical concerns, particularly in domains where AI decisions can have serious consequences for individuals and society. Future research must prioritize the development of explanation methods that not only make AI systems more transparent but also help ensure that these systems are fair, unbiased, and accountable [116]. For example, one important area of focus is the development of methods that can explain and mitigate bias in large models [117]. Many large models, especially those trained on massive, unfiltered datasets, can inadvertently learn and propagate harmful biases [118]. Developing techniques to identify, explain, and reduce bias in model predictions will be crucial for ensuring that AI systems are ethical and do not perpetuate discrimination [119]. Moreover, the accountability of AI systems, especially in high-stakes domains, requires that explanations are not only accurate but also provide insight into the model's decision-making process in a way that allows for meaningful oversight. The development of XAI

methods that can be used to audit and review AI systems for ethical compliance will play a critical role in ensuring that large foundation models are deployed responsibly [120].

5.7 Conclusion

The future of explainable AI for large foundation models is an exciting and rapidly evolving area of research [121]. As these models continue to grow in size and complexity, the need for more effective, scalable, and interpretable methods becomes even more critical. Self-explainable models, multimodal explainability, improved post-hoc explanation methods, interpretable neural architecture design, and human-in-the-loop approaches represent some of the most promising directions for future research [122]. By addressing the challenges of model complexity, fairness, and accountability, we can ensure that large foundation models remain both powerful and transparent, facilitating their responsible use in real-world applications [123].

6 Conclusion

The rapid development and deployment of large foundation models, such as GPT-3, BERT, and other transformer-based architectures, have revolutionized the field of artificial intelligence across a broad range of applications, from natural language processing and computer vision to multimodal systems. However, the complexity and size of these models pose significant challenges to transparency, interpretability, and trust, which are essential for the responsible adoption of AI technologies. This paper has explored the growing need for explainable AI (XAI) in the era of large foundation models, outlining the challenges and limitations of current approaches, discussing state-of-the-art techniques for providing interpretability, and highlighting promising future directions in the field.

The landscape of explainability for large foundation models is rapidly evolving, with a diverse range of methods available to interpret these complex systems. Attribution techniques, such as Integrated Gradients, SHAP, and LIME, have been foundational in providing insights into how individual features contribute to model predictions. Additionally, model-agnostic methods, such as counterfactual explanations and LIME, offer flexible ways to understand and explain the decision-making process of any model, including large transformer-based architectures. These techniques provide a valuable means of gaining insight into the black-box nature of large models, but they also present challenges related to computational complexity, scalability, and the need to capture higher-order feature interactions.

The growing importance of multimodal models, which process text, images, audio, and other types of data simultaneously, has introduced new challenges for explainability. As these models become increasingly complex, the development of integrated techniques that can explain decisions across different modalities will be critical. Furthermore, the exploration of self-explainable models—models

that inherently generate explanations for their predictions—is an exciting and promising avenue for advancing the field of XAI. By building explanations directly into the architecture of the model, it is possible to create systems that are both accurate and transparent by design.

Looking ahead, the integration of XAI methods into the design and training of large foundation models is expected to play a key role in advancing the transparency of AI systems. Additionally, the inclusion of human-in-the-loop methodologies, where human experts collaborate with AI systems to verify and refine model explanations, will likely enhance trust and accountability in decision-making. Ethical considerations, such as ensuring fairness, transparency, and accountability, will also remain a priority as AI systems are deployed in high-stakes applications such as healthcare, finance, and law enforcement.

In conclusion, while significant progress has been made in the development of explainable AI techniques for large foundation models, much work remains to be done. The future of XAI lies in creating models that balance performance with interpretability, ensuring that AI systems are not only powerful but also transparent, ethical, and accountable. By addressing the challenges associated with model complexity, fairness, and human oversight, researchers and practitioners can build AI systems that are more trustworthy and better aligned with societal values. Ultimately, the continued advancement of XAI for large foundation models will play a crucial role in fostering broader acceptance and responsible use of AI technologies in real-world applications.

References

1. Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.
2. Jiajin Li, Steve King, and Ian Jennions. Intelligent fault diagnosis of an aircraft fuel system using machine learning—a literature review. *Machines*, 11(4):481, 2023.
3. Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.
4. Jörn Lötsch, Dario Kringel, and Alfred Ultsch. Explainable Artificial Intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics*, 2(1):1–17, 2021.
5. KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of Artificial Intelligence*, pages 603–649, 2020.
6. AR Troncoso-García, M Martínez-Ballesteros, Francisco Martínez-Álvarez, and A Troncoso. Explainable machine learning for sleep apnea prediction. *Procedia Computer Science*, 207:2930–2939, 2022.
7. Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5):593, January 2021. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

8. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
9. Timo Speith. A review of taxonomies of Explainable Artificial Intelligence (XAI) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.
10. Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.
11. Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. LLMs for XAI: Future Directions for Explaining Explanations, May 2024. arXiv:2405.06064 [cs].
12. Simi Job, Xiaohui Tao, Lin Li, Haoran Xie, Taotao Cai, Jianming Yong, and Qing Li. Optimal treatment strategies for critical patients with deep reinforcement learning. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–22, 2024.
13. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners, January 2023. arXiv:2205.11916 [cs].
14. Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021.
15. Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pages 272–281. PMLR, 2019.
16. Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
17. Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.
18. Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Transactions on Artificial Intelligence*, 2023.
19. Ruchi Verma, Joshita Sharma, and Shagun Jindal. Time Series Forecasting Using Machine Learning. In *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4*, pages 372–381. Springer, 2020.
20. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
21. Daniel Holliday, Stephanie Wilson, and Simone Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168, 2016.
22. Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. Can Large Language Models be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate, January 2024. arXiv:2401.16788 [cs].
23. Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai):

What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

24. Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.
25. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.
26. Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, Honolulu HI USA, May 2024. ACM.
27. Cynthia Rudin and Joanna Radin. 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*, 1(2):1–9, 2019.
28. Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations, October 2023. arXiv:2310.11207 [cs].
29. Erika Puiutta and Eric MSP Veith. Explainable reinforcement learning: A survey. In *International Cross-domain Conference for Machine Learning and Knowledge Extraction*, pages 77–95. Springer, 2020.
30. Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2022.
31. Daniel V McGehee, Mark Brewer, Chris Schwarz, Bryant Walker Smith, et al. Review of automated vehicle technology: Policy and implementation implications. Technical report, Iowa. Dept. of Transportation, 2016.
32. Guang Yang, Felix Raschke, Thomas R Barrick, and Franklyn A Howe. Manifold Learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering. *Magnetic Resonance in Medicine*, 74(3):868–878, 2015.
33. Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of Similarity-based Explanations, March 2021. arXiv:2006.04528 [cs, stat].
34. Erico Tjoa and Cuntai Guan. A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2020.
35. Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017.
36. Dylan Molho, Jiayuan Ding, Wenzhuo Tang, Zhaoheng Li, Hongzhi Wen, Yixin Wang, Julian Venegas, Wei Jin, Renming Liu, Runze Su, et al. Deep learning in single-cell analysis. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–62, 2024.
37. Yassine Zniyed, Thanh Phuong Nguyen, et al. Efficient tensor decomposition-based filter pruning. *Neural Networks*, 178:106393, 2024.
38. Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022.

39. Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021.
40. Saqib Aziz, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. Machine learning in finance: A topic modeling approach. *European Financial Management*, 28(3):744–770, 2022.
41. Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.
42. Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.
43. Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023.
44. Jodie Crocker, Krishna Kumar, and B Cox. Using explainability to design physics-aware CNNs for solving subsurface inverse problems. *Computers and Geotechnics*, 159:105452, 2023.
45. Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
46. Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
47. Kyungtae Lee, Mukil V Ayyasamy, Yangfeng Ji, and Prasanna V Balachandran. A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys. *Scientific Reports*, 12(1):11591, 2022.
48. Ajay Thampi. *Interpretable AI: Building explainable machine learning systems*. Simon and Schuster, 2022.
49. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.
50. Francesco Carlo Morabito, Cosimo Ieracitano, and Nadia Mammone. An explainable Artificial Intelligence approach to study MCI to AD conversion via HD-EEG processing. *Clinical EEG and Neuroscience*, 54(1):51–60, 2023.
51. Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 2021.
52. Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2020.
53. Jordan D Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, and Maryellen L Giger. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics*, 49(1):1–14, 2022.
54. Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. Open Source Language Models Can Provide Feedback: Evaluating LLMs' Ability to Help Students Using GPT-4-As-A-Judge, May 2024. arXiv:2405.05253 [cs].

55. Wei Sun. *Stability of machine learning algorithms*. PhD thesis, Purdue University, 2015.
56. Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. Od-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences*, 12(11):5310, 2022.
57. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
58. Wei Song, Lu Liu, Minghao Liu, Wenxiang Wang, Xiao Wang, and Yu Song. Representation learning with deconvolution for multivariate time series classification and visualization. In *Data Science: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2020, Taiyuan, China, September 18-21, 2020, Proceedings, Part I 6*, pages 310–326. Springer, 2020.
59. Sina Mohseni, Jeremy E Block, and Eric D Ragan. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*, 2018.
60. Nick Bostrom and Eliezer Yudkowsky. The ethics of Artificial Intelligence. In *Artificial Intelligence Safety and Security*, pages 57–69. Chapman and Hall/CRC, 2018.
61. Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors*, 23(2):634, 2023.
62. Malvika Sharma, Carl Savage, Monika Nair, Ingrid Larsson, Petra Svedberg, and Jens M Nygren. Artificial intelligence applications in health care practice: scoping review. *Journal of Medical Internet Research*, 24(10):e40238, 2022.
63. Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Regulation (eu)*, 679:2016, 2016.
64. Zichen Chen, Jianda Chen, Mitali Gaidhani, Ambuj Singh, and Misha Sra. XplainLLM: A QA Explanation Dataset for Understanding LLM Decision-Making, November 2023. arXiv:2311.08614 [cs].
65. Kanika Goel, Renuka Sindhgatta, Sumit Kalra, Rohan Goel, and Preeti Mutreja. The effect of machine learning explanations on user trust for automated diagnosis of covid-19. *Computers in Biology and Medicine*, 146:105587, 2022.
66. Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
67. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1441–1448, 2014.
68. Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.
69. Alnour Alharin, Thanh-Nam Doan, and Mina Sartipi. Reinforcement learning interpretation methods: A survey. *IEEE Access*, 8:171058–171077, 2020.
70. Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*, 2023.
71. Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

72. Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.
73. Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, and Eugenio di Sciascio. XAI4LLM. Let Machine Learning Models and LLMs Collaborate for Enhanced In-Context Learning in Healthcare, June 2024. arXiv:2405.06270 [cs].
74. Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects, February 2019. arXiv:1812.04608 [cs].
75. Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li. Designing ecg monitoring healthcare system with federated transfer learning and explainable AI. *Knowledge-Based Systems*, 236:107763, 2022.
76. Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1398, 2019.
77. Matti Tedre, Tapani Toivonen, Juho Kahila, Henriikka Vartiainen, Teemu Valtonen, Ilkka Jormanainen, and Arnold Pears. Teaching machine learning in k–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE access*, 9:110558–110572, 2021.
78. Shafie Gholizadeh and Nengfeng Zhou. Model explainability in deep learning based natural language processing. *arXiv preprint arXiv:2106.07410*, 2021.
79. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.
80. Bo Wang, Jianlong Zhou, Yiqiao Li, and Fang Chen. Impact of fidelity and robustness of machine learning explanations on user trust. In *Australasian Joint Conference on Artificial Intelligence*, pages 209–220. Springer, 2023.
81. Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.
82. Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, pages 93–105, August 2023. arXiv:2306.05715 [cs].
83. Ghanshyam Pilania. Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science*, 193:110360, 2021.
84. Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
85. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. arXiv:2203.02155 [cs].
86. Quang-Hien Kha, Viet-Huan Le, Truong Nguyen Khanh Hung, Ngan Thi Kim Nguyen, and Nguyen Quoc Khanh Le. Development and validation of an explain-

able machine learning-based prediction model for drug–food interactions from chemical structures. *Sensors*, 23(8):3962, 2023.
87. Ryuji Hamamoto. Application of artificial intelligence for medical research, 2021.
88. Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 53–56, 2018.
89. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExt-GPT: Any-to-any multimodal LLM. *arXiv preprint arXiv:2309.05519*, 2023.
90. Guido Schryen. Speedup and efficiency of computational parallelization: A unifying approach and asymptotic analysis. *arXiv preprint arXiv:2212.11223*, 2022.
91. Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.
92. Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s):1–42, December 2023.
93. Tobias Harren, Hans Matter, Gerhard Hessler, Matthias Rarey, and Christoph Grebner. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *Journal of Chemical Information and Modeling*, 62(3):447–462, 2022.
94. Ričards Marcinkevičs and Julia E Vogt. Interpretability and explainability: A machine learning zoo mini-tour. *arXiv preprint arXiv:2012.01805*, 2020.
95. Yuhong Mo, Hao Qin, Yushan Dong, Ziyi Zhu, and Zhenglin Li. Large language model (llm) ai text generation detection based on transformer deep learning algorithm. *arXiv preprint arXiv:2405.06652*, 2024.
96. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
97. Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.
98. Daniel P Malone and John F Creamer. NHTSA and the next 50 years: Time for congress to act boldly (again). Technical report, SAE Technical Paper, 2016.
99. Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, February 2021.
100. Ali AlShami, Terrance Boult, and Jugal Kalita. Pose2Trajectory: Using transformers on body pose to predict tennis player's trajectory. *Journal of Visual Communication and Image Representation*, 97:103954, 2023.
101. Dongyuan Wu, Liming Nie, Rao Asad Mumtaz, and Kadambri Agarwal. A llm-based hybrid-transformer diagnosis system in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 2024.
102. Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018.
103. Thu Trang Nguyen, Thach Le Nguyen, and Georgiana Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 77–94. Springer, 2020.

104. AS Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, OS Albahri, AH Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and Explainable Artificial Intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 2023.
105. A Shrivastava, P Kumar, Anubhav, C Vondrick, W Scheirer, DS Prijatelj, M Jafarzadeh, T Ahmad, S Cruz, R Rabinowitz, et al. Novelty in image classification. In *A Unifying Framework for Formal Theories of Novelty: Discussions, Guidelines, and Examples for Artificial Intelligence*, pages 37–48. Springer, 2023.
106. Maya Grace Torii, Takahito Murakami, and Yoichi Ochiai. Expanding Horizons in HCI Research Through LLM-Driven Qualitative Analysis, January 2024. arXiv:2401.04138 [cs].
107. Yassine Zniyed, Thanh Phuong Nguyen, et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
108. Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.
109. Jinkyu Kim, Anna Rohrbach, Zeynep Akata, Suhong Moon, Teruhisa Misu, Yi-Ting Chen, Trevor Darrell, and John Canny. Toward explainable and advisable model for self-driving cars. *Applied AI Letters*, 2(4):e56, 2021.
110. Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.
111. Kushagra Agrawal, Nirmit Desai, and Tanmoy Chakraborty. Time series visualization using t-SNE and UMAP. *Journal of Big Data*, 8(1):1–21, 2021.
112. Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.
113. Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.
114. Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.
115. Uche Onyekpe, Yang Lu, Eleni Apostolopoulou, Vasile Palade, Eyo Umo Eyo, and Stratis Kanarachos. Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP. In *Explainable AI: Foundations, Methodologies and Applications*, pages 157–183. Springer, 2022.
116. Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*, 2024.
117. Shaker El-Sappagh, Jose M Alonso, SM Riazul Islam, Ahmad M Sultan, and Kyung Sup Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports*, 11(1):2660, 2021.
118. Sangwon Lee, Junho Hong, Ling Liu, and Wonik Choi. Ts-fastformer: Fast transformer for time-series forecasting. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–20, 2024.

119. Zhangxing Bian, Siyu Xia, Chao Xia, and Ming Shao. Weakly supervised vitiligo segmentation in skin image through saliency propagation. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 931–934. IEEE, 2019.
120. Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.
121. Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.
122. Bereket A Yilma, Chan Mi Kim, Gerald C Cupchik, and Luis A Leiva. Artful path to healing: Using machine learning for visual art recommendation to prevent and reduce post-intensive care syndrome (pics). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.
123. Anirban Roy, Laurens van der Maaten, and Daniela Witten. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 16(3):e1009043, 2020.